HAL
open science

Analyse computationnelle des éléments cis-régulateurs dans les génomes des drosophiles et des mammifères

Marc Santolini

To cite this version:

Marc Santolini. Analyse computationnelle des éléments cis-régulateurs dans les génomes des drosophiles et des mammifères. Analyse de données, Statistiques et Probabilités [physics.data-an]. Université Paris-Diderot - Paris VII, 2013. Français. NNT : . tel-00865159

HAL Id: tel-00865159

https://theses.hal.science/tel-00865159

Submitted on 24 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Département de Physique
École Normale Supérieure

Laboratoire de Physique Statistique

THÈSE de DOCTORAT de l'UNIVERSITÉ PARIS 7

Spécialité : Physique Théorique

présentée par

Marc SANTOLINI

pour obtenir le grade de DOCTEUR de l'UNIVERSITÉ PARIS 7

Analyse computationnelle des éléments cis-régulateurs dans les génomes des drosophiles et des mammifères

---

Thèse soutenue le 19 septembre 2013
devant le jury composé de :

M.	**Emmanuel BARILLOT**	Examinateur
M.	**Vincent HAKIM**	Directeur de thèse
M.	**Pascal MAIRE**	Membre invité
M.	**Massimo VERGASSOLA**	Rapporteur
M.	**Martin WEIGT**	Rapporteur
M.	**Alain ZIDER**	Président du Jury

Remerciements

Je tiens d'abord à remercier mon directeur de thèse, Vincent Hakim, qui m'a procuré un encadrement d'une très grande qualité au cours de ces quatre années passées au Laboratoire de Physique Statistique de l'École Normale Supérieure. Sa disponibilité, son discernement, sa curiosité, son esprit critique, l'étendue de sa culture scientifique dans les disciplines aux modes de réflexion et aux langages si différents que sont la physique fondamentale et la biologie, sa capacité à extraire d'un problème l'essence qui satisfait l'intuition et à chasser le superflu, sa démarche toujours rigoureuse vers le plus testable, le plus vérifiable, le plus dicible, mêlé à une grande humilité intellectuelle en font un chercheur exemplaire qui nourrit l'admiration, et je suis fier d'avoir pu apprendre à ses côtés.

Je tiens par ailleurs à remercier Pascal Maire qui m'a accueilli à l'Institut Cochin et m'a permis de réaliser une partie des expériences sur la différenciation musculaire exposées dans cette thèse. Pascal a réussi à m'introduire avec beaucoup de pédagogie à la génétique et au développement, sujets qui m'étaient initialement très étrangers. Son calme, sa patience, la clarté de ses raisonnements et son ouverture d'esprit ont grandement facilité cette expérience interdisciplinaire et je lui en suis gré.

J'ai eu la chance de réaliser plusieurs collaborations au cours de cette thèse, et par là de goûter à la satisfaction des travaux en groupe, l'apanage de l'interdisciplinarité. D'abord, je remercie beaucoup Hervé Rouault, qui finissait sa thèse lors de mon arrivée, et qui m'a énormément apporté sur les plans informatique (c'est peu dire), scientifique et humain. Ses connaissances encyclopédiques en programmation me furent essentielles, et je le remercie de sa disponibilité pour m'y avoir introduit. La collaboration avec Thierry Mora au LPS m'a permis de me plonger dans la physique des verres de spins et de garder un pied dans la physique pure et le langage qui lui est associé, dans une thèse autrement très portée sur la biologie. Je remercie Thierry pour sa rigueur et sa clarté, ainsi que pour sa capacité à savoir prendre du recul sur l'objet scientifique et le mettre en perspective. Sa connaissance d'un large corpus de philosophie des sciences en a fait un locuteur passionant. Je remercie par ailleurs Serge Plaza et François Payre du Centre de Biologie du Développement de l'Université Paul Sabatier pour les nombreuses et fructueuses interactions que nous avons eues. J'ai beaucoup bénéficié de leur dynamisme et les remercie de leur accueil très chaleureux lors de mes venues à Toulouse. Je remercie aussi Delphine Menoret qui finit sa thèse avec Serge et avec qui ce fut un plaisir de collaborer. Enfin, je remercie grandement Iori Sakakibara, post-doctorant à l'Institut Cochin, qui m'a accompagné dans ma découverte des méandres et de la complexité de la biologie expérimentale. Ses capacités de travail n'ont d'égal que l'étendue de la confiance qu'on peut lui accorder.

J'ai pu bénéficier au cours de cette thèse d'un cadre particulièrement agréable, que ce soit au LPS comme à l'Institut Cochin. Je remercie le directeur du LPS Eric Perez pour sa proximité et sa disponibilité, les secrétaires Marie Gefflot, Annie Ribaudeau, Alinh Rin-Tybenszky et Nora Sadaoui ainsi que les ingénieurs système Zaire Dissi, Rémy Portier et Frédéric Ayrault pour leur efficacité. Je suis aussi redevable à Maryse Derand, secrétaire de l'Institut Cochin, pour sa gentillesse et sa rapidité.

Ces années ont été financées en partie par un contrat doctoral avec l'université Paris Diderot, et par l'Association Française contre les Myopathies (AFM), et je les remercie de leur

confiance. Je remercie aussi l'Association DEPHY et notamment son président Édouard Brézin pour leur aide plus que bienvenue lors de la jonction cahoteuse des deux contrats.

Cette longue période de travail fut l'occasion de nouer des amitiés précieuses. Je remercie pour cela mes acolytes de la DC21, et en particulier Martin Retaux, Guilhem Sommeria-Klein et Léo Blondel avec qui j'ai passé des moments formidables. Leur diversité de culture et de points de vue mêlée à leur ouverture d'esprit en ont fait des compagnons idéaux, et nos discussions prolixes vont profondément me manquer. J'ai aussi pu bénéficier d'un entourage agréable à l'Institut Cochin, et je remercie entre autres Alice, Laura, Morgane, Fabien, Philippe, Christophe, Julien, Romain, Josiane et Maud.

Je remercie enfin mes parents pour les valeurs précieuses qu'ils m'ont inculquées, pour m'avoir laissé la plus grande liberté dans les choix qui furent les miens et m'avoir toujours soutenu quelque furent mes décisions. Je remercie aussi ma compagne Aya pour sa présence et son soutien tout au long de cette thèse.

Il est évident qu'il resterait une longue prose à écrire pour remercier tous ceux qui ne figurent pas ici, famille, amis, professeurs ou chercheurs, dont l'apport affectif ou intellectuel a tacitement participé à la réalisation de ce travail. Ils sauront que je les en remercie et les garde à l'esprit.

Table des matières

Liste des figures vii

Principales abbréviations utilisées ix

Avant-propos 1

Chapitre 1 - Introduction générale. 3

1.1 Le phénotype cellulaire 4
1.2 Les réseaux de régulation génétique 7
1.3 Les interactions protéine-ADN : modèles mathématiques 15
1.4 Les interactions protéine-ADN : mesures expérimentales 21
1.5 Les modules de cis-régulation (CRMs) 28
1.6 Prédiction et validation des CRMs 37
1.7 Bases de données 43

Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN. 49

2.1 Observations de corrélations au sein des TFBS 50
2.2 Modèles existants permettant de décrire la statistique des TFBS 51
2.3 Modèles de maximum d'entropie 54
2.4 Article 58
2.5 Analyse thermodynamique des modèles 90
2.6 Conclusion et perspectives du chapitre 2 91

Chapitre 3 - *Imogene* : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle 95

3.1 Quelques approches existantes pour la recherche de motifs et de modules de régulation 96
3.2 Article 105
3.3 Calcul de la moyenne de la postérieure par une méthode MCMC 132
3.4 Conclusion et perspectives du chapitre 3 139

Chapitre 4 - Étude de la différenciation épidermale chez la drosophile 141

4.1 Concept d'optimum de Pareto 143
4.2 Article 144
4.3 Conclusion et perspectives du chapitre 4 215

Chapitre 5 - Étude de la différenciation musculaire chez la souris **217**

5.1 Introduction à la myogenèse squelettique 219

5.2 Intégration de données bioinformatiques et expérimentales sur le muscle 223

5.3 Prédictions et validations de la régulation par Six 229

5.4 Synergie entre Six et MyoD au cours de la myogenèse 265

5.5 Conclusion et perspectives du chapitre 5 280

Annexe A - Statistiques génomiques **283**

Bibliographie **286**

Liste des figures

Introduction générale. 3
1.1 Le paysage de la différenciation cellulaire . . . 5
1.2 Spécification spatio-temporelle du type cellulaire . . . 6
1.3 Différents exemples de reprogrammation cellulaire . . . 7
1.4 Vision cybernétique du traitement de l'information par la cellule . . . 8
1.5 Un réseau de régulation génétique type . . . 9
1.6 Caractéristiques de l'épigénome . . . 10
1.7 Exemples de motifs dans les réseaux de régulation génétique . . . 12
1.8 Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique. . . 14
1.9 Différents états du facteur de transcription . . . 16
1.10 Construction et utilisation du modèle PWM . . . 17
1.11 Étapes d'une expérience de ChIP-on-chip et ChIP-seq . . . 25
1.12 Résolution des expériences ChIP-on-chip et ChIP-seq . . . 26
1.13 Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près . . . 27
1.14 Les différents types de CRMs et leurs marques épigénétiques . . . 29
1.15 Différents *enhancers* conduisent à différents patterns d'expression . . . 30
1.16 Deux modèles d'enhancers : enhanceosome et billboard . . . 32
1.17 L'enhanceosome de l'interferon-β . . . 33
1.18 Flexibilité du code de cis-régulation au cours de l'évolution chez les *Drosophiles* 34
1.19 Évolution de la fixation de HNF4α chez les mammifères . . . 35
1.20 « Shadow enhancer » du gène de segmentation *Hunchback* . . . 36
1.21 De l'enhancer au super-enhancer . . . 37
1.22 Différentes approches pour la prédiction des CRMs . . . 38
1.23 Méthodes de validation des CRMs par transfection et transgenèse . . . 42
1.24 Impact physiologique de la délétion et de la mutation d'un enhancer . . . 43
1.25 Évolution du coût de séquençage . . . 44
1.26 Visualisation de données ChIP-seq *via* le site UCSC . . . 46
1.27 Les différentes données obtenues par le projet ENCODE . . . 47

Modèles de fixation des Facteurs de Transcription à l'ADN. 49
2.1 Différents modèles pour décrire les corrélations entre nucléotides dans les sites de fixation de facteurs de transcription . . . 51
2.2 Illustration d'un système dont on veut maximiser l'entropie . . . 55
2.3 Chaleur spécifique pour différents TFs . . . 92
2.4 Histogrammes des valeurs des champs h et couplages J . . . 93

***Imogene* : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle** 95
3.1 Illustration de l'approche Espérance-Maximisation . . . 97
3.2 Conditions initiales . . . 135

3.3 Corrélations entre les échantillons 136
3.4 Estimation de la convergence 137
3.5 Comparaison des approches par MCMC et par descente de gradient 138

Étude de la différenciation épidermale chez la drosophile **141**
4.1 Frontière d'efficacité de Pareto 144

Étude de la différenciation musculaire chez la souris **217**
5.1 Formation du muscle squelettique chez l'embryon de souris 220
5.2 Motifs d'expression et localisation génomique des gènes de la famille Six 221
5.3 Rôle des protéines Six au cours de la myogenèse 222
5.4 Courbe ROC de MEF3 225
5.5 PWM MEF3 obtenue avec le modèle d'évolution 226
5.6 Visualisation de données musculaires sur UCSC 227
5.7 Prédiction de sites MEF3 sur les séquences régulatrices CE et DRR de *Myod1* . . 230
5.8 Validation des sites MEF3 prédits dans les éléments régulateurs de *Myod1* 231
5.9 Visualisation du locus des myosines sur UCSC 232
5.10 Description de l'expérience permettant d'étudier la synergie Six+MyoD 266
5.11 Données d'expression des gènes activés par la synergie Six+MyoD 267
5.12 Définition d'enhancers putatifs MyoD+MEF3 269
5.13 Rôle des enhancers putatifs MyoD+MEF3 dans la synergie Six+MyoD 270
5.14 Liste finale de gènes synergiques associés à un enhancer MyoD+MEF3 271
5.15 Illustration de la validation des enhancers par test Luciferase 272
5.16 Enhancers MyoD+MEF3 « positifs » 274
5.17 Enhancers MyoD+MEF3 « négatifs » 275
5.18 Motifs sur-représentés sur les enhancers synergiques 276
5.19 Mutagenèse systématique des sites prédits sur l'enhancer de *Srl* 277
5.20 Variation de la quantité de vecteur MyoD électroporé 278
5.21 Expression Luciférase de multimères synthétiques de sites de fixation 279

Statistiques génomiques **283**
A.1 Distribution des tailles intergéniques et introniques chez différentes espèces . . 284

Principales abbréviations utilisées

ARNm	ARN messager
bHLH	*basic Helix-Loop-Helix*
bp	Paire de base
ChIP	Immunoprécipitation de la chromatine (*Chromatin immunoprecipitation*)
CRM	Module de cis-régulation (*Cis-Regulatory Module*)
DHS	Hypersensible à la DNase I (*DNaseI-hypersensistive*)
ESC	Cellule souche embryonnaire (*Embryonic Stem Cell*)
ISH	Hybridation *in situ* (*In-Situ* Hybridization)
kb	kilobases (1000bp)
MRF	Facteur de régulation myogénique (*Myogenic Regulatory Factor*)
nt	Nucléotide
PCR	Réaction en chaîne par polymérase (*Polymerase Chain Reaction*)
PWM	Matrice de poids (*Position Weight Matrix*)
TF	Facteur de transcription (*Transcription Factor*)
TFBS	Site de fixation d'un facteur de transcription (*Transcription Factor Binding Site*)
TSS	Site d'initiation de la transcription (*Transcription Start Site*)

Avant-propos

Cette thèse se décompose en cinq chapitres. Le premier chapitre consiste en une introduction générale à la régulation transcriptionnelle chez les organismes eucaryotes, avec un intérêt particulier pour les drosophiles et les mammifères. Les deux chapitres suivants traitent de la modélisation mathématique de la régulation transcriptionnelle. D'abord, le chapitre 2 utilise des données *in vivo* extensives de fixation de facteurs de transcription sur l'ADN et compare la capacité de différents modèles à les décrire correctement. Puis le chapitre 3 présente Imogene, un algorithme de prédiction de motifs et de modules de régulation à partir de séquences d'ADN possédant une fonction de régulation similaire. Cet algorithme est une extension au cas des mammifères d'un algorithme développé précédemment par Hervé Rouault dans le cas des Drosophiles. Nous en présentons par ailleurs une utilisation nouvelle comme classifieur de séquences d'ADN conduisant différents motifs d'expression. Le chapitre 4 présente une application directe de cet algorithme au cas de la différenciation des trichomes chez la Drosophile. La validation des prédictions a nécessité de nombreuses expériences dont les mérites reviennent à Delphine Menoret de l'équipe de Serge Plaza à l'Université Paul Sabatier de Toulouse. Enfin, le chapitre 5 présente diverses prédictions et validations expérimentales concernant la différenciation musculaire chez la souris. La partie expérimentale a été en grande partie l'œuvre de Iori Sakakibara de l'équipe de Pascal Maire à l'Institut Cochin, et j'ai aussi pu réaliser certaines des expériences (qPCR, tests Luciférase, mutagenèses...) sous sa direction ainsi que celle de Pascal Maire.

Les chapitres 2, 3, 4 et 5 sont organisés autour d'articles en anglais décrivant le travail réalisé. Afin de faciliter leur lecture, les concepts sous-jacents et les travaux antérieurs sont introduits en début de chapitre.

Enfin, la version pdf de ce manuscrit contient de nombreux liens hypertexte rendant la lecture plus aisée. Dans la table des matières et la liste des figures, les titres pointent vers l'item décrit. Dans le corps du texte, les travaux cités pointent vers la référence correspondante dans la bibliographie, et il est possible de revenir au texte en cliquant sur le numéro de page côtoyant la référence.

Chapitre 1

Introduction générale.

1.1 Le phénotype cellulaire **4**
1.1.1 Qu'est-ce que le phénotype d'une cellule ? 4
1.1.2 La différenciation cellulaire 4
1.1.3 La cellule dans l'organisme : une spécification spatio-temporelle 6
1.1.4 La reprogrammation cellulaire 6
1.2 Les réseaux de régulation génétique **7**
1.2.1 Vision cybernétique de la cellule 8
1.2.2 Divers modes de régulation 8
1.2.3 Câblage du réseau et fonction 13
1.2.4 Évolution des réseaux génétiques 13
1.3 Les interactions protéine-ADN : modèles mathématiques **15**
1.3.1 Modes de recherche du site de fixation par le TF 15
1.3.2 Modèle PWM 16
1.3.3 Modèle biophysique 18
1.3.4 Modèle thermodynamique 19
1.4 Les interactions protéine-ADN : mesures expérimentales **21**
1.4.1 Approches *in vitro* : MITOMI, SPR, PBM, CSI, SELEX, et HT-SELEX 21
1.4.2 Approche clonale : la technique de simple hybride 23
1.4.3 Approches *in vivo* : ChIP-on-chip, ChIP-seq, DNAse I 24
1.5 Les modules de cis-régulation (CRMs) **28**
1.5.1 Les différents types de CRMs 28
1.5.2 Grammaire des enhancers : enhanceosome vs billboard 31
1.5.3 Évolution des enhancers 33
1.5.4 Les « shadow enhancers » 36
1.5.5 Par delà les enhancers : les « super-enhancers » 36
1.6 Prédiction et validation des CRMs **37**
1.6.1 Méthodes utilisant la concentration en sites de fixation 37
1.6.2 Méthodes utilisant la phylogénie 39
1.6.3 Méthodes utilisant les marques épigénétiques et de ChIP-seq pour des TFs 41
1.6.4 Validation expérimentale 41
1.6.5 Implication des CRMs dans les maladies humaines 42
1.7 Bases de données **43**
1.7.1 Obtention de données génomiques 44
1.7.2 Obtention de données sur les TFs 45
1.7.3 Outils de visualisation 45
1.7.4 Le projet ENCODE 47

1.1 Le phénotype cellulaire

1.1.1 Qu'est-ce que le phénotype d'une cellule ?

Les organismes vivants sont constitués de cellules de l'ordre de quelques microns, facilement observables à l'aide d'un simple microscope optique. Chaque cellule contient un certain nombre de constituants (gènes, protéines, métabolites...) enclos par une membrane. Il existe des organismes unicellulaires (bactérie, levure) et multicellulaires (mouche, souris, homme). Ce sont ces derniers auxquels nous nous intéressons dans cette thèse. Les cellules qui les constituent sont majoritairement eucaryotes, c'est-à-dire qu'elles possèdent un noyau renfermant le matériel génétique. [1]

Bien que possédant toutes le même matériel génétique (à quelques variations près), les cellules d'un organisme apparaissent d'emblée comme hétérogènes, que ce soit dans leur forme ou dans leurs constituants. Par exemple, chez l'homme, les neurones sont composés d'un corps d'une dizaine de microns de diamètre contenant le noyau, de milliers de prolongements dont certains peuvent atteindre une taille de plus de 1.5 mètre, et sont riches en ions sodium et potassium, tandis que les fibres musculaires squelettiques sont de forme longue et tubulaire, possèdent plusieurs dizaines de noyaux et expriment actine et myosine.

Cette diversité semble néanmoins limitée. Aussi, parmi les $\sim 6\ 10^{13}$ cellules du corps humain, on peut distinguer ~ 320 différents types cellulaires (Brazma et al., 2001). Bien entendu, ce nombre dépend du seuil de similarité choisi : deux cellules d'un même type n'expriment pas *exactement* le même nombre de molécules. Classiquement, la classification d'un type cellulaire se base sur des propriétés morphologiques observables au microscope ou encore sur l'analyse de molécules présentes à la surface des cellules. Par ailleurs, différents types cellulaires sont associés à différentes fonctions : dans notre exemple, la transmission du signal nerveux dans le cas des neurones, la contraction dans le cas des fibres musculaires.

Ces différentes propriétés observables caractérisent le *phénotype* cellulaire (étymologiquement « exhiber un type » en grec). Nous allons le voir, ce phénotype est le résultat de la modulation par des facteurs environnementaux de l'expression génétique qui conditionne le contenu en protéines de la cellule.

1.1.2 La différenciation cellulaire

L'acquisition d'un phénotype cellulaire particulier au sein d'un organisme est le sujet de la biologie du développement. Cette acquisition passe par différentes étapes de différenciation cellulaire. Schématiquement, au cours du développement d'un organisme, des cellules non différenciées ou souches empruntent un chemin unidirectionnel de différenciation qui restreint peu à peu le nombre de types cellulaires qu'elles peuvent potentiellement devenir, passant de l'état souche totipotent à des états pluripotents successifs avant la différenciation finale. Ainsi, la formation des cellules somatiques, qui sont les cellules d'un organisme n'étant ni souches ni germinales (les cellules qui donnent naissance aux gamètes ou cellules sexuelles), est le résultat d'un processus de différenciation initial lors du développement embryonnaire au cours duquel les cellules issues de l'œuf donnent naissance à trois couches de tissus distinctes : l'endoderme (feuillet interne), l'ectoderme (feuillet externe) et le mésoderme (feuillet intermédiaire). Des différenciations successives ont ensuite lieu au sein de ces couches pour

1. Il existe cependant quelques cas connus d'organismes multicellulaires procaryotes, dont les cellules ne possèdent pas de noyau, par exemple chez les bactéries magnétotactiques (Keim et al., 2004).

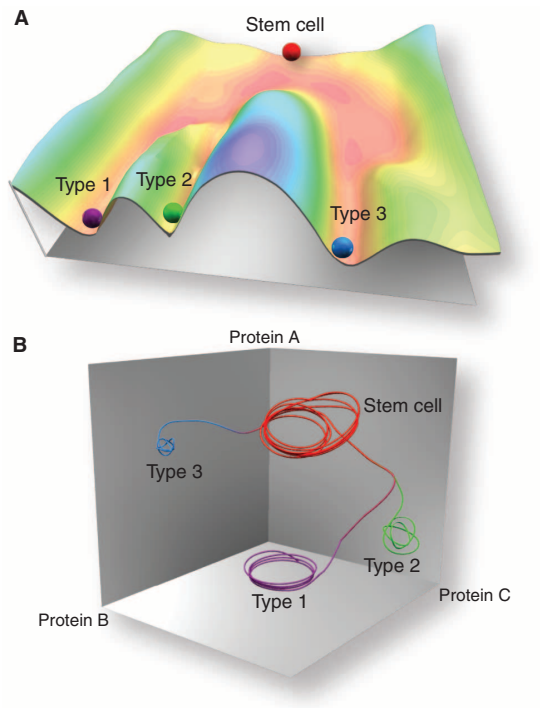


FIGURE 1.1 – **Le paysage de la différenciation cellulaire.**
Figure tirée de (Furusawa and Kaneko, 2012). **A.** Paysage épigénétique tel qu'imaginé par Waddington (Waddington et al., 1957) en résonance avec la notion de paysage énergétique en physique. Le développement cellulaire est représenté par une bille dévalant un paysage composé de différentes vallées séparées par des barrières difficilement franchissables, représentant les différents types cellulaires et leur robustesse face aux fluctuations. **B.** Représentation dynamique de l'évolution des états cellulaires. Le phénotype est ici caractérisé par l'expression de trois protéines A, B et C, dont l'évolution dans le temps peut être représentée par une trajectoire dans un espace tridimensionnel. Les états souche et différenciés sont caractérisés par des bassins d'attraction correspondant à différents types cellulaires.

former divers organes tels que le tube digestif (endoderme), les muscles et les os (mésoderme), ou encore la peau et le système nerveux (ectoderme).

Dans un écrit aujourd'hui célèbre datant de 1957 (Waddington et al., 1957), Waddington proposa une représentation de ces différentes étapes sous la forme d'un paysage épigénétique semblable aux paysages énergétiques dont sont coutumiers les physiciens (fig 1.1A). Dans cette représentation, le processus de différenciation cellulaire est comparé à une bille dévalant une pente et dont la trajectoire suit les multiples embranchements de vallées escarpées, chacune représentant un état de développement différent. Les vallées sont séparées par des pics dont la hauteur reflète la difficulté de passer d'un état à un autre, et les destinations finales possibles de la bille correspondent aux différents types cellulaires.

La notion de trajectoire de différenciation peut être rendue plus parlante en adoptant une représentation de système dynamique. Comme nous l'avons vu en 1.1.1, la cellule contient de nombreux composants : gènes, protéines ou autres métabolites, qui pris dans leur ensemble déterminent à un instant donné l'état cellulaire. Il est ainsi possible de représenter l'état cellulaire à un temps donné comme un point dans un espace de grande dimension dans lequel chaque axe représente l'abondance d'un certain composant (fig 1.1B). De par leur rôle primor-

FIGURE 1.2 – **Spécification spatio-temporelle du type cellulaire.**
Hybridation *in situ* de l'ARN du gène *Myog*, marqueur de la différenciation des progéniteurs du muscle squelettique, chez des embryons de souris âgés de 9.5, 10.5 et 11.5 jours (de gauche à droite), observés sous un même grossissement de 10. Le motif (*pattern*) de spécification du muscle squelettique est clairement visible au niveau des somites, lieu des futures vertèbres. Images tirées de la base de donnée Embrys (`http://embrys.jp`).

dial dans la définition de l'état cellulaire, l'expression des protéines (et donc des gènes qui les produisent) domine généralement ces composants, et on parle de « niveau d'expression génétique » pour décrire leur abondance. Les changements d'expression génétique, au cours desquels certains gènes vont être activés et d'autres réprimés, induisent un changement de l'état cellulaire, ce qui se traduit par une trajectoire dans l'espace des états. Ces changements d'expression restreignent finalement l'état cellulaire à une certaine région, définie comme un « attracteur » de la dynamique. Une fois au sein d'un attracteur, l'état cellulaire est robuste aux perturbations du niveau d'expression génétique des différentes composantes. Les attracteurs peuvent alors être vu comme des types cellulaires distincts correspondant aux différentes vallées de la représentation de Waddington (Kaufmann, 1993).

1.1.3 La cellule dans l'organisme : une spécification spatio-temporelle

Au sein de l'organisme, la différenciation cellulaire opère à un rythme précis et dans un contexte environnemental bien défini. Aussi, les trajectoires dans l'espace d'expression génétique que nous avons présentées précédemment sont fonction de l'espace – la position de la cellule dans l'organisme, qui détermine en particulier la concentration des signaux qu'elle reçoit de son environnement – et du temps – le stade de développement de l'organisme –. Il est ainsi possible d'observer chez l'embryon certains motifs ou *patterns* spatio-temporels d'expression génétique correspondant à des organes en formation et révélés par hybridation *in situ* de l'ARN de certains gènes spécifiques d'un type cellulaire. Par exemple, dans le cas de la formation des muscles squelettiques, le gène de différenciation terminale *Myog* est exprimé chez la souris dès 8 jours embryonnaires au niveau des somites, segments correspondant aux futures vertèbres de la souris adulte, puis commence à être exprimé au niveau des bourgeons de membres à 11.5 jours (voir fig 1.2).

1.1.4 La reprogrammation cellulaire

Depuis plusieurs décennies, différentes expériences ont exhibé la plasticité des états différenciés, élargissant ainsi considérablement la vision classique selon laquelle des cellules souches totipotentes se différencient de manière irréversible en des cellules de moins en moins plastiques, jusqu'à atteindre un état différencié stable. Par exemple, (Blau et al., 1985) ont mon-

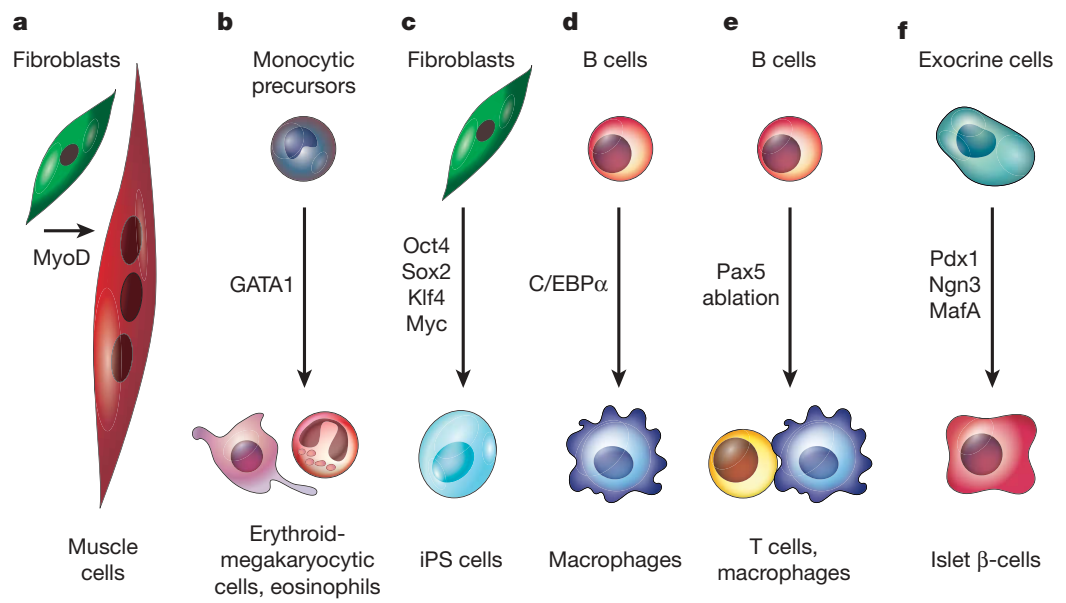


FIGURE 1.3 – **Différents exemples de reprogrammation cellulaire.**
Figure tirée de (Graf and Enver, 2009), montrant plusieurs exemples d'expériences de surexpression ou d'ablation de facteurs de transcription aboutissant à des changements de types cellulaires.

tré en 1985 que des programmes d'expression génétique dormants peuvent être exprimés de manière dominante dans des cellules différenciées par la fusion de différents types cellulaires : ainsi, la fusion de cellules musculaires avec des cellules non musculaires permet l'activation des gènes de type musculaire dans le type cellulaire non musculaire. Puis différents travaux ont montré qu'il était possible de convertir des lignées de cellules différenciées en un autre type cellulaire en introduisant certaines protéines régulatrices de la transcription, ou Facteurs de Transcription (TFs) (Davis et al., 1987; Kulessa et al., 1995) : on parle alors de transdifférenciation, dont l'un des exemples canoniques est la différenciation de cellules de la peau ou fibroblastes en cellules musculaires par l'introduction du facteur de différenciation myogénique MyoD (voir fig 1.3). Parallèlement, des expériences réalisées chez plusieurs espèces de mammifères ont montré que le transfert de noyaux de cellules différenciées embryonnaires ou adultes dans un oeuf énucléé peut mener à la formation d'un organisme complet, montrant de manière univoque que l'identité des cellules différenciées peut être complètement renversée (Gurdon and Melton, 2008). Enfin, l'avancée la plus récente dans ce domaine a été la démonstration que des cellules somatiques différenciées peuvent être reprogrammées en cellules souches puripotentes par simple introduction d'un « cocktail » de 4 facteurs de transcription : Oct4, Sox2, c-Myc et Klf4 (Takahashi and Yamanaka, 2006) (fig 1.3C).

1.2 Les réseaux de régulation génétique

Afin de pouvoir mieux comprendre les mécanismes de différenciation et de reprogrammation exposés en 1.1, il convient de se plonger dans les mécanismes internes de la cellule qui régissent ses changements d'états.

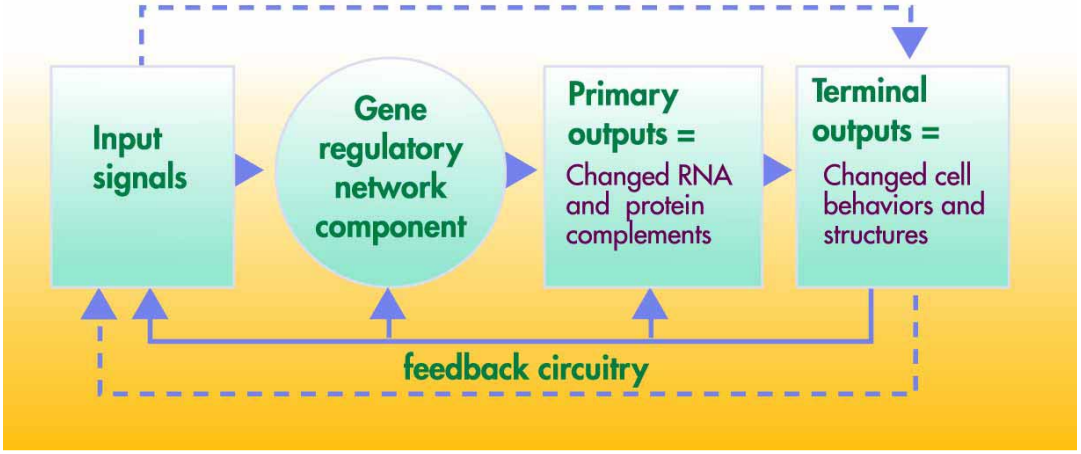


FIGURE 1.4 – **Vision cybernétique du traitement de l'information par la cellule.** Figure tirée du programme 'Genomes to life' du Département de l'Énergie des États-Unis datant de 2001 (U.S. Department of Energy, 2001) schématisant un réseau de régulation cellulaire comme un système de traitement entrée/sortie, possédant trois composantes fondamentales : (1) un système de réception et de transduction des signaux d'entrées qui peuvent être intra- ou extra-cellulaires (plusieurs signaux pouvant affecter un même gène cible), (2) un « composant central » (*core component*) composé du réseau de régulation génétique traitant les signaux, et (3) d'un signal de sortie consistant en l'expression moléculaire des ARNs et protéines des gènes cibles. Le processus résulte en la modification du phénotype de la cellule. Des boucles de régulation (*feedback*) assurent le contrôle et la stabilité des différentes étapes.

1.2.1 Vision cybernétique de la cellule

Le paradigme qui règne sur la biologie moléculaire depuis plus d'un demi siècle est celui des réseaux génétiques. L'expression des gènes est en effet régulée par des protéines, les facteurs de transcription, qui sont eux-mêmes issus de l'expression d'autres gènes, créant ainsi un réseau d'interactions entre gènes. Certaines protéines peuvent par ailleurs directement réguler l'activité d'autres protéines, et certains ARNs issus de la transcription de gènes non codants jouent aussi un rôle fondamental dans la régulation de l'activité génétique, le tout formant un réseau complexe d'interactions à différents niveaux. La compréhension de ce réseau et des fonctions qui en résultent forme le socle de la biologie des systèmes. Dans ce cadre, la cellule est vue comme une unité de traitement d'information qui interprète différents signaux reçus en entrée, les traite par un réseau interne de régulation, et réagit en sortie en modifiant son état ou son comportement (fig 1.4). L'intérêt d'une telle description mécanistique est qu'elle permet d'opérer quantifications mathématiques et prédictions, ce qui l'a rendue extrêmement fertile au cours des dernières décennies (Nurse and Hayles, 2011).

1.2.2 Divers modes de régulation

Les modes de régulation qui permettent à la cellule d'interpréter des signaux afin de changer d'état sont nombreux. Nous allons nous concentrer ici sur ceux affectant la production d'ARNs ou de protéines (fig. 1.5).

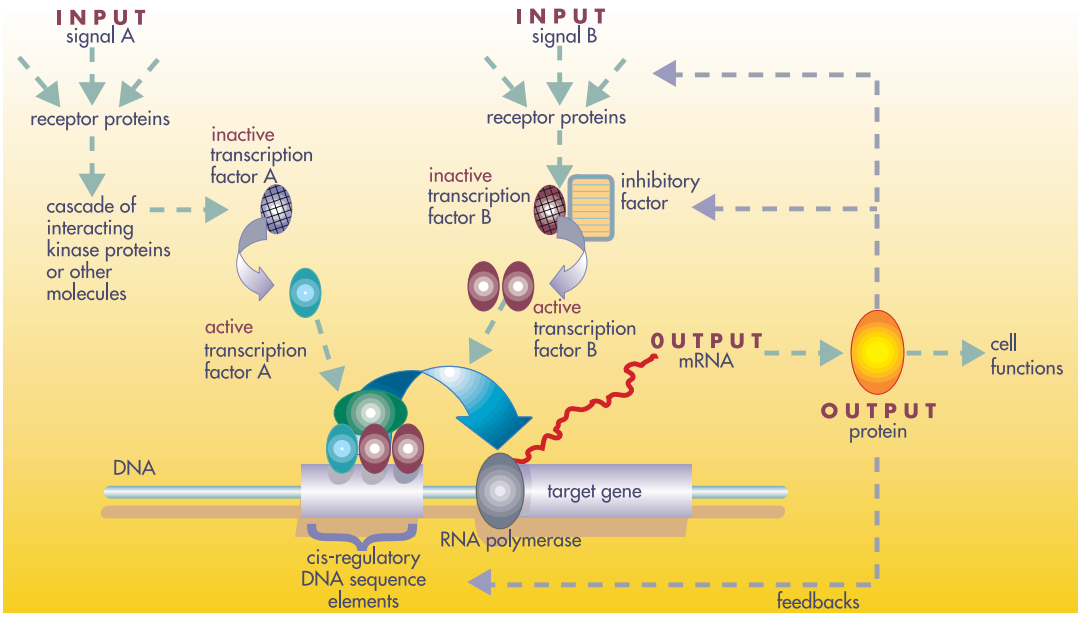


FIGURE 1.5 – **Un réseau de régulation génétique type.**
Dans cette représentation schématique tirée du rapport du U.S. Department of Energy (2001), deux voies de signalisation A et B transmettent des signaux d'entrée (qui peuvent être intra ou extra cellulaires) en rendant des facteurs de transcription actifs. Une fois activés, ces derniers interagissent avec des séquences d'ADN proches d'un gène cible en se fixant sur des sites de petite taille ($\sim$ 10bp). Les différents facteurs de transcription interagissent entre eux pour former des complexes occupant des régions de $\sim$ 1000bp appelées modules de cis-régulation ou CRMs (voir section 1.5). Lorsque les facteurs de transcription sont fixés sur le CRM de leur gène cible, il peuvent activer ou inhiber la transcription d'ARN et donc la production de la protéine correspondante.

- **Régulation génétique**

Tout d'abord, un réseau d'expression génétique est caractérisé par un jeu d'interactions entre différents gènes. Ici, nous centrons notre attention sur les gènes codant pour des protéines, mais on pourrait inclure les gènes codant pour des ARNs non traduits, ceux-ci pouvant être impliqués dans la régulation. Dans ce réseau, les interactions se font par l'intermédiaire de protéines régulatrices appelées facteurs de transcription ou TFs, qui sont au nombre de $\sim$ 1400 chez l'homme (Vaquerizas et al., 2009), soit 6% des protéines encodées. Les gènes qui les expriment représentent donc $\sim$ 3% de l'ensemble des 30,000 gènes connus à ce jour. Pour réguler (activer ou inhiber) la transcription d'un gène cible, les TFs se fixent sur des sites de reconnaissance spécifiques sur l'ADN de $\sim$ 10bp et interagissent avec la machinerie transcriptionnelle au niveau du promoteur du gène cible. Les TFs peuvent se fixer sur le promoteur même, comme c'est souvent le cas chez la bactérie, ou dans des régions distales allant jusqu'à plusieurs centaines de kb, comme on trouve plus couramment chez les organismes complexes. Par ailleurs, différents TFs peuvent se combiner sur certaines régions de régulation contenant de multiples sites de fixation pour former des complexes protéiques. Ces régions, appelées modules de cis-régulation (CRMs) ou plus communément *enhancers*, sont d'une taille typique de $\sim$ 1000bp et ont la particularité de conduire à une expression spatio-temporelle très spécifique du gène cible. Ces différents points seront amplement développés en section 1.5.

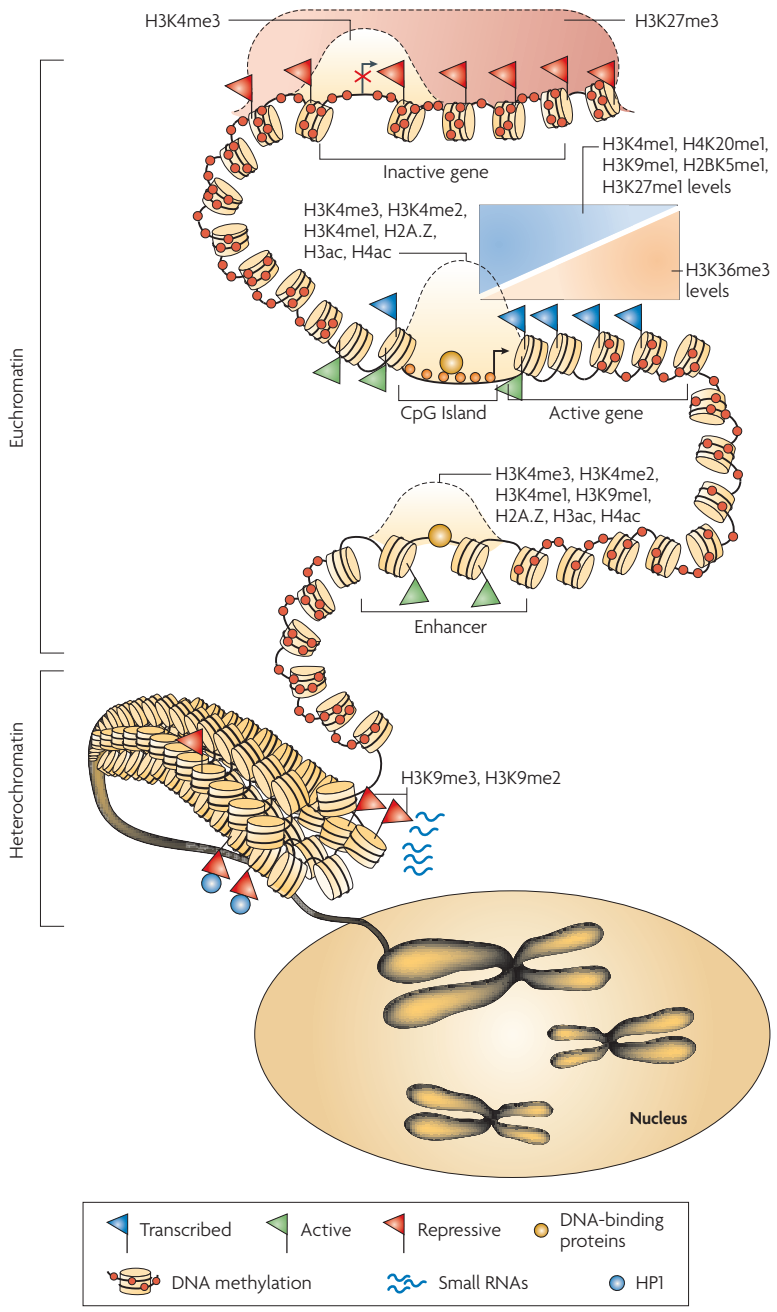


FIGURE 1.6 – **Caractéristiques de l'épigénome.**
Figure tirée de (Schones and Zhao, 2008). Les chromosomes sont partagés entre régions accessibles d'euchromatine et régions difficilement accessibles d'hétérochromatine. Les régions hétérochromatiques sont marquées par la di- et triméthylation de la lysine 9 de l'histone H3 (H3K9me2 et H3K9me3). La méthylation de l'ADN est répandue à travers tout le génome, mais est absente de certaines régions comme les îlots CpG, les promoteurs et les CRMs. La modification H3K27me3 couvre de larges régions englobant des gènes inactifs. Les marques H3K4me3, H3K4me2, H3K4me1 et l'acétylation des histones marquent les TSSs des gènes actifs. Les marques H3K4, H3K9, H3K27, H4K20 et H2BK5 marquent les régions transcrites activement à proximité de la région 5′ des gènes (en amont), alors que la marque H3K36 marque les gènes transcrits dans leur région 3′ (en aval).

- **Régulation épigénétique**

Outre la régulation génétique, due à l'action de protéines issues de séquences codantes et se fixant sur des séquences d'ADN – régulation qui est donc entièrement encodée dans le génome et transmise à la descendance –, il existe un autre mode de régulation de la transcription des gènes qui permet notamment d'acquérir une modification d'expression génétique transmise à la descendance sans qu'il y ait modification du code génétique : c'est la régulation épigénétique. Cette régulation passe notamment par la modification des propriétés chimiques de l'ADN et des histones sur lesquelles il s'enroule pour former la chromatine (fig. 1.6) [2]. Ainsi, la méthylation des dimères CpG de l'ADN [3] au niveau des régions riches en CG, ou îlots CpG, situées près de nombreux promoteurs et habituellement dépourvues de ces marques conduit à une inactivation du gène cible (Bird, 2002). Par ailleurs, la méthylation des histones au niveau des résidus lysines entraîne la fermeture de la chromatine, empêchant l'expression du ou des gène(s) situés à leur niveau, alors que l'acétylation des mêmes lysines entraîne au contraire une ouverture de la chromatine, favorisant ainsi la transcription génétique (Greer and Shi, 2012). Ce mode de régulation sera développé plus en détail en section 1.5.1.

- **Régulation post-transcriptionnelle**

Les modifications post-transcriptionnelles affectent les ARNs issus de la transcription des gènes. Ces modifications peuvent être causées par des microARNs ou miRNAs qui sont des ARNs de $\sim$ 23 nts issus d'ARNs se repliant en structure double brin de type « épingles à cheveux » ou *hairpins*. Les miRNAs s'associent à la protéine *Argonaute* du complexe RISC (*RNA-induced silencing complex*) pour entraîner la dégradation spécifique d'ARNms (Bartel, 2009). De manière similaire, certains *hairpins* de taille plus importante sont clivés par la protéine Dicer pour former plusieurs petits ARNs de taille similaire aux miRNAs : ce sont les siRNAs (*small interfering RNAs)*). Ceux-ci recrutent aussi le complexe protéique RISC et ciblent spécifiquement des ARNm (Hammond et al., 2001; Hannon, 2002). Ce phénomène est connu sous le nom d'interférence ARN (RNAi) et a donné lieu à une méthode aujourd'hui couramment utilisée pour inhiber l'expression d'un gène.

- **Régulation post-traductionnelle**

Les modifications post-traductionnelles affectent les protéines issues de la traduction des ARNs. Elles passent par une modification chimique des protéines, typiquement la phosphorylation, ou comme nous l'avons vu pour la régulation épigénétique, la méthylation ou l'acétylation. Ces modifications peuvent avoir pour effet de changer l'activité de la protéine, que ce soit en modifiant son activité enzymatique ou en déclenchant sa relocalisation nucléaire. Il existe aussi des modifications de structure de la protéine, comme c'est le cas du facteur de transcription *Shavenbaby* chez la Drosophile : dans sa forme native, cette protéine inhibe la transcription de ses gènes cibles ; cependant ses résidus terminaux peuvent être clivés par des petits peptides de 11 à 32 acides aminés encodés par le gène *Pri*, rendant la protéine transcriptionnellement active (Kondo et al., 2010).

2. Il est à noter que certains emploient le terme épigénétique pour qualifier la fixation des TFs sur l'ADN. Ici, le terme épigénétique réfère seulement aux modifications chimiques affectant les histones et l'ADN, et donc l'accessibilité du génome

3. Les dimères C-G sont appelés CpG, où p caractérise le phosphore liant les deux bases, pour les différencier du CG utilisé pour parler de la statistique en C et G de l'ADN

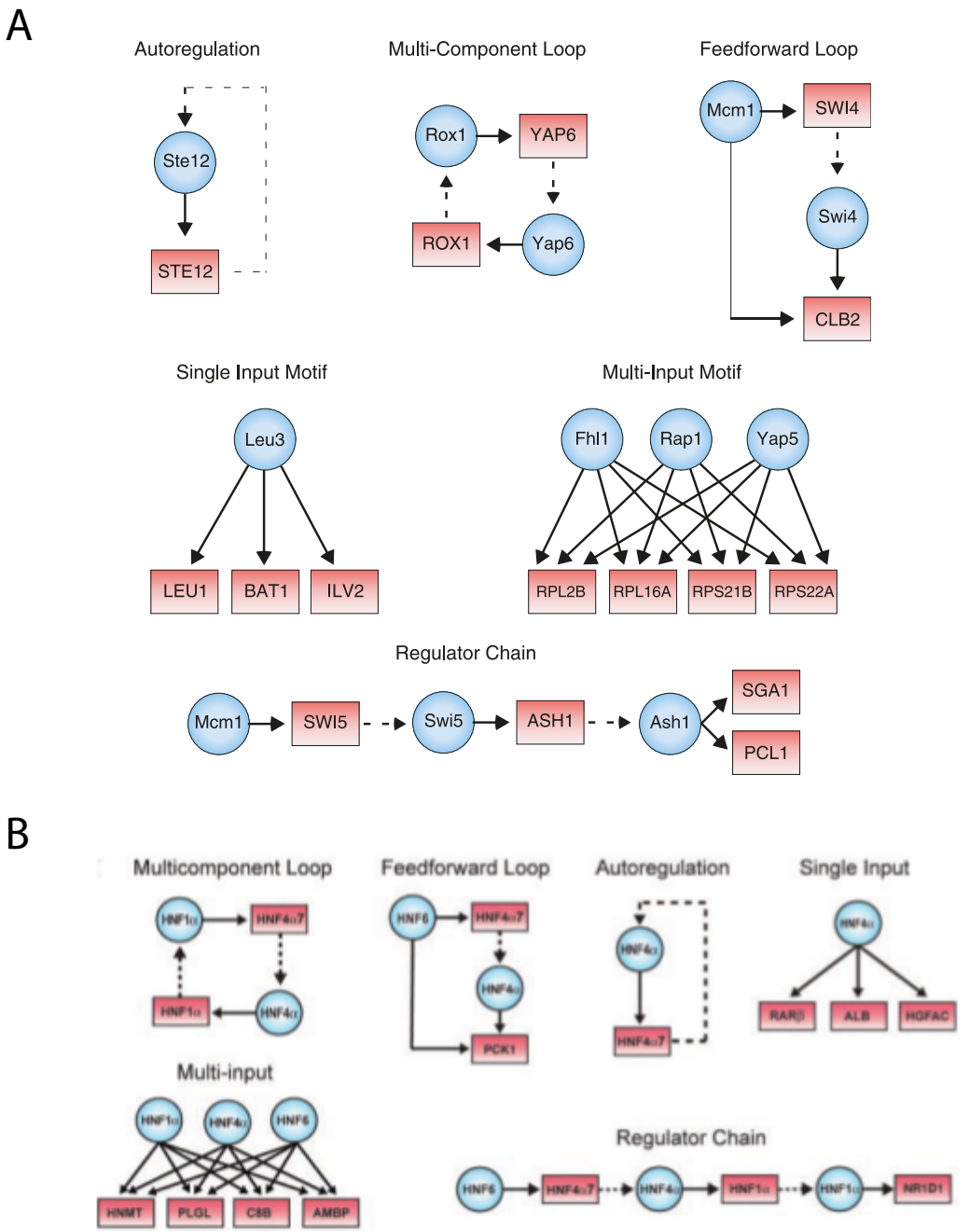


FIGURE 1.7 – **Exemples de motifs dans les réseaux de régulation génétique.**
Ces exemples sont issus d'analyses d'interactions entre facteurs de transcription (cercles bleus) et promoteurs (rectangles rouges) (A) chez la levure (Lee et al., 2002) et (B) chez l'homme (Odom et al., 2004). Les flèches solides indiquent la fixation d'un facteur de transcription à un promoteur, et les flèches en pointillé indiquent l'expression d'un facteur de transcription à partir de son gène.

1.2.3 Câblage du réseau et fonction

Maintenant que nous avons vu la nature des interactions au sein des réseaux génétiques, nous pouvons nous pencher sur leur structure. Notamment, plusieurs études réalisées chez divers organismes de la bactérie à l'homme ont révélé que les réseaux de transcription contiennent un petit ensemble de motifs de régulation récurrents, appelés motifs de réseaux (Alon, 2007a; Shen-Orr et al., 2002; Milo et al., 2002) (fig. 1.7). De tels motifs furent d'abord détectés de manière systématique chez la bactérie *Escherichia coli* en remarquant qu'ils apparaissaient dans le réseau de transcription bien plus souvent qu'on ne l'attendrait dans un réseau aléatoire (Shen-Orr et al., 2002). Les mêmes motifs ont ensuite été trouvés chez la levure (Milo et al., 2002; Lee et al., 2002) et chez l'homme (Odom et al., 2004). Une explication possible de la récurrence de ces motifs est liée aux fonctions qu'ils remplissent. Par exemple, la boucle d'autorégulation négative, qui est trouvée chez la moitié des répresseurs d'*Escherichia coli*, possède deux fonctions : l'une est de parvenir rapidement à un état d'équilibre en utilisant un promoteur fort, l'autre est de servir de tampon au bruit d'expression (Alon, 2007b). Un autre motif récurrent est la boucle feedforward. Celle-ci consiste en 3 gènes : un régulateur X, qui régule Y, tous deux régulant Z. Dans le cas où des interactions sont des activations et que X et Y sont requis pour activer Z, cette boucle peut servir de tampon au bruit d'expression de X, évitant que des fluctuations de son niveau d'expression n'entraîne par erreur l'activation de Z.

1.2.4 Évolution des réseaux génétiques

L'importance des motifs est rendue plus claire lorsque l'on s'intéresse à l'évolution des réseaux. En effet, au cours de l'évolution, les réseaux de régulation génétique changent : modification des constituants, recâblage du réseau, duplication d'éléments... Néanmoins, certaines modifications sont plus défavorisées du point de vue évolutif que des autres. Ainsi, les motifs tels que les boucles d'autorégulation ou les boucles feedforward, de par leur importance fonctionnelle, auront tendance à être conservés. Pour ce qui est des éléments du réseau, la modification d'un régulateur, par exemple la mutation d'un acide aminé au sein d'un facteur de transcription, aura des conséquences sur l'ensemble des éléments régulés par ce facteur de transcription et pourra donc être fortement délétère. Par contre, la modification d'un site de reconnaissance de ce facteur de transcription sur l'ADN n'aura qu'une portée locale sur la régulation du gène associé.

À titre d'exemple, prenons le cas du réseau de différenciation du muscle squelettique présenté en figure 1.8, que nous étudierons plus en détail dans le chapitre 5 de ce manuscrit. Au coeur de ce réseau génétique se trouvent les facteurs de régulation myogéniques ou MRFs, des facteurs de transcription de type bHLH qui ont la capacité de convertir des cellules non mesodermales, c'est-à-dire n'étant pas destinées à devenir des progéniteurs musculaires, en cellules ayant des propriétés musculaires (Weintraub et al., 1989). Ces facteurs sont dits « régulateurs maîtres » de la différenciation musculaire. Chez les vertébrés il y a quatre MRFs : *Myf5*, *Mrf4*, *Myod1*, qui ont des roles redondant dans la spécification des progéniteurs musculaires, et *Myog*, qui conduit à la différenciation terminale. Chez la Drosophile c'est le TF *Twist* qui semble être le principal MRF, mais contrairement aux MRFs des vertébrés, son rôle ne s'arrête pas au contrôle de la différenciation musculaire mais est plus général dans le développement du mésoderme (Baylies et al., 1998). C'est cependant le gène *Nautilus* qui possède la séquence d'acides aminés la plus proche de celle des MRFs vertébrés. Ce dernier permet la spécification des progéniteurs myogéniques, et son expression est restreinte au développement musculaire.

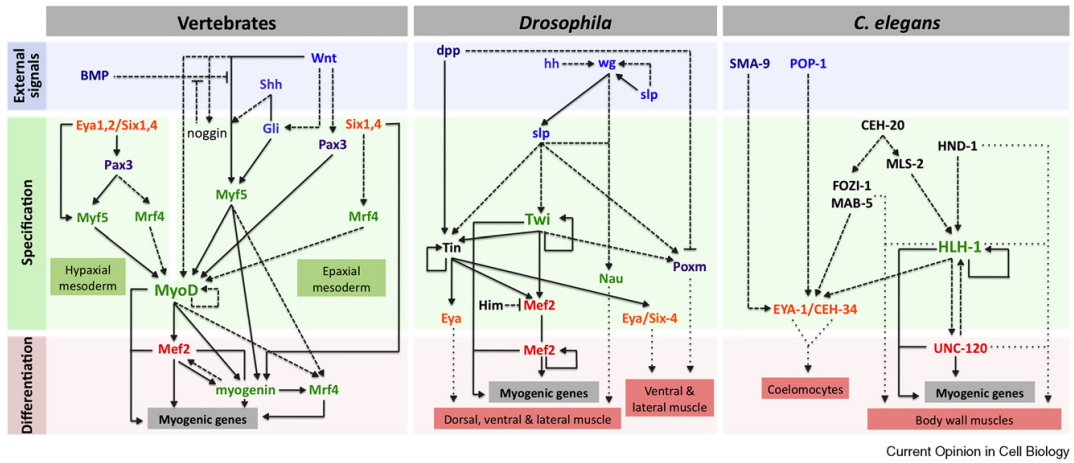


FIGURE 1.8 – **Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique..**
Figure tirée de (Ciglar and Furlong, 2009). Les lignes pleines indiquent une régulation transcriptionnelle directe, les tirets indiquent des interactions génétiques dont le caractère direct n'a pas été montré, et les lignes pointillés indiquent que l'interaction est suggérée par des donnés d'expression. *Myod1* est l'analogue structurel de *Nautilus* mais est l'analogue fonctionnel de *Twist* chez la Drosophile et de *HLH-1* chez *C. elegans*. Les couleurs indiquent respectivement des voies de signalisation externes au mésoderme (bleu), les régulateurs myogéniques de la famille bHLH (vert), les protéines Six/Eya (orange), les protéines de la famille *paired-box* Pax (violet), les TFs des familles MADS/SRF (rouge), et les régulateurs d'autres familles de protéines (noir).

Néanmoins, les mutants *nautilus* sont viables et son rôle semble mineur comparé aux MRFs vertébrés. Enfin, chez le ver *Caenorhabditis elegans*, c'est l'orthologue de *Myod1*, *hlh-1*, qui tient rôle de MRF.

Malgré ces différences (nombre de MRFs, membre de la famille bHLH tenant ce rôle), on retrouve dans les trois cas une boucle feedforward conservée au niveau de la régulation des cibles des MRFs (fig. 1.8). Ainsi, MyoD régule l'expression de Mef2 et l'activité de MAPK p38 en même temps que l'expression de plusieurs cibles initiales, et par la suite MyoD et phospho-Mef2 co-régulent des gènes plus tardifs. Ce mécanisme permet ainsi de réguler l'aspect temporel de l'expression génétique. Chez la Drosophile, le même motif est observé avec Twist et Mef2 et chez *C. elegans* avec HLH-1 et le TF UNC-129, de la même famille que Mef2.

Le coeur du réseau est donc conservé dans la forme (topologie), même s'il y a des divergences dans le fond (membres de la famille de TFs impliqués). Néanmoins, les éléments régulateurs en amont, ainsi que les membres périphériques du réseau ont rapidement évolué. Par exemple, chez les vertébrés le TF Pax3 est très en amont dans la hiérarchie génétique et permet l'activation des MRFs et la spécification myogénique, alors que chez la Drosophile son homologue *poxm* est en aval des MRFs et sa perte de fonction n'a que des effets mineurs sur la myogenèse. Par ailleurs, le complexe composé de protéine Six et de leur cofacteur Eya, initialement découvert comme régulateur majeur de la différenciation oculaire chez la Drosophile, est chez les vertébrés un régulateur essentiel situés en amont des MRFs. Chez la Drosophile,

il possède aussi un rôle dans la spécification myogénique, mais bien plus en aval que chez les vertébrés. Enfin, chez *C. elegans* ce complexe est aussi en aval des MRFs mais il participe en plus à la détermination de cellules non myogéniques.

Nous voyons donc que l'évolution d'un réseau génétique possède de multiples facettes : conservation de motifs de réseau fonctionnellement importants (dans notre exemple, la boucle feedforward au coeur du réseau régissant l'aspect temporel de l'expression des cibles), recâblage des interactions pour traiter différents signaux d'entrée... Par ailleurs, il apparaît que plus qu'à des TFs particuliers, c'est à des familles de TFs que nous avons affaire. Aussi un même rôle au sein du réseau peut-il être rempli par différents membres d'une même famille, comme c'est le cas pour *Myod1* et *Twist*. Ceci s'explique par le fait que les membres d'une même famille partagent des propriétés d'interaction avec l'ADN semblables. Ces interactions sont à la source du fonctionnement du réseau, et nous allons maintenant présenter plus en avant leurs propriétés.

1.3 Les interactions protéine-ADN : modèles mathématiques

Nous l'avons vu, les interactions entre facteurs de transcription et ADN sont une composante essentielle des réseaux génétiques. Les TFs se fixent sur des sites spécifiques de ~ 10 bp dans le voisinage des gènes qu'ils régulent. Trouver ces sites est donc un premier pas vers la reconstruction des réseaux de régulation sous-jacents. Dans cette section nous présentons les modèles d'interactions protéine-ADN qui ont été proposés, et leur application concrète à la recherche de sites de fixation.

1.3.1 Modes de recherche du site de fixation par le TF

Un facteur de transcription peut être dans plusieurs états : en diffusion tridimensionnelle, auquel cas il est dit « libre », ou bien fixé sur l'ADN. Dans ce dernier cas, il interagit avec l'ADN selon deux modes : une attraction non spécifique d'énergie E_{ns} indépendante de la position sur l'ADN, et une interaction spécifique $E_s(r)$ qui dépend de la séquence de taille $l \sim 10$ à la position r sur l'ADN. L'interaction non spécifique est due à l'interaction électrostatique entre la protéine chargée positivement et l'ADN chargé négativement, alors que l'interaction spécifique implique des liaisons hydrogènes entre le domaine de fixation de la protéine et les nucléotides du site de fixation. Le facteur de transcription peut ainsi se trouver dans trois états thermodynamiques représentés en figure 1.9 : en diffusion tridimensionnelle libre, fixé non spécifiquement (diffusion unidimensionnelle le long de la structure d'ADN), et fixé spécifiquement sur l'ADN. Ces trois modes contribuent à la cinétique de la recherche d'un site fonctionnel (Berg et al., 1981; Winter and von Hippel, 1981; Winter et al., 1981). Ainsi, l'attraction non spécifique conduit la protéine à passer à peu près autant de temps fixé sur l'ADN qu'en diffusion libre. La recherche de site de reconnaissance est donc un processus mixte de diffusion unidimensionnelle sur l'ADN et de diffusion tridimensionnelle dans le milieu. Lorsqu'il est fixé sur l'ADN, le facteur diffuse dans un paysage d'énergie E_{ns} plat lorsqu'il est dans sa conformation de fixation non spécifique, ou dans un paysage d'énergie $E_s(r)$ dans sa conformation de fixation spécifique. Cela permet au facteur d'échantillonner les sites de faible énergie $E_s(r)$ tout en évitant d'être bloqué par les barrières de haute énergie en passant en mode de recherche non spécifique. Ce processus s'avère au final très efficace : les temps de recherche sont typiquement inférieurs à une minute, ce qui est petit devant les

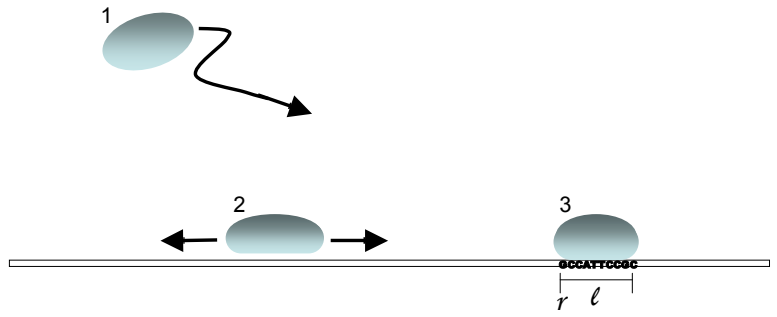


FIGURE 1.9 – **Différents états du facteur de transcription.** Figure tirée de (Lässig, 2007). Lors de sa recherche de site de fixation, le TF peut se trouver dans trois états distincts : (1) un état libre de diffusion tridimensionnelle, (2) un état de diffusion unidimensionnelle sur l'ADN par fixation non spécifique, et (3) un état de fixation spécifique. L'énergie de fixation dépend du site de fixation, de taille l et de coordonnée r.

processus de régulation de la cellule qui se déroulent au mieux sur quelques minutes (Gerland et al., 2002; Slutsky and Mirny, 2004). Il est donc pertinent de décrire l'effet d'un site de fixation sur la régulation d'un gène cible par la probabilité qu'il a de fixer un facteur de transcription à l'équilibre thermodynamique.

1.3.2 Modèle PWM

Présenté en 1987 par Berg et von Hippel (Berg and von Hippel, 1987), le modèle PWM est le modèle le plus simple décrivant l'énergie de fixation spécifique entre un facteur de transcription et un site de fixation sur l'ADN. Ce modèle repose sur plusieurs hypothèses. Tout d'abord, il y a l'hypothèse importante que les sites de fixation des TFs sur l'ADN ont été sélectionnés au cours de l'évolution pour leur propriété de sites de reconnaissance, quelle que soit la concentration du TF dans la cellule. En d'autres termes, le processus de sélection discrimine les sites de fixation sur la seule base de leur énergie de fixation à un TF donné : les sites ayant une énergie de fixation dans une certaine gamme sont retenus, les autres rejetés. Par ailleurs, au sein de cette gamme d'énérgie « utile », toutes les séquences sont équiprobables. Enfin, la dernière hypothèse est que chaque nucléotide d'un site de fixation contribue de manière indépendante, c'est-à-dire additive à l'énergie totale du site. Cette hypothèse permet de simplifier le problème en gardant le nombre de paramètres petit.

L'argument de Berg et von Hippel est que ce problème est analogue à celui de physique statistique consistant à déduire les taux d'occupation des niveaux d'énergie de particules indépendantes sachant que l'énergie totale doit avoir une certaine valeur moyenne E. La solution de ce problème est donnée par la formule de Boltzmann reliant énergie et taux d'occupation :

$$f_{i,b} = \exp(-\lambda E_{i,b}) / \mathcal{Z}_i \tag{1.1}$$

où $f_{i,b}$ est la probabilité d'observer la base b à la position i du site de fixation, $E_{i,b}$ est l'énergie associée (en $k_B T$), $\mathcal{Z}_i$ est la fonction partition qui permet de normaliser la distribution à la position i, et λ est un facteur sans dimension, analogue du β de la thermodynamique, et lié au processus de sélection. Dans la suite, nous intégrerons ce facteur à l'énergie.

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

1	2	3	4	5	6	7	8	9	10	11	12	13	14
B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

e Site scoring

0.45	−0.66	0.79	1.68	0.45	−0.66	0.79	0.45	−0.66	0.79	0.00	1.68	−0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum

f

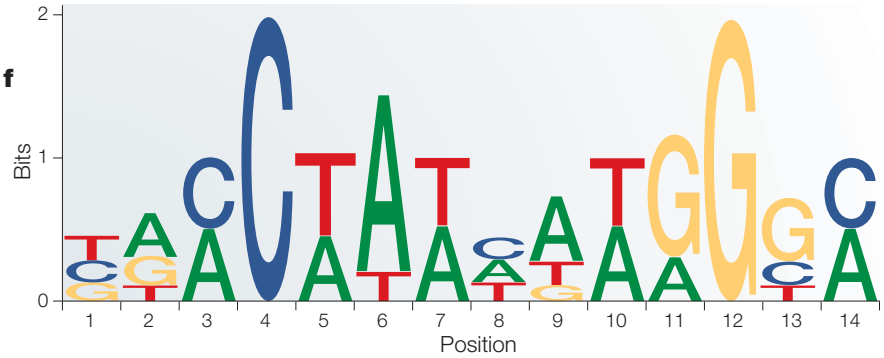


FIGURE 1.10 – **Construction et utilisation du modèle PWM.**
Figure tirée de (Wasserman and Sandelin, 2004). (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice de fréquence est construite, indiquant pour chaque nucléotide sa multiplicité à une position donnée dans le site. (d) La PWM est simplement construite en prenant le logarithme relatif des fréquences PWMs par rapport aux fréquences *a priori* des nucléotides. (e) Le score (ou énergie) d'une séquence d'ADN donnée est calculé en additionnant les poids PWM correspondants. (f) La PWM peut être représentée sous forme de logo (Giocomo et al., 2011). Dans cette représentation, la hauteur d'une colonne représente le contenu en information ou information relative moyenne d'une position, et la taille des bases reflète leur fréquence.

La connaissance des fréquences des bases permet de définir une autre quantité utile caractérisant la variabilité des séquences de fixation, l'information relative des sites par rapport à une séquence d'ADN aléatoire (Stormo and Fields, 1998) :

$$\mathcal{I} = \sum_{i=1}^{L} \sum_{b=A,C,G,T} f_{i,b} \ln\left(\frac{f_{i,b}}{\pi_b}\right) \tag{1.2}$$

où L est la taille du site de fixation et π_b correspond à la probabilité *a priori* d'observer la base b dans le génome. Il est usuel de définir l'énergie relativement au fond génomique :

$$\tilde{E}_{i,b} = \ln\left(\frac{f_{i,b}}{\pi_b}\right) \tag{1.3}$$

L'énergie totale d'un site S_i est alors

$$\begin{aligned} E &= \sum_{i=1}^{L} \tilde{E}_{i,b} \\ &= \sum_{i=1}^{L} \ln\left(\frac{f_{b(i)}}{\pi_b}\right) \\ &= \ln\left(\frac{\prod_{i=1}^{L} f_{b(i)}}{\prod_{i=1}^{L} \pi_b}\right) \\ &= \ln\left(\frac{P(S_i|\mathrm{TF})}{P(S_i|\text{fond génomique})}\right) \end{aligned} \tag{1.4}$$

où $b(i)$ est la base située à la position i du site de fixation. Cette énergie quantifie simplement à quel point la séquence S_i est plus ($E > 0$) ou moins ($E < 0$) probablement un site de fixation (de probabilité $P(S_i|\mathrm{TF})$) qu'un site tiré au hasard dans le génome (de probabilité $P(S_i|\text{fond génomique})$). On parle aussi de *score* de la séquence. L'information relative $\mathcal{I}$, qui est le score moyen des séquences fixées par le TF, peut alors être vue comme quantifiant à quel point l'ensemble des sites de fixation se distingue d'un ensemble de même taille de sites tirés au hasard.

Avec ces outils en main, il devient alors simple de bâtir un modèle PWM et de l'utiliser pour prédire des séquences fixées (fig. 1.10). Étant donnés des sites de fixation connus, il suffit d'évaluer la fréquence d'occurrence de chaque base à chaque position. La comparaison avec les probabilités génomiques *a priori* d'occurrence permet alors de bâtir une matrice de score, la PWM. Cette matrice peut alors être utilisée pour attribuer un score à une séquence d'ADN en additionnant les scores à chaque position. Finalement, les séquences ayant un score dépassant un certain seuil sont considérées comme des séquences de fixation.

1.3.3 Modèle biophysique

Le modèle PWM est basé sur une hypothèse forte, celle que les sites de fixation ont été sélectionnés sur la base de leur seule affinité ou énergie envers un TF. Néanmoins, à aucun moment n'intervient la concentration du TF dans la cellule, dont dépend pourtant la probabilité de fixation. C'est ce que tente de capturer le modèle biophysique (Gerland et al., 2002; Djordjevic et al., 2003; Zhao et al., 2009).

Considérons l'interaction entre un TF et une séquence d'ADN S_i :

$$TF + S_i \rightleftharpoons TF : S_i \tag{1.5}$$

où $TF : S_i$ dénote le complexe entre le TF et le site S_i. La constante d'équilibre de cette réaction s'écrit selon la loi d'action de masse :

$$K_i = \frac{[TF : S_i]}{[TF][S_i]} \tag{1.6}$$

Le site peut être dans deux états : occupé par le TF où libre. Aussi, la probabilité que le TF soit fixé au site s'écrit simplement

$$P(\text{fixation}|S_i) = \frac{[TF : S_i]}{[TF : S_i] + [S_i]} = \frac{1}{1 + \frac{1}{K_i[TF]}} = \frac{1}{1 + e^{\beta(E_i - \mu)}} \tag{1.7}$$

où $E_i = -kT\ln(K_i)$ est l'énergie libre standard de fixation (souvent notée ΔG), $\mu = kT\ln[TF]$ est le potentiel chimique, k est la constante de Boltzmann, T la température et $\beta = 1/kT$. Ici nous avons considéré qu'il n'y avait qu'un seul site de fixation. De manière générale, le site est en compétition avec le fond génomique, ce qui ajoute une contribution à μ (voir section 1.3.4). À l'instar du modèle PWM, l'énergie E_i est généralement prise comme étant une fonction additive des énergies individuelles des différentes bases du site. Ainsi, lorsque le TF est à faible concentration ($\mu \to -\infty$), le modèle biophysique écrit en équation 1.7 se réduit au modèle PWM.

1.3.4 Modèle thermodynamique

La description biophysique peut être réécrite en termes thermodynamiques en utilisant des raisonnements simples sur le nombre d'états possibles et leur énergie (et donc poids de Boltzmann) associée. Nous adoptons ici l'approche de (Gerland et al., 2002). On pourra par ailleurs se référer à l'excellente revue (Lässig, 2007). Considérons le cas simple d'un seul facteur de transcription interagissant avec un génome de taille $L \gg 1$ ne contenant qu'un seul site fonctionnel, le reste de la séquence étant aléatoire. Nous l'avons vu, l'expérience montre que la protéine se fixe à l'ADN avec une probabilité $1/2$. Lorsqu'elle est fixée, elle est à l'équilibre entre le mode spécifique et le mode non spécifique. Nous désirons savoir avec quelle probabilité elle est fixée de manière spécifique. La fonction de partition, énumérant tous les poids de Boltzmann associés aux différents états accessibles au TF fixé s'écrit :

$$\mathcal{Z} = \sum_{r=1}^{L} e^{-\beta E_s(r)} + Le^{-\beta E_{ns}} \tag{1.8}$$

Notons i la position du site fonctionnel. On peut écrire :

$$\begin{aligned} \mathcal{Z} &= e^{-\beta E_s(i)} + e^{-\beta E_{ns}} + \sum_{r \neq i} e^{-\beta E_s(r)} + (L-1)e^{-\beta E_{ns}} \\ &\simeq e^{-\beta E_i} + \mathcal{Z}_0 \end{aligned} \tag{1.9}$$

où Z_0 est la fonction de partition d'une séquence aléatoire, et nous avons introduit l'énergie E_i définie par

$$e^{-\beta E_i} = e^{-\beta E_s(i)} + e^{-\beta E_{ns}} \tag{1.10}$$

Dans le cas d'un site de reconnaissance, $E_{ns} \gg E_s(i)$ de sorte que $E_i \simeq E_s(i)$ (Gerland et al., 2002). La probabilité que le facteur soit fixé sur le site fonctionnel s'écrit finalement :

$$P(\text{fixation spécifique}|E_i) = \frac{e^{-\beta E_i}}{\mathcal{Z}} = \frac{1}{1+e^{\beta(E_i-F_0)}} \tag{1.11}$$

où $F_0 = -kT \log \mathcal{Z}_0$ est l'énergie libre d'une séquence génomique aléatoire. On reconnait une fonction de Fermi, avec un seuil d'énergie à F_0 : pour $E_i < F_0$, la protéine est essentiellement fixée de manière spécifique à son site de reconnaissance, alors que pour $E_i > F_0$, elle ne distingue plus le site du fond génomique et y est faiblement fixée.

Généralisons à présent au cas de plusieurs facteurs de transcription et sites de reconnaissance. Nous négligeons le recouvrement entre facteurs de transcription fixés sur des sites proches, qui poserait des problèmes stériques et corrèlerait les sites de fixation dans un certain voisinage (la présence d'un TF empêchant la présence d'un autre), et considérons que le nombre de TFs est grand devant le nombre de sites de reconnaissance pour éviter les problèmes de saturation : ainsi, le génome est composé de L séquences indépendantes, chacune pouvant être soit non occupée, soit occupée de manière non spécifique, soit occupée de manière spécifique. Notons μ le potentiel chimique du TF en solution. La fonction de partition totale est le produit des fonctions de partition des sites indépendants,

$$\mathcal{Z}(\mu) = \prod_{r=1}^{L} \mathcal{Z}(\mu, r) \tag{1.12}$$

où la fonction de partition d'un site s'écrit :

$$\mathcal{Z}(\mu, r) = e^{-\beta\mu} + e^{-\beta E_s(r)} + e^{-\beta E_{ns}} \tag{1.13}$$

En utilisant à nouveau la définition de E_i en éq.1.10, la probabilité de fixation d'un site à la position i s'écrit

$$P(\text{fixation spécifique}|E_i) = \frac{e^{-\beta E_i}}{\mathcal{Z}(\mu, i)} = \frac{1}{1+e^{\beta(E_i-\mu)}} \tag{1.14}$$

La valeur de μ est liée à la fois au nombre de TFs ainsi qu'à la possibilité de se fixer dans le fond génomique. Elle est fixée implicitement par l'équation :

$$n = \sum_{r=1}^{L} \frac{1}{1+e^{\beta(E_r-\mu)}} \tag{1.15}$$

qui signifie simplement que le nombre de TFs n dans le système est égal à la somme sur tous les sites de fixation possibles pondérée par la probabilité que le TF y soit fixé. Lorsque $\mu \to -\infty$ et que la fonction de Fermi peut être approximée par la loi de Boltzmann, l'équation peut s'inverser et l'on trouve (Aurell et al., 2007)

$$\mu = F_0 + kT \log n \tag{1.16}$$

où F_0 est l'énergie libre du fond génomique introduite en éq. 1.11. Ainsi, la prise en compte d'une multiplicité de TFs ajoute un facteur $kT \log n$ au seuil de la fonction de Fermi par rapport au cas d'un seul TF. Par ailleurs, cette approche thermodynamique nous a permis de généraliser le modèle biophysique simple introduit en section 1.3.3.

1.4 Les interactions protéine-ADN : mesures expérimentales

Ces dernières années, des avancées technologiques considérables ont permis d'une part d'établir des modèles de fixation spécifique pour de nombreux TFs, d'autre part de localiser leurs sites de fixation dans le génome. Ces avancées ont eu lieu autant sur le plan *in vitro*, utilisant protéines purifiées et séquences nucléiques artificielles pour déduire l'affinité protéine-ADN, que sur le plan *in vivo*, mesurant l'interaction de la protéine avec l'ADN génomique (Stormo and Zhao, 2010).

1.4.1 Approches *in vitro* : MITOMI, SPR, PBM, CSI, SELEX, et HT-SELEX

- **Approche microfluidique : MITOMI**

En 2007, Maerkl et Quake ont mis au point une technique appelée MITOMI (Mechanically Induced Trapping Of Molecular Interactions) permettant une mesure directe de l'affinité d'un TF à des centaines de séquences d'ADN à la fois (Maerkl and Quake, 2007). Cette technique repose sur l'utilisation d'un système microfluidique composé de chambres dans lesquelles un fluide dont on peut facilement modifier la composition circule dans des canaux d'un diamètre de l'ordre de 1μm dont le microenvironnement est finement contrôlé. Le fluide contient des gènes synthétiques codant pour le TF ainsi que du matériel permettant la synthèse de la protéine directement au sein de la chambre, ce qui évite de purifier préalablement le TF. Chaque chambre du système contient des anticorps fixés à la surface permettant de capturer le TF et une certaine concentration d'une séquence d'ADN spécifique contenant une marque fluorescente. Le système contient ainsi des centaines de séquences d'ADN différentes, chacune étant présente à différentes concentrations. Lorsque le TF est fixé par les anticorps, il recrute des séquences d'ADN selon leur affinité. Celles qui ne se fixent pas sont lavées. Au final, les séquences fixées produisent un signal de fluorescence. La comparaison des signaux pour différentes concentrations d'ADN donne accès au rapport des constantes d'équilibre K_{eq} (eq. 1.6). La comparaison avec une séquence référence dont la constante K_{eq} est connue permet alors de déterminer le K_{eq} absolu pour chaque séquence de fixation.

En utilisant 17 systèmes de ce type, ils ont ainsi pu mesurer l'affinité de 4 TFs de type bHLH à 464 séquences d'ADN différentes : les séquences consensus et des séquences ayant une, deux, trois ou quatre mutations. À titre de comparaison, ils ont construit une PWM à partir des séquences contenant une seule mutation, puis ont prédit les énergies attendues des séquences à plusieurs mutations. La prédiction de la PWM s'est avérée bonne dans seulement 56% des cas pour les séquences à deux mutations, 10% pour les séquences à 3 mutations et 0% des cas pour les séquences à 4 mutations, montrant les limites de ce modèle indépendant confronté à des données d'interactions d'ordre supérieur. Un modèle plus raffiné prenant en compte l'énergie d'interaction non spécifique et incluant des interactions entre nucléotides voisins permet néanmoins de rendre compte des valeurs observées (Stormo and Zhao, 2007). Nous reviendrons sur la nécessité de prendre en compte les interactions entre paires de nucléotides lors de l'interaction spécifique entre TF et ADN dans le chapitre 2.

- **Approche physique : la microscopie SPR**

La méthode de résonance des plasmons de surface (*Surface Plasmon Resonance* ou SPR) est habituellement utilisée pour étudier l'interaction d'une protéine avec un ligand (qui peut être une autre protéine), mais elle peut aussi être utilisée pour mesurer les interactions entre une protéine et quelques centaines de séquences d'ADN différentes (Shumaker-Parry et al., 2004;

Campbell and Kim, 2007). Le principe de la microscopie SPR est que l'angle de réflection de la lumière sur une fine surface d'or, par exemple, dépend de la masse de molécules fixées de l'autre côté de sa surface. Si de l'ADN est lié à la surface, la fixation du TF induit un changement de masse et donc d'angle de reflection lumineuse mesurable au cours du temps. Ainsi, la cinétique de fixation du TF jusqu'à l'atteinte de l'équilibre est accessible. On peut de même étudier la dissociation du TF lors du lavage de la surface. Ces mesures donnent directement accès aux taux d'association k_{on} et de dissociation k_{off} que la simple mesure de la constante d'équilibre $K_{eq} = k_{on}/k_{off}$ ne permet habituellement pas de déterminer.

- **Approches basées sur des puces à ADN : PBM et CSI**

L'analyse de fixation des protéines par puce à ADN (*Protein-Binding Microarray* ou PBM) est une technologie haut débit qui a été développée au cours des 10 dernières années (Berger et al., 2006). Les puces sont composées de 44,000 puits auxquels sont liés des brins d'ADN. Une puce contient tous les sites de fixation de 8bp possibles ($4^8/2 = 32,768$ séquences en prenant en compte le fait qu'il y a un site sur chacun des deux brins d'ADN) plus deux bases flanquantes (une à chaque extrémité) qu'il est possible de faire varier. Un TF purifié à partir de cellules ou synthétisé *in vitro* est ajouté à la puce, qui est ensuite lavée pour se débarrasser des fixations non spécifiques. La quantité de protéine fixée à un puits donné est déterminée grâce à un anticorps fluorescent contre la protéine. L'enrichissement en protéine est calculé relativement au bruit de fond (anticorps non spécifique par exemple). Il est alors possible d'utiliser ces mesures pour bâtir une PWM du TF (voir par exemple Kinney et al. (2007)).

Une autre méthode utilise aussi des puces à ADN : c'est l'identification de site apparenté (*Cognate Site Identifier* ou CSI) (Warren et al., 2006). Une différence technique avec les PBMs est que l'ADN est d'abord synthétisé en simple brin puis se replie en double brin pour former le site de fixation, évitant ainsi de devoir générer l'ADN double brin à partir de précurseurs. Par ailleurs, le TF est en compétition avec un marqueur fluorescent qui peut se fixer à l'ADN : il n'est donc pas nécessaire d'utiliser un marquage spécifique sur le TF ou sur un anticorps, ce qui rend la procédure plus généralisable. Finalement, la spécificité du TF est représentée par un « paysage de spécificité » qui encapsule l'information de fluorescence de l'ensemble des variations par rapport à une séquence consensus dans une représentation simple (Carlson et al., 2010).

- **Approche par purification des séquences fixées : SELEX et HT-SELEX**

Mise au point il y a plus de 20 ans, la méthode SELEX (*Systematic Evolution of Ligands by EXponential enrichment*) repose sur la sélection de séquences d'ADN aléatoires par un TF *in vitro* (Oliphant et al., 1989; Tuerk and Gold, 1990; Blackwell and Weintraub, 1990; Wright et al., 1991). Une bibliothèque de sites de fixation potentiels est d'abord générée en synthétisant des séquences d'ADN aléatoires ou en utilisant des séquences génomiques. Les extrémités de ces séquences contiennent des précurseurs permettant l'amplification exponentielle par PCR. Le TF purifié est ajouté aux sites et les séquences fixées sont séparées des séquences non fixées, par exemple par retard sur gel. Après un cycle de sélection, les séquences récupérées sont encore enrichies en séquences de basse affinité pour le TF, car celles-ci sont simplement initialement bien plus abondantes que les séquences de haute affinité. Afin d'augmenter la proportion de séquence de grande affinité, les séquences filtrées sont amplifiées puis filtrées à nouveau, ceci sur plusieurs cycles. À la fin de ce processus, les séquences sélectionnées sont clonées et séquencées, résultant en un nombre typique de moins de ~ 100 séquences indépendantes (Fields et al., 1997). Si les séquences initiales sont issues d'ADN génomique, il est

possible d'utiliser l'hybridation des séquences à des puces à ADN. La présence de plusieurs cycles de sélection rend néanmoins la détermination des énergies de fixation moins directe qu'avec les techniques précédentes. Une variante de la technique appelée SELEX-SAGE utilise des multimères de sites à la place de sites uniques et permet de réduire le nombre de cycles de sélection et d'augmenter ainsi le nombre de séquences de fixation obtenues (Roulet et al., 2002), permettant de réaliser des modèles plus précis (Nagaraj et al., 2008).

Depuis la mise au point de la méthode SELEX, des avancées considérables ont été réalisées dans les techniques de séquençage, permettant l'obtention de millions de séquences à la fois : on parle de séquence haut-débit (*high-throughput*) ou encore séquençage massivement parallèle. L'utilisation de ces nouvelles techniques dans l'expérience SELEX a mené à la méthode HT-SELEX (Nagaraj et al., 2008), aussi appelée Bind-n-Seq (Zykovich et al., 2009). Il est alors possible d'estimer un modèle d'énergie à partir des fréquences d'observation des différentes séquences dès le premier cycle (Nagaraj et al., 2008). Des cycles supplémentaires permettent d'obtenir plus d'information sur les séquences les plus spécifiques, notamment sur la présence de contributions non indépendantes à l'énergie, ou de compenser la faible spécificité d'un TF. L'avantage de cette technique est que la taille des sites de fixation n'est pas limitée. Ainsi, avec une nanomole d'ADN ($\sim 10^{15}$ séquences) on peut couvrir l'ensemble des sites de 25bp possibles. Le séquençage haut-débit permet d'en échantillonner $\sim 10^8$, ce qui est largement suffisant pour contraindre des modèles d'énergie indépendants, même pour des TFs ayant des sites de fixations de taille > 15bp comme c'est souvent le cas chez la bactérie. Cette technique a récemment été poussée encore plus loin (Jolma et al., 2010). En utilisant des protéines marquées, les auteurs ont réalisé un HT-SELEX à partir d'extraits cellulaires, et en ajoutant un code barre aux séquences d'ADN de chaque expérience, ils ont pu analyser les sites de fixation pour plusieurs TFs en parallèle. Ils ont ensuite utilisé cette technique pour obtenir des modèles de spécificité pour 411 TFs humains, la plus grande étude de ce genre réalisée à ce jour (Jolma et al., 2013).

1.4.2 Approche clonale : la technique de simple hybride

Contrairement aux approches précédentes, la technique de simple hybride (*Bacterial one-hybrid* ou B1H) n'est pas purement *in vitro*, au sens où l'interaction protéine-ADN est testée au sein d'une bactérie. Néanmoins, parce que l'interaction n'est pas testée dans son contexte cellulaire d'origine, nous la considérerons comme telle. Cette approche repose sur l'intégration par une bactérie hôte de deux vecteurs d'expression génétique, ou plasmides. Le premier exprime le facteur de transcription d'intérêt fusionné à une sous-unité de l'ARN polymérase (l'appât), c'est la protéine « hybride ». L'autre contient une région de séquence aléatoire représentant un site de fixation potentiel (la proie) en amont d'un promoteur à faible activité. La fixation de cette région par la protéine hybride permet l'activation d'un gène de sélection, généralement *HIS3*, un gène de la levure requis pour la biosynthèse de l'histidine et dont l'homologue bactérien est absent de la souche d'*Escherichia coli* utilisée. La croissance des cellules a lieu dans un milieu ne contenant pas l'histidine. Dans ces conditions, les bactéries n'exprimant pas *HIS3* ne peuvent croître. Ainsi, seules les bactéries au sein desquelles le facteur de transcription se fixe à la proie expriment *HIS3*, croissent et forment des colonies, d'où la notion de gène de sélection. Par ailleurs, la stringence de la sélection peut être modulée en ajoutant au milieu différentes concentrations de 3-amino-triazole (3-AT), un inhibiteur de *HIS3*. De cette façon l'affinité du site de fixation peut être estimée plus finement.

Dans les études de ce type, les sites de fixation présents au sein des colonies sont sé-

quencés individuellement, ce qui permet d'obtenir environ 50 séquences pour une expérience de sélection donnée. Néanmoins, il semble possible d'utiliser les nouvelles technologies de séquençage pour récupérer l'ensemble des sites de fixation des bactéries présentes sur une plaque (Stormo and Zhao, 2010). À l'instar de la méthode HT-SELEX, on obtient des millions de sites, ceux ayant une plus grande affinité étant présents à plusieurs centaines de milliers d'exemplaires, et ceux ayant une faible affinité étant présent en un seul voire aucun exemplaire.

Notons qu'il est aussi possible d'adopter la démarche inverse, c'est-à-dire de partir de quelques sites de fixation présumés fonctionnels mais pour lesquels on ne connaît pas le TF associé. En utilisant une bibliothèque de plasmides codant pour différents TFs hybrides, il est alors possible de déterminer si l'un d'entre eux possède une affinité importante avec les sites testés.

1.4.3 Approches *in vivo* : ChIP-on-chip, ChIP-seq, DNAse I

Dans cette section, nous nous intéressons aux techniques permettant d'identifier les sites de fixation d'un facteur de transcription sur le génome. Ces méthodes se basent sur des extraits cellulaires (de 10^4 à 10^8 cellules) qui peuvent provenir d'un tissu homogène (un seul type de cellule) ou hétérogène (plusieurs types de cellules), voire de l'organisme entier si la dissection est impossible (embryon de mouche par exemple). L'information obtenue est donc toujours conditionnée par ce matériau de départ, et l'on n'obtient que les sites *accessibles* étant donnés le type cellulaire et la période de développement étudiés.

- **Immunoprécipitation de la chromatine : ChIP-on-chip et ChIP-seq**

La technique d'immunoprécipitation de la chromatine (ChIP) (fig. 1.11) consiste dans un premier temps à induire la réticulation (*crosslink*) des protéines se liant à l'ADN en traitant les cellules avec de la formaldéhyde. Cette étape permet de transformer les liaisons faibles protéine-ADN en liaisons covalentes. Une fois les protéines fixées, la chromatine est découpée par digestion enzymatique ou en la soumettant à des ultrasons (c'est la sonication), résultant en des fragments de taille variant entre 200 et 600bp. Ces fragments sont ensuite immunoprécipités en présence d'un anticorps spécifique d'un facteur de transcription ou d'un isoforme d'histone (dans le cas d'une étude du paysage épigénétique) d'intérêt, permettant ainsi de récupérer tous les sites de fixation dans le génome. Après purification des fragments précipités, l'échantillon peut être analysé soit par hybridation sur puce (ChIP-on-chip) ou par séquençage haut débit (ChIP-seq).

Dans le cas du ChIP-on-chip, l'échantillon immunoprécipité et l'ADN de départ (*input*) sont marqués avec des colorants fluorescents et hybdridés sur une puce à ADN composée de très nombreux puits contenant des oligonucléotides (courtes séquences d'ADN) correspondant à différentes régions du génome. Dans le meilleur cas, ces oligonucléotides couvrent l'ensemble du génome. Les sites de liaison sont identifiés par l'écart d'intensité entre les signaux de fluorescence des conditions d'immunoprécipitation et d'*input*.

Dans le cas du ChIP-seq, l'échantillon immunoprécipité est analysé par séquençage à haut débit, résultant en une librairie de *reads* d'une longueur typique variant entre 27 et 50bp issus des extrémités des séquences. Ces *reads* sont ensuite alignés sur un génome de référence. À chaque position du génome correspond ainsi un certain nombre de séquences précipitées et

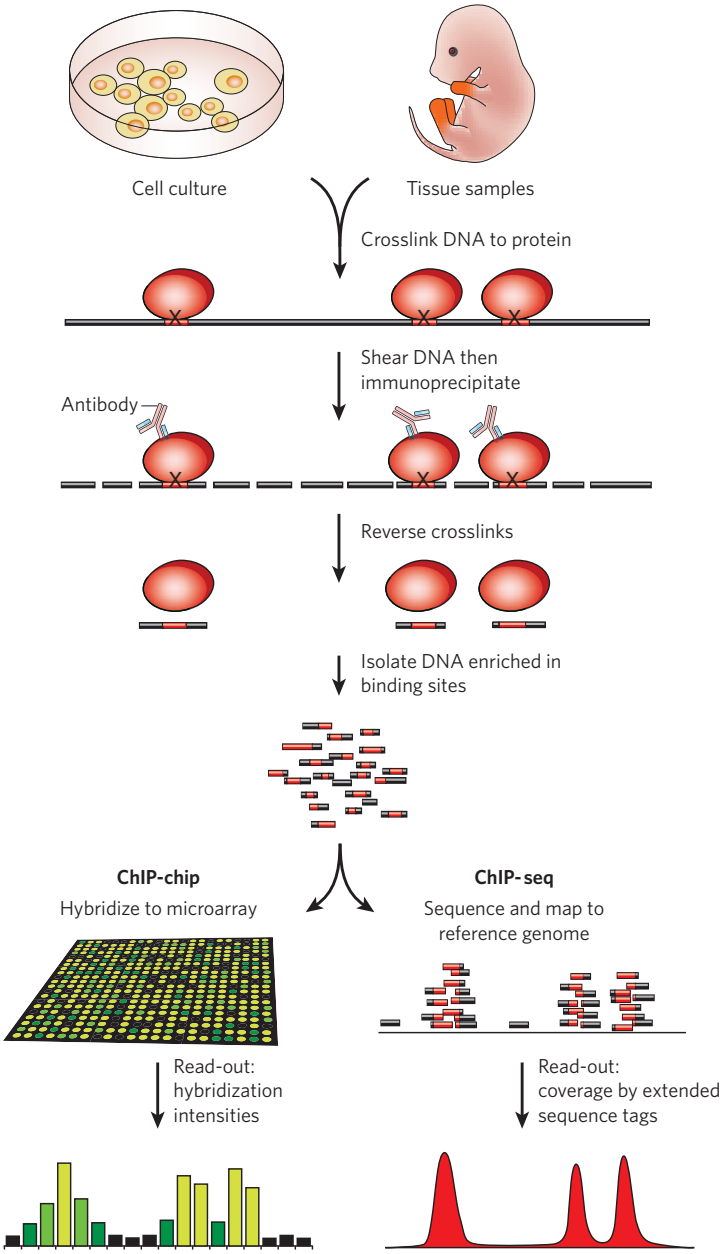


FIGURE 1.11 – **Étapes d'une expérience de ChIP-on-chip et ChIP-seq.**
Figure tirée de Visel et al. (2009b). À partir d'extraits cellulaires issus de cultures *in vitro* ou prélevés *in vivo*, plusieurs étapes permettent de récupérer les régions fixées par un TF d'intérêt. D'abord, les protéines, liées de manière non covalente à l'ADN, sont fixées par utilisation d'un agent de réticulation, le formaldéhyde. Puis la chromatine est découpée en fragments de $\sim$ 500bp par utilisation d'ultrasons : on parle de sonication. L'utilisation d'un anticorps spécifique au TF d'intérêt permet de précipiter les fragments d'ADN fixés. Les fragments résultants sont alors soit hybridés à une puce à ADN contenant de nombreuses séquences génomiques (ChIP-on-chip), soit directement séquencés et alignés à un génome de référence (ChIP-seq). La comparaison à un échantillon d'ADN non précipité (« input ») permet alors de définir les régions significativement fixées, avec une résolution plus précise dans le cas du ChIP-seq, où l'ensemble du génome est couvert, par rapport au ChIP-on-chip, pour lequel seulement un sous ensemble du génome est analysé.

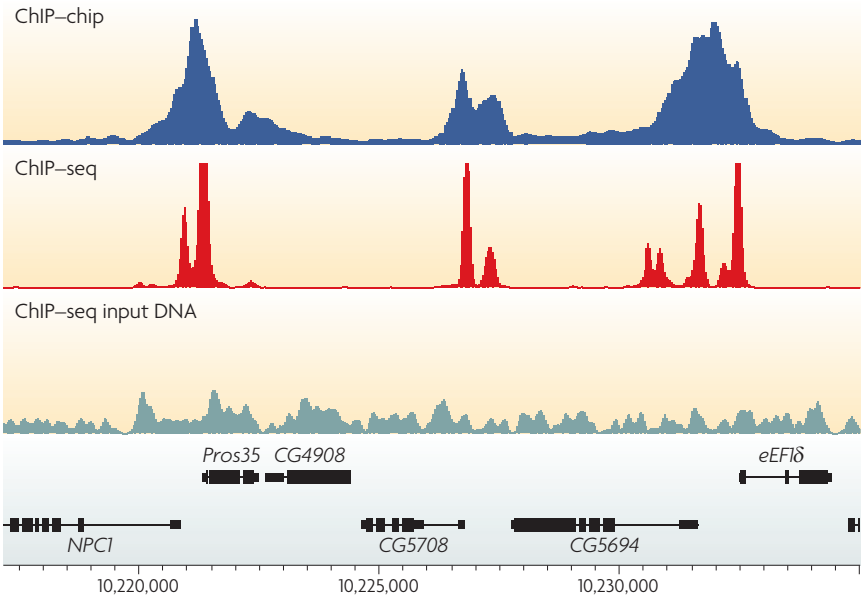


FIGURE 1.12 – **Résolution des expériences ChIP-on-chip et ChIP-seq.** Figure tirée de (Park, 2009), montrant les profils de fixation de la protéine Chromator générés à partir d'expériences de ChIP-on-chip (intensité relative par rapport au contrôle, bleu) et de ChIP-seq (densité de séquences, rouge) dans la lignée cellulaire S2 de *Drosophila melanogaster*. On peut noter la plus grande résolution de l'expérience ChIP-seq pour déterminer les sites de liaison. L'ADN utilisé en *input* de l'expérience de ChIP-seq et servant de contrôle est montré en gris, et les gènes du locus indiqués en noir.

d'*input*. En comparant ce nombre au nombre moyen dans le locus et à l'*input*, il est possible d'identifier des pics correspondant à la fixation du facteur (voir par exemple le programme d'appel de pics ChIP-seq MACS (Zhang et al., 2008)).

Dans les deux cas, il faut noter que l'on a affaire à la fixation *moyenne* du facteur sur l'ADN dans la population de cellules étudiée. Ainsi, un petit pic peut représenter aussi bien une fixation forte dans un petit sous-ensemble de cellules (par exemple celles qui sont à un certain état d'avancement du cycle cellulaire) qu'une fixation moyenne dans l'ensemble de la population. L'expérience de ChIP-seq offre une résolution bien plus précise ($\leq$ 100bp) que la méthode ChIP-on-chip (fig. 1.12). En effet, dans ce dernier cas la résolution est limitée par le nombre d'oligonucléotides utilisés, qui sont dans le meilleur des cas répartis sur le génome avec $35 - 100$ nucléotides d'écart entre deux instances. Pour se comparer à la ChIP-seq, il faudrait que tous les oligonucléotides se superposent à une base près, ce qui demanderait un trop grand nombre de puces.

- **Empreinte à la DNase I (*DNase I footprinting*)**

Contrairement aux techniques précédentes, l'empreinte à la DNase I ne repose pas sur l'étude d'un facteur de transcription précis, mais permet au contraire d'obtenir l'ensemble des sites de fixation dans le génome pour un type cellulaire donné, avec une précision au nucléotide près. Cette méthode repose sur le fait que la fixation stable des facteurs de transcription à l'ADN n'est possible que si la région est pauvre en nucléosomes, les protéines

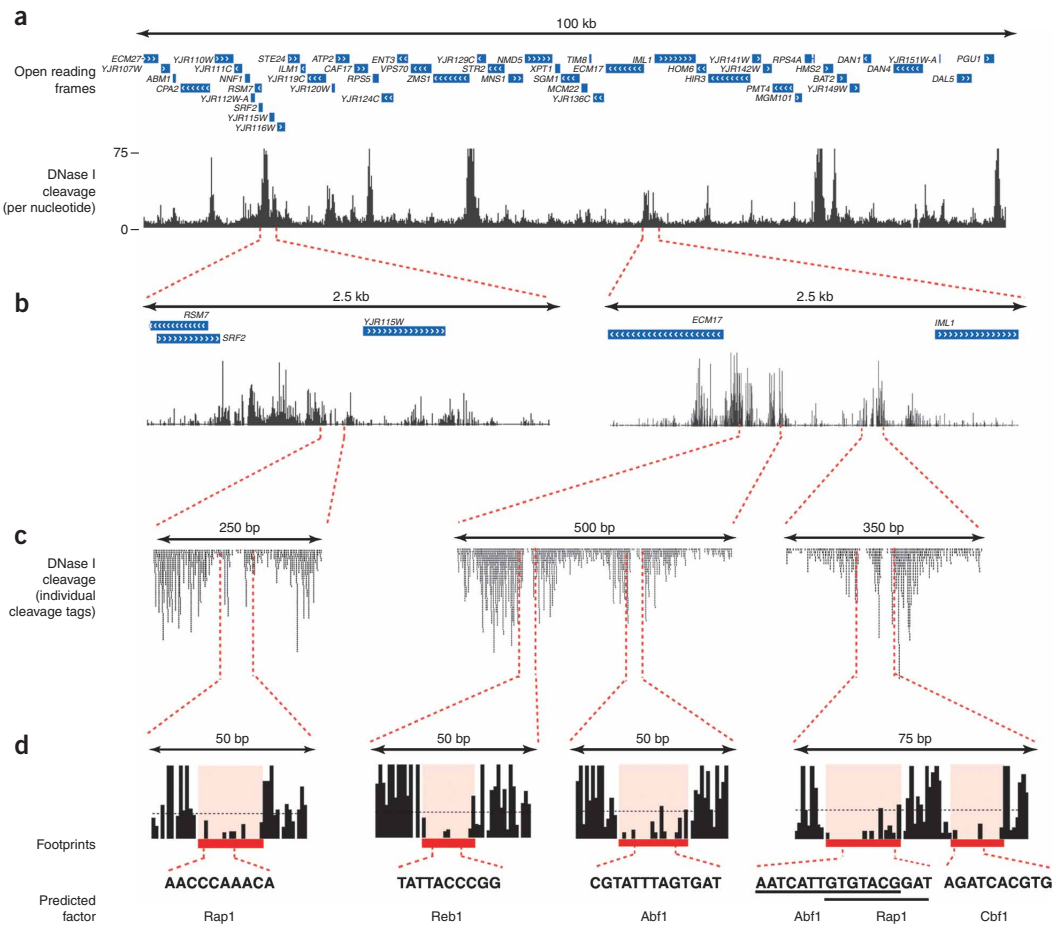


FIGURE 1.13 – **Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près.** Figure tirée de (Hesselberth et al., 2009). (a) Densité de digestion de la DNase I par nucléotide dans une région de 100kb du génome de la levure contenant $\sim$ 50 gènes (boîtes bleues). On voit clairement certaines régions promotrices marquées par la DNase I. (b) Zoom sur deux régions de 2.5kb. (c) Zoom sur des régions de 250bp. Le nombre d'événements de digestion par nucléotide est marqué par des empilements de boîtes noires, révélant des régions protégées : ce sont les empreintes à la DNase I. (d) Les empreintes sont associées à des motifs de régulation de la levure connus. Les pointillés noirs indiquent le nombre moyen de digestion par nucléotide dans le génome ($\sim$ 2 digestions par base).

autour desquelles s'enroule l'ADN : on parle de région de chromatine ouverte. Ces régions sont préférentiellement digérées par l'endonucléase DNase I. Étant donné que la majorité de l'ADN est enroulé autour de nucléosomes, les sites hypersensibles à la digestion par DNase I (*DNase I-hypersensitive* ou DHS) correspondent essentiellement à des régions de chromatine ouverte ayant des roles de régulation génétique : promoteurs, enhancers...

En combinant la technique de DHS avec le séquençage à haut débit, l'expérience de DNase-seq permet d'identifier tous les types de région de régulation à l'échelle du génome (Thurman et al., 2012). Les régions riches en sites de digestion identifient alors les sites DHS. Par ailleurs,

au sein d'un site DHS, il y a de petites régions ($\sim$ 15bp) qui sont protégées de la digestion par DNase I : ce sont les empreintes à la DNase I ou *DNase I footprints* (fig. 1.13). Ces empreintes sont dues à la présence de protéines ou de complexes fixés à l'ADN. Cette technique de détection de sites de liaison par empreinte à la DNase I existe depuis 30 ans mais n'a que récemment été porté à l'échelle génomique. En comparant à des données ChIP-seq ou en utilisant des bases de données de motifs de facteurs de transcription, il est possible d'identifier le facteur correspondant dont les sites de fixation sont alors connus au nucléotide près.

1.5 Les modules de cis-régulation (CRMs)

Nous l'avons vu en section 1.2.2, les séquences d'ADN régulant l'expression génétique – CRMs pour *Cis-Regulatory Modules* – jouent un rôle prépondérant au cours du développement des organismes. Ces CRMs assurent en effet l'orchestration de l'expression de gènes spécifiques aux différentes étapes du développement et aux divers types cellulaires. Ils sont au coeur de l'évolution des réseaux génétiques, car ils dictent les interactions entre gènes. De plus, leur altération peut conduire à de nombreuses pathologies, liées pour la plupart à une expression génétique aberrante. Notamment, la majeure partie des variants génétiques qui sont associés de manière significative à une susceptibilité envers une maladie sont situés hors des régions codant pour des protéines, suggérant qu'un certain nombre affectent non pas la forme de la protéine engendrée mais l'expression du gène la produisant en détruisant une activité CRM. Dans cette partie, nous présentons les différents types de CRMs, leur structure, et leur évolution.

1.5.1 Les différents types de CRMs

Selon leur rôle dans la régulation de l'expression génétique, les CRMs peuvent être distingués en trois catégories.

- **Promoteurs**

Les promoteurs permettent la fixation de l'ARN polymérase pour débuter la formation d'un transcrit ARN au site d'initiation de transcription (*Transcription Start Site* ou TSS). Dans les promoteurs fixant l'ARN polymérase II (la majorité des promoteurs eucaryotes), des facteurs de transcription généraux se fixent à un coeur de $\sim$ 100bp autour du TSS afin de faciliter la fixation du complexe de polymérase. Ces coeurs de promoteurs contiennent pour certains des motifs stéréotypés, comme la boîte TATA, et ont un TSS bien déterminé ; néanmoins la plupart des promoteurs des génomes mammifères sont des régions riches en GC et en dinuclétoides CpG (les « îlots CpG ») qui ne possèdent pas de boîte TATA et permettent l'initiation de la transcription dans un interval d'environ 100 bases (Carninci et al., 2006). Au niveau épigénétique, les promoteurs actifs sont caractérisés par une région pauvre en nucléosomes en amont du TSS, flanquée de nucléosomes possédant la marque de méthylation H3K4me3.

- ***Enhancers* et *silencers***

Les *enhancers* et *silencers* sont respectivement définis par leur effet positif ou négatif sur l'expression d'un gène cible. Cet effet peut notamment être observé par transfert d'un plasmide contenant l'élément régulateur en amont d'un gène rapporteur dans un animal transgénique ou dans des cultures cellulaires transfectées (voir 1.6.4). Leur activité ne dépend généralement pas de leur position et de leur orientation sur le plasmide. Selon l'environnement cellulaire, une région régulatrice peut être soit *enhancer* soit *silencer*, en fonction de la nature

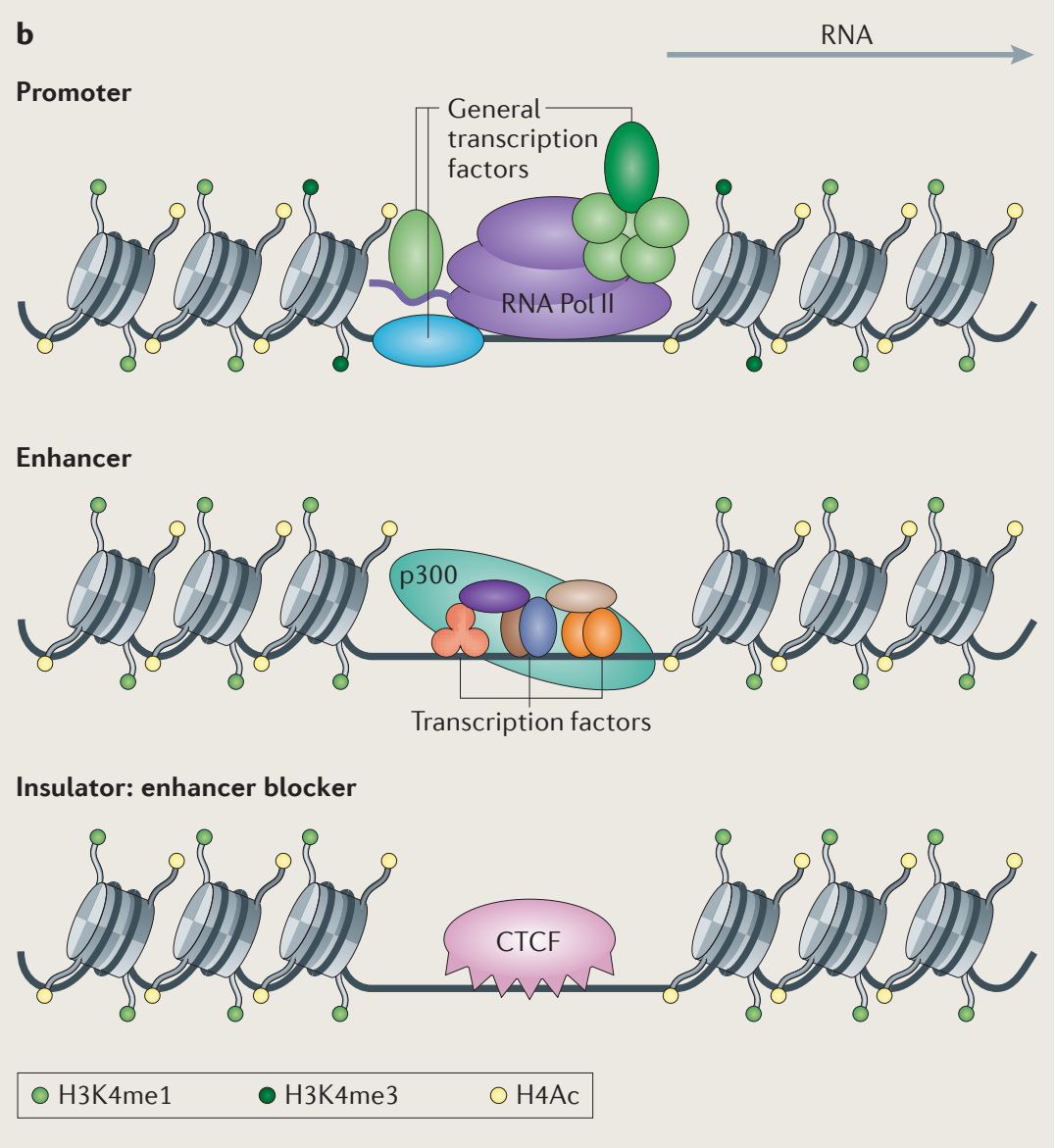


FIGURE 1.14 – **Les différents types de CRMs et leurs marques épigénétiques.** Figure tirée de (Hardison and Taylor, 2012). La notion de CRM renvoie à un regroupement de sites de liaison pour un ou plusieurs facteurs de transcription. Les CRMs peuvent être regroupés en plusieurs classes : les promoteurs, les *enhancers/silencers*, et les insulateurs. Les CRMs des différentes classes partagent les marques d'acétylation H3Ac et H4Ac, les promoteurs actifs sont spécifiquement marqués par H3K4me3, et les enhancers et insulateurs par H3K4me1. Les enhancers sont par ailleurs souvent fixés par le co-activateur p300. Enfin, chez les mammifères les insulateurs recrutent CTCF pour bloquer l'activation par les enhancers.

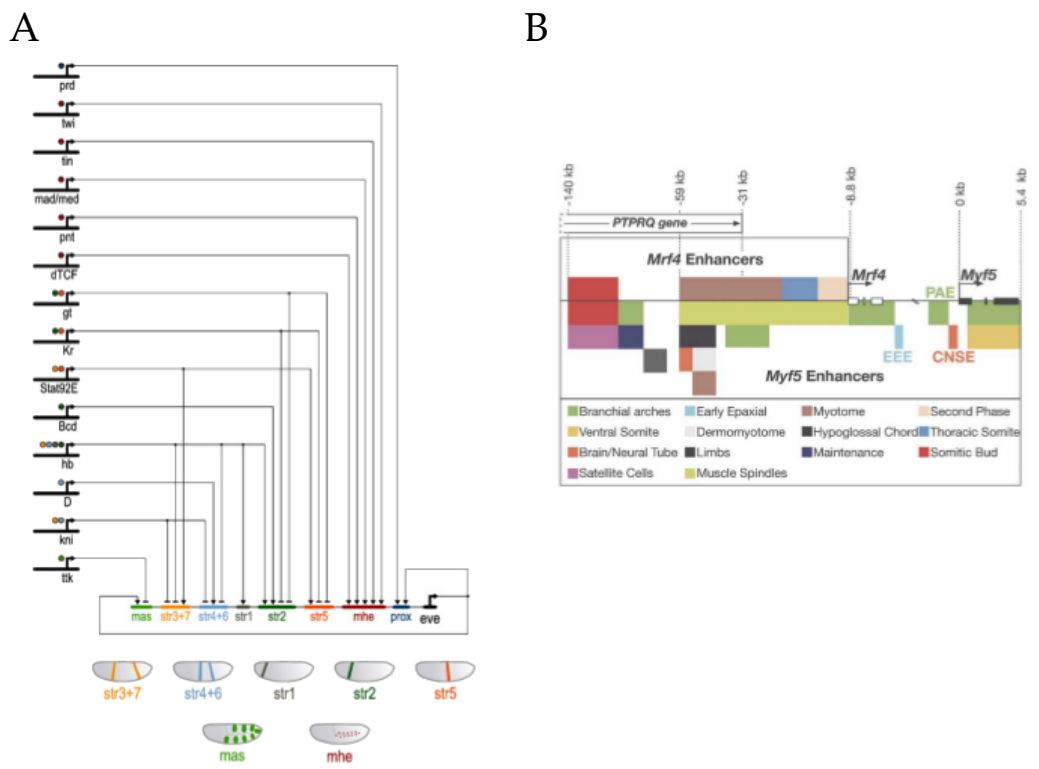


FIGURE 1.15 – **Différents *enhancers* conduisent à différents patterns d'expression.**
(A) Figure tirée de (Wilczynski and Furlong, 2010), montrant le réseau de cis-régulation du gène *even-skipped* chez la *Drosophile*. Des enhancers individuels (boîtes colorées) conduisent à des motifs d'expression distincts (indiqués dans les embryons par la même couleur). Les cercles colorés au sein des régulateurs indiquent les couleurs des CRMs régulés.
(B) Figure tirée de (Carvajal et al., 2008), représentant les différents éléments régulateur du locus *Mrf4/Myf5*. Chaque boîte colorée représente une régulation de la transcription à un stade de développement particulier ou dans une région anatomique spécifique du muscle embryonnaire ou adulte.

de co-activateurs ou de co-répresseurs des TFs recrutés. Il y a néanmoins relativement peu de *silencers* caractérisés et l'on utilise le terme d'*enhancers* pour désigner de manière générale ces régions régulatrices.

Les *enhancers* peuvent se situer à des distances variables du gène qu'ils régulent (Maniatis et al., 1987), pouvant parfois aller jusqu'à 1 Mb comme dans le cas de *Shh* chez la souris (Lettice et al., 2003) (voir fig. 1.24). Les enhancers contiennent de multiples sites de fixations de TFs. Cette multiplicité est requise pour l'activité enhancer, comme cela l'a été montré pour le premier enhancer découvert : celui du virus simien 40 (SV40) (Schirm et al., 1987; Ondek et al., 1988). Un gène peut par ailleurs posséder plusieurs enhancers distincts conduisant à des expressions spécifiques dans différents tissus, comme cela l'a été montré dans le cas du gène *eve* chez la *Drosophile* (Wilson and Odom, 2009) ou dans le cas du cluster de gènes de détermination myogénique *Myf5* et *Mrf4* chez les mammifères (Carvajal et al., 2008) (fig. 1.15). Ainsi, les différents enhancers d'un même gène peuvent être vus comme autant de points d'entrée d'un réseau de régulation génétique, représentant diverses fonctions logiques et intégrant différentes information spatio-temporelles pour produire en sortie une expression génétique

spatio-temporelle finement contrôlée (Bolouri and Davidson, 2002; Buchler et al., 2003).

Enfin, comme décrit en fig. 1.14, les enhancers sont associés à de hauts niveaux de marque épigénétique H3K4me1 (Heintzman et al., 2009) et sont souvent fixés par le co-activateur p300 (Wang et al., 2005; Heintzman et al., 2009).

- **Insulateurs**

Les insulateurs sont des CRMs qui restreignent l'effet des enhancers sur leur gène cible (Wallace and Felsenfeld, 2007). Ainsi, certains insulateurs possèdent une activité de blocage d'enhancers. Situés entre un enhancer et un promoteur cible, ces insulateurs bloquent l'activité de l'enhancer, conduisant à une réduction de l'expression du gène cible (Chung et al., 1993). Chez les mammifères, la fixation de la protéine CTCF est nécessaire à cette activité de blocage de l'activité enhancer (Bell et al., 1999), alors que chez la *Drosophile* et plusieurs autres insectes il existe au moins quatre protéines additionnelles qui sont suffisantes à la réalisation de cette activité (Schoborg and Labrador, 2010). Par ailleurs, les insulateurs peuvent servir de barrière de protection contre des marques d'hétérochromatine répressives. De tels insulateurs permettent notamment d'éviter les effets de positions – la modification de l'expression d'un gène selon sa position dans le chromosome – lorsqu'ils entourent un gène rapporteur intégré au hasard dans le génome (Recillas-Targa et al., 2002). Cette activité passe notamment par le recrutement de *USF*, protéine qui recrute des enzymes de modification de la chromatine. Un insulateur peut combiner les activités de barrière de protection et de blocage d'enhancer.

De même que les enhancers, les insulateurs peuvent se situer à des distances variables des gènes qu'ils régulent. Il est à noter que la protéine CTCF possède d'autres fonctions que celle d'isolation, et tous les sites de CTCF ne correspondent pas forcément à des insulateurs (Phillips and Corces, 2009).

1.5.2 Grammaire des enhancers : enhanceosome vs billboard

Nous l'avons vu, les CRMs contiennent en général de multiples sites de liaisons (TFBS) pour un ou plusieurs TFs. On parle de *clustering* (regroupement). Lorsque les TFBS correspondent à plusieurs TFs différents, on parle de CRM hétérotypique, et dans le cas où ils correspondent à un même TF, on parle de CRM homotypique. Cette distinction est surtout utile pour décrire les différentes méthodes de prédiction de CRM, car la plupart des CRMs identifiés chez les Métazoaires sont hétérotypiques (Aerts, 2012). L'organisation de ces sites de liaison relève de deux types d'architecture principaux (fig. 1.16).

- **Le modèle "enhanceosome"**

Dans ce modèle, l'architecture des sites de liaison est de première importance. Le paradigme en est l'enhancer du gène humain interferon-β, sur lequel 8 TFs se lient pour former une surface de reconnaissance continue (Panne, 2008). Les TFBS de cet enhancer se recouvrent les uns les autres, créant au final un complexe de TFs fixés à l'ADN agissant comme une seule unité de régulation (fig. 1.17).

- **Le modèle "billboard"**

La majorité des CRMs adhèrent à ce type d'organisation. L'architecture y est libre : les sites de liaisons n'ont pas de contrainte de nombre, d'ordre, de sens, ou d'espacement (Kulkarni and Arnosti, 2003). De tels CRMs sont propices à une détection informatique basée sur la densité en sites de liaisons pour différents TFs.

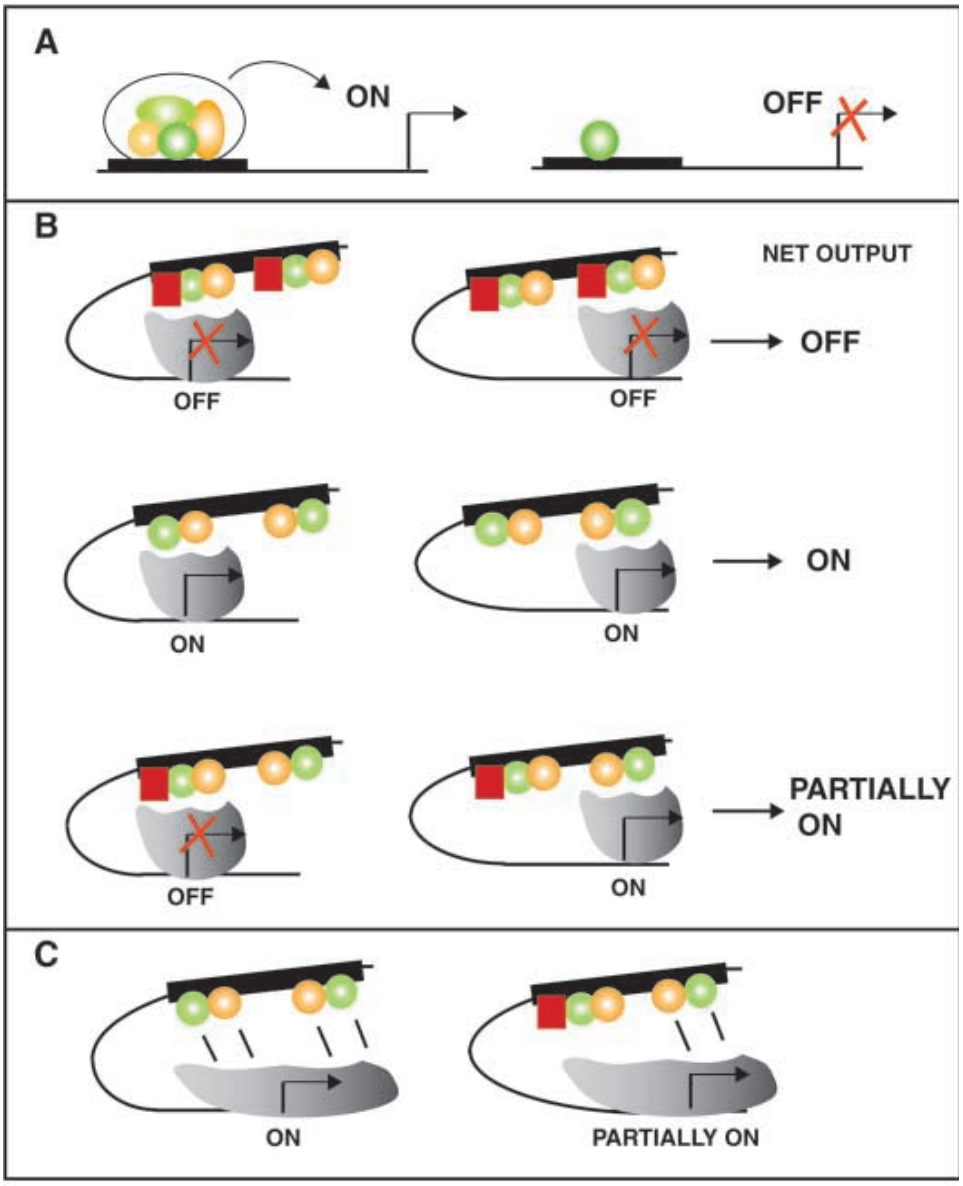


FIGURE 1.16 – **Deux modèles d'enhancers : enhanceosome et billboard.**
Figure tirée de (Kulkarni and Arnosti, 2003). (A) Dans le modèle enhanceosome, l'enhancer traite l'information des multiples TFs qui le fixent. Un complexe très structuré crée une interface qui recrute la machinerie de transcription basale. L'enhancer peut être vu comme un ordinateur moléculaire qui produit à partir d'entrées multiples un seul signal vers la machinerie de transcription. Le gène cible n'est activé qu'en cas de formation du complexe entier, ce qui fournit un interrupteur binaire on/off seulement activé en cas de stimulus adéquat. La déstabilisation du complexe en changeant par exemple la concentration d'une des protéines permettrait alors d'obtenir une réponse graduelle. (B,C) Modèle d'enhancer « billboard ». Dans ce cas, l'enhancer ne consiste pas en une seule unité de régulation, mais en des sous-unités pouvant contenir différentes informations (répression ou activation par exemple) que la machinerie basale échantillonne soit itérativement (B), soit simultanément (C).

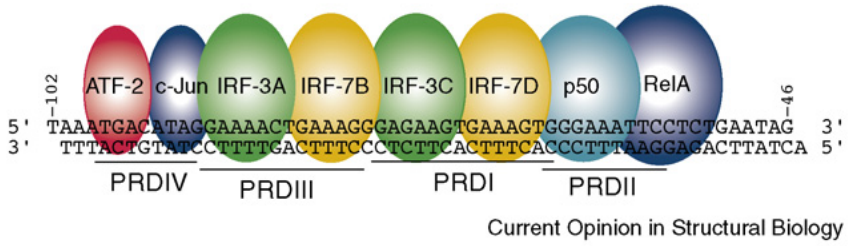


FIGURE 1.17 – **L'enhanceosome de l'interferon-β.**
Figure tirée de (Panne, 2008) représentant le complexe de TFs assemblés sur l'enhanceosome IFN-β, formant une surface de reconnaissance agissant comme une seule unité de régulation.

1.5.3 Évolution des enhancers

La fonction centrale que jouent les enhancers dans la régulation de l'expression génétique laisse à penser que ceux-ci seraient sous sélection et leur séquence serait donc plus conservée que celle des régions non codantes du génome. De fait, la comparaison de séquences non-codantes entre espèces proches s'avère être un mode de détection puissant des régions de régulation (Prabhakar et al., 2006). Ainsi, l'utilisation de la conservation entre des espèces lointaines comme l'homme et le poisson *Fugu* ou de l'extrême conservation entre des espèces proches comme l'homme, la souris et le rat, permet de détecter des régions ayant une activité enhancer *in vivo* avec un succès proche de 50% (Pennacchio et al., 2006). À l'instar de la régulation de l'interféron-β, de telles séquences très contraintes obéissent à une logique de type « enhanceosome » où la fonction est intimement liée à la séquence.

Contrastant avec cette vision d'enhancers très contraints, plusieurs études pointent vers une plus grande flexibilité des séquences enhancers (Ludwig et al., 2000; Dermitzakis and Clark, 2002; Moses et al., 2006). Supportant l'idée que la plupart des enhancers se comportent selon le modèle « billboard », la grammaire des sites de fixation dans des séquences orthologues apparaît comme étant loin d'être figée (Liberman and Stathopoulos, 2009). Ainsi, l'enhancer régulant le gène *short gastrulation* (*sog*), bien que présentant chez différentes espèces de *Drosophiles* une architecture variable des sites de fixation le composant, conduit à un même motif d'expression (fig. 1.18). Cette idée qu'une panoplie de grammaires conduisent à une même régulation est confortée par les résultats de Zinzen et al. (2009) où des enhancers ayant des « entrées » différentes (i.e étant fixés par des TFs différents pendant des durées variables) produisent des « sorties » similaires, dans ce cas une expression spécifique à un tissu donné.

Supportant l'idée d'une flexibilité de la régulation, plusieurs études ont exhibé l'évolution rapide des sites de liaison de TFs dans le génome (Wilson and Odom, 2009). Une étude de la fixation génomique des facteurs de transcription CEBPα et HNF4α dans les cellules du foie de 5 espèces de vertébrés (l'homme, deux espèces de souris, le chien et le poulet) a notamment montré que les événements de fixation conservés chez les 5 espèces sont très rares ($\sim$ 0.3% des pics humains) et correspondent à des régions ultraconservées proches de gènes importants dans la spécification du foie (Schmidt et al., 2010). Par ailleurs, lors de la perte de fixation dans l'une des espèce, un gain de fixation proche ($\pm$10kb) est observé dans la moitié des cas. Étonnamment, ces changements rapides du câblage du réseau affectent peu l'expression génétique globale (Tirosh et al., 2008; Odom et al., 2007).

Cette évolution est en grande partie due à une évolution de séquence de fixation. Ainsi,

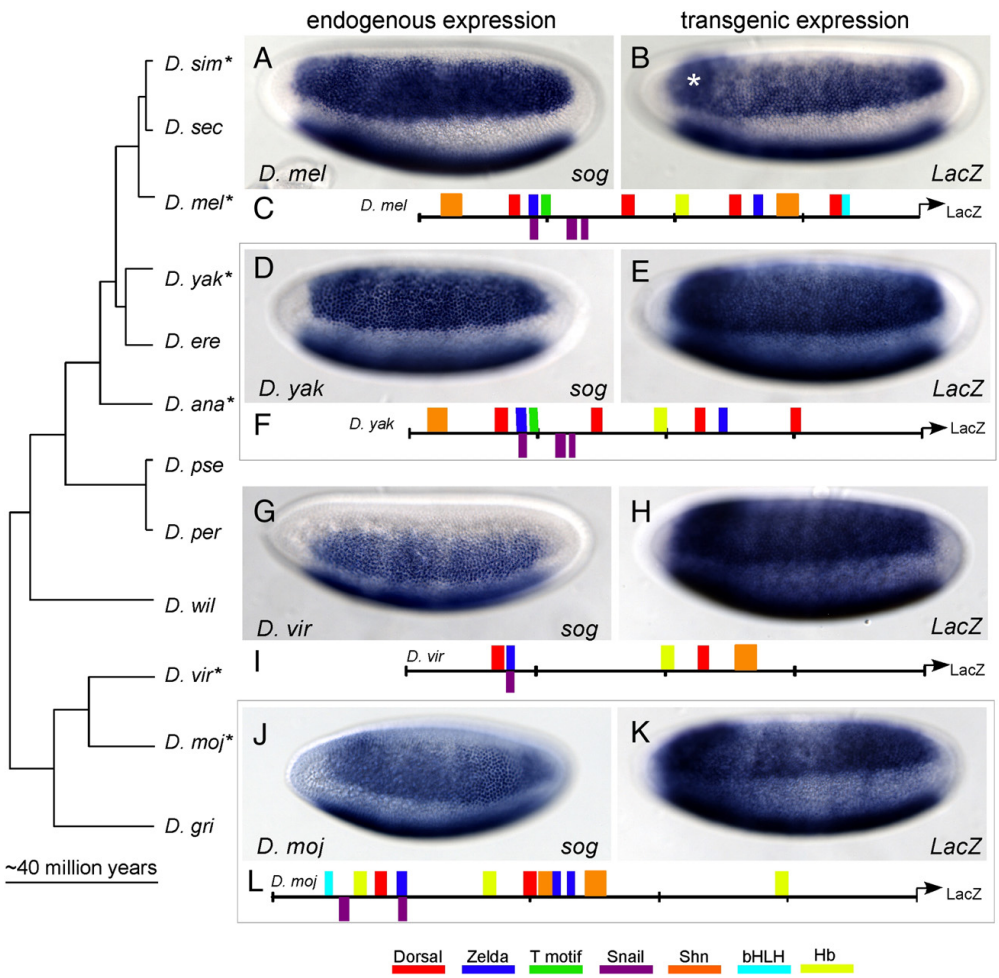


FIGURE 1.18 – **Flexibilité du code de cis-régulation au cours de l'évolution chez les *Drosophiles*.**
Figure tirée de (Liberman and Stathopoulos, 2009), montrant l'expression du gène *sog* par ISH chez plusieurs espèces de *Drosophilidæ* (A,D,G,J), ainsi que l'expression d'un rapporteur *LacZ* sous contrôle de la région de régulation de *sog* chez l'espèce considérée (B,E,H,K), dont une représentation schématique est donnée pour chaque cas (C,F,I,L). Les sites de fixation pour différents TFs sont indiqués par un rectangle de couleur.

une étude récente a utilisé une souris portant le chromosome 21 de l'homme pour comparer la fixation du facteur HNF4α dans un contexte murin par rapport au contexte original (Wilson et al., 2008). Le paysage de fixation sur le chromosome 21 exogène a très précisément récapitulé celui observé chez l'homme (fig. 1.19), montrant que le contexte cellulaire est sensiblement le même chez les deux espèces. Par ailleurs, des modifications épigénétiques ainsi que l'expression des ARNm ont pu être récapitulées.

Reste la question du mécanisme permettant cette évolution rapide. Une étude portant sur 7 facteurs de transcription chez les mammifères a montré qu'une proportion importante ($\sim$ 20%) des régions de fixation de ces TFs se situent au sein de différentes familles de transposons (Bourque et al., 2008) (fig. 1.19). Ces transposons, ou éléments transposables, sont des anciens rétrovirus intégrés dans les génomes mammifères ayant la capacité de se dupliquer pour

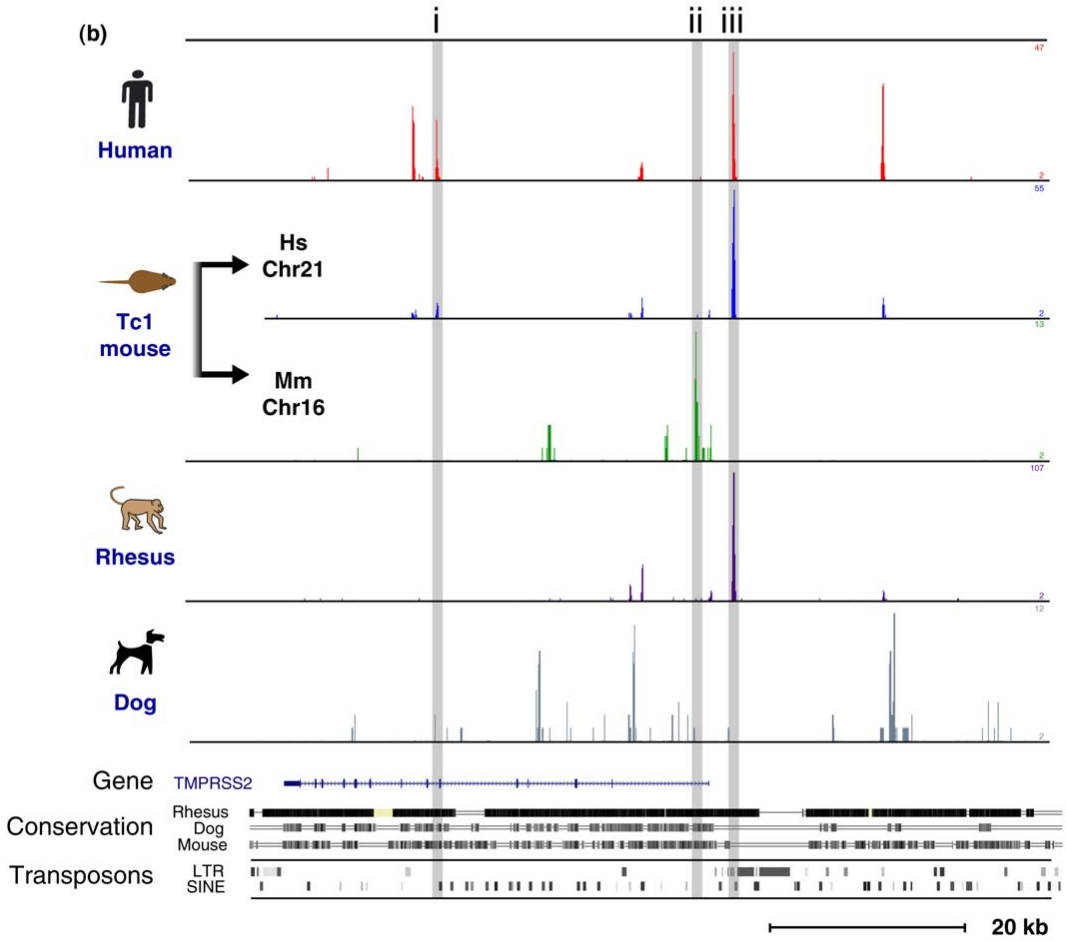


FIGURE 1.19 – **Évolution de la fixation de HNF4α chez les mammifères.**
Figure tirée de (Wilson and Odom, 2009), représentant la fixation par ChIP-seq (pics de couleur) du facteur de transcription humain HNF4α (ou son homologue murin HNF4a) chez l'homme, la souris, le macaque, le chien, ainsi qu'une souris transgénique contenant le chromosome 21 humain. Les zones grisées indiquent : (i) une fixation de HNF4α chez l'homme retrouvée sur le chromosome 21 humain de la souris mais pas chez le macaque malgré la proximité de séquence, (ii) une fixation de HNF4a spécifique à la souris, et (iii) une fixation spécifique aux primates qui a lieu sur des éléments transposables.

s'intégrer dans une autre région du génome et jouent un rôle fondamental dans l'évolution des génomes (Cordaux and Batzer, 2009). Leur accumulation dans le génome a vraisemblablement permis d'obtenir un matériau de base permettant de produire par mutations ponctuelles des éléments de régulation *de novo* (Feschotte, 2008). Par ailleurs, les transposons peuvent permettre de diffuser par « copier-coller » des éléments de régulation existant. Ainsi, des vagues d'expansion de transposons spécifiques à différentes espèces de mammifères sont à l'origine de la variabilité des régions de fixation observée dans le cas du facteur CTCF (Schmidt et al., 2012).

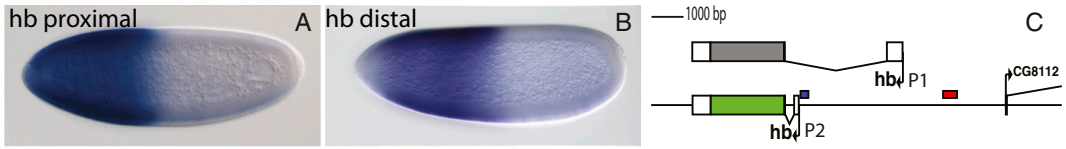


FIGURE 1.20 – **« Shadow enhancer » du gène de segmentation *Hunchback*.**
Figure tirée de (Perry et al., 2011), montrant des hybdridations *in situ* de transgènes contenant le gène rapporteur *LacZ* sous contrôle de l'enhancer proximal (A) ou du « shadow enhancer » plus distal (B) du gène de segmentation *Hb*. (C) Position des enhancers par rapport à *Hb* (bleu : proximal, rouge : distal).

1.5.4 Les « shadow enhancers »

L'évolution des éléments de cis-régulation est un mécanisme majeur permettant la diversité animale. Néanmoins, de tels changements pourraient compromettre certaines activités génétiques essentielles. Des expériences de ChIP-on-chip ont suggéré que plusieurs gènes de développement actifs lors du développement précoce de l'embryon de Drosophile possèdent des CRMs secondaires, qui conduisent à des motifs d'expression génétique comparables à ceux produits par des CRMs « primaires » plus proximaux (Zeitlinger et al., 2007). L'expression de « shadow enhancer » a été proposée par Michael Levine en 2008 pour décrire ces CRMs redondants et souvent distaux de plusieurs dizaines de kb du gène régulé (Hong et al., 2008). Il est probable que de tels CRMs soient apparus au cours de l'évolution par duplication du CRM primaire, à l'instar du phénomène de duplication des séquences codant pour des protéines. L'avantage évident que peut conférer la redondance d'un élément de régulation est d'offrir de la robustesse face aux mutations. Par ailleurs, une telle redondance permet de faciliter la divergence et donc la spécialisation des différents CRMs. Ainsi les « shadow enhancers » semblent évoluer plus rapidement que les CRMs primaires auxquels ils sont apparentés (Hong et al., 2008) pour fournir de nouveaux sites de fixation et conduire à de nouvelles activités de régulation sans bloquer la fonction critique de certains gènes de développement.

Un exemple mêlant robustesse et divergence est le cas des multiples CRMs régulant le gène *Svb* chez la Drosophile. Chaque CRM est lié à la production d'un motif distinct de trichomes (excroissances de l'épithélium comparables à des poils) sur la larve : ainsi, plusieurs mutations dans ces différents CRMs sont nécessaires pour observer un changement morphologique conséquent (McGregor et al., 2007). Dans ce même système, il a été montré que deux CRMs supplémentaires, des « shadow enhancers », sont dispensables dans des conditions de température usuelles, mais requis lorsque les embryons se développent dans des conditions de température extrêmes (Frankel et al., 2010).

Par ailleurs, il a été montré que les gènes de segmentation (ou gènes *gap*) de la Drosophile possèdent tous des « shadow enhancers » (fig. 1.20). Leur rôle semble être d'assurer une plus grande précision spatiale du motif d'expression du gène régulé : la perte de l'un des CRMs, proximal aussi bien que « shadow », conduisant à une expression trop restreinte ou trop répandue spatialement selon le cas (Perry et al., 2011).

1.5.5 Par delà les enhancers : les « super-enhancers »

Récemment, il a été montré que certains groupements d'enhancers peuvent agir comme une même unité de régulation : on parle de *super-enhancers* (Whyte et al., 2013). Ces régions

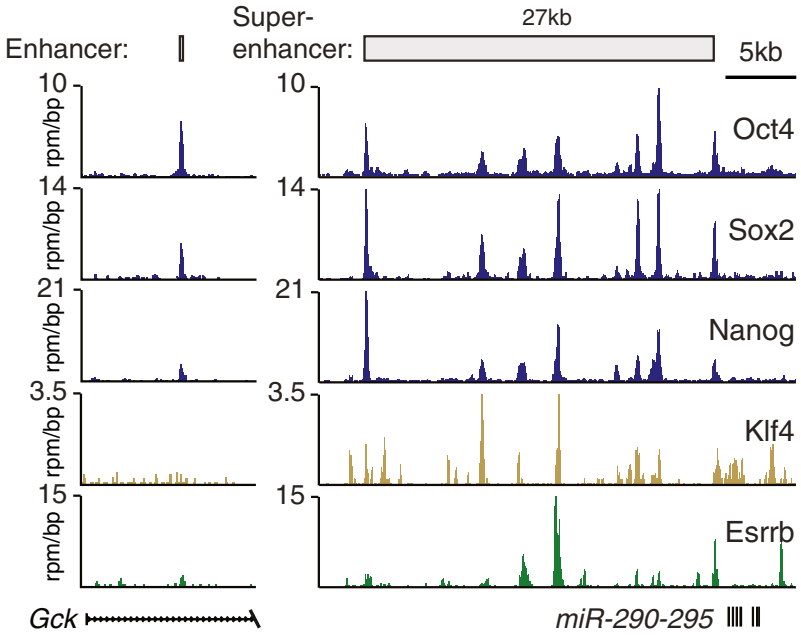


FIGURE 1.21 – **De l'enhancer au super-enhancer.**
Figure tirée de (Whyte et al., 2013), montrant les profils de ChIP-seq des TFs maîtres Oct4, Sox2, Nanog, Klf4 et Esrrb aux loci de *Gck* et *miR-290-295* dans les cellules souches embryonnaires. Le super-enhancer se distingue du simple enhancer par sa taille (27kb), sa grande concentration en TFs maîtres, notamment Klf4 et Esrrb, et la fixation de la protéine Med1 du complexe Mediator.

de taille typique $\sim$ 10kb (fig. 1.21), sont fixées par des TFs maîtres et sont associées à des gènes encodant des régulateurs clés de l'identité cellulaire. Identifiés dans les cellules souches embryonnaires (ESCs), ces ensembles d'enhancers sont fixés par le complexe co-activateur Mediator, qui interagit avec la cohésine pour former un anneau permettant de connecter la région de régulation au promoteur (Kagey et al., 2010). Par ailleurs, les gènes associés aux super-enhancers possèdent un niveau particulièrement élevé d'expression et leur knock-down est associée à une perte de l'état souche des cellules.

Ainsi, ce second niveau d'organisation de la régulation pourrait simplifier la modélisation de la régulation du type cellulaire, en passant de millier de traces de fixation pour différents TFs à quelques centaines de super-enhancers contrôlant les gènes clés de l'identité cellulaire.

1.6 Prédiction et validation des CRMs

1.6.1 Méthodes utilisant la concentration en sites de fixation

Nous l'avons vu, une propriété des CRMs est leur grande concentration en TFBS. Ceci a motivé des approches de prédiction de promoteurs et d'enhancers basées sur leur contenu ou *clustering* en motif (fig. 1.22a). L'avantage de telles approches est qu'elles peuvent être réalisées avec seulement la séquence d'ADN génomique et des modèles de TFs ou motifs (par exemple des PWMs, voir fig. 1.10) représentant les facteurs de transcription impliqués

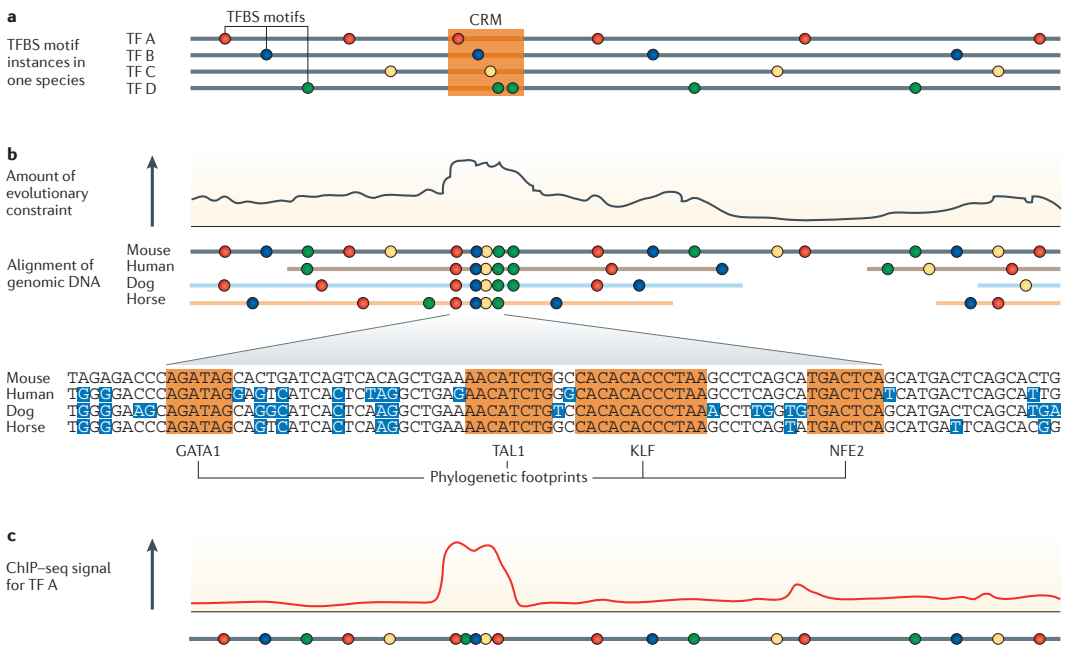


FIGURE 1.22 – **Différentes approches pour la prédiction des CRMs.**
Figure tirée de (Hartwell et al., 1999), résumant les différentes approches de prédiction des CRMs. (a) Prédiction de sites sites de fixations (TFBS) pour différents facteurs de transcription. Le regroupement ou *clustering* de TFBS définit un CRM putatif. (b) Le CRM, possédant des contraintes fonctionnelles, est conservé chez plusieurs espèces proches. La séquence d'ADN correspondant au CRM (un enhancer connu de la β-globine) est montrée chez les différentes espèces. Les nucléotides qui diffèrent par rapport à la séquence de référence (la souris) sont indiqués en bleu, et les sites de fixation de TFs connus ou prédits sont indiqués en orange. Ceux-ci correspondent à des séquences complètement conservées : on parle d'empreinte phylogénétique ou *phylogenetic footprint*. (c) Une expérience de ChIP-seq spécifique au TF A produit un pic au niveau du CRM. Il est à noter que des pics peuvent exister même si aucun site de fixation n'est prédit, dû par exemple à une interaction protéine-protéine entre le TF A et un autre TF fixé à l'ADN.

dans le processus étudié. Cependant, les clusters de motifs sont très répandus dans les grands génomes, et sans l'ajout d'informations supplémentaires comme les marques épigénétiques ou l'expression des gènes voisins, ces approches produisent un grand nombre de faux positifs (éléments prédits comme positifs mais étant en réalité négatifs). Par ailleurs, les TFs impliqués ne sont pas toujours connus, et il faut alors apprendre des motifs putatifs à partir de séquences fonctionnelles.

- **Approches utilisant des motifs connus**

L'une des premières investigations basée sur le regroupement de TFBS utilisait 5 motifs connus de la détermination musculaire pour prédire par régression linéaire les CRMs actifs dans le muscle (Wasserman and Fickett, 1998). Le taux de validation était relativement bas, autour de 20%. De même, chez *Drosophila melanogaster*, plusieurs études ont utilisé le clustering de motifs pour prédire des CRMs de différents processus développementaux (par ex Berman et al. (2002)). Ces études ont trouvé de nouveaux enhancers validés expérimentalement (bonne sensibilité) mais avaient des taux de prédiction relativement bas, entre 15 et 30%.

L'algorithme *Ahab* (Rajewsky et al., 2002), utilisant un modèle thermodynamique de fixation des TFs sur les CRMs, a quant à lui réussi à prédire un nombre bien plus important de régions fonctionnelles : $\sim 80\%$ des modules prédits à proximité de 29 gènes de segmentation chez la drosophile ont effectivement récapitulé le motif d'expression du gène associé (Schroeder et al., 2004). Ce succès semble notamment être dû au fait que ce modèle thermodynamique, basé sur une prise en compte exhaustive de toutes les segmentations possibles des CRMs en motifs et en ADN « background », permet de donner plus de poids au cas où plusieurs sites de faibles affinité pour un TF se trouvent au sein d'un même module, alors que les autres méthodes utilisent généralement un seuil de probabilité relativement élevé (afin d'éviter les faux positifs) à partir duquel une séquence est considérée comme fixée par un TF. Par ailleurs, cette étude s'est restreinte à un ensemble de gènes connus pour lesquels les régions à proximité riches en TFBS ont *a priori* plus de chances d'être fonctionnelles. De manière générale, plus le domaine de recherche est étendu (par exemple, le génome entier), plus le nombre de faux positifs augmente.

- **Approches *de novo* où les motifs ne sont pas connus**

Lorsque les motifs (PWMs) ne sont pas connus à l'avance, il faut les générer *de novo* à partir de leur surreprésentation dans des CRMs connus. Par exemple, l'algorithme CisModule permet de générer des motifs et des modules simultanément en utilisant un modèle de mélange hiérarchique (Zhou and Wong, 2004). Lorsqu'il est appliqué aux CRMs musculaires introduits précédemment, il permet de retrouver certains motifs connus et permet de retrouver $\sim 70 - 80\%$ des séquences connues lorsqu'elle sont mélangées avec un nombre similaire de séquences aléatoires. Par ailleurs, l'apprentissage de modèles permettant de discriminer différentes classes de CRMs entre elles plutôt qu'une classe de CRMs par rapport à des séquences aléatoires ou intergéniques peut s'avérer plus fructueux. Ainsi, (Smith et al., 2006) ont utilisé des motifs connus ainsi que des motifs appris *de novo* avec le programme DME (Smith et al., 2005) pour leur capacité à discriminer des séquences appartenant à différents jeux de données de régions promotrices pour bâtir un modèle de régression logistique permettant de prédire l'activité tissu-spécifique dans 45 des 56 tissus humains et murins considérés. Il existe aussi plusieurs méthodes qui n'utilisent pas de motifs du type PWM, mais de purs modèles probabilistes tels que des chaînes de Markov d'ordre 5 ou des regroupements de « mots » de k nucléotides ou k-mers selon des critères de distance de Hamming et surreprésentés dans les séquences d'intérêt, par exemple (Cao et al., 2010). Ces méthodes sont passées en revue dans (Kantorovitz et al., 2009), et elles peuvent atteindre des sensibilités de $\sim 60\%$ pour la prédiction de CRMs mammifères. L'intérêt est que ces études ne présument pas d'un modèle de fixation des TFs à l'ADN. C'est aussi un désavantage, puisqu'elles sont moins informatives quant au réseau génétique sous-jacent et aux mécanismes de régulation impliqués.

1.6.2 Méthodes utilisant la phylogénie

Les approches utilisant la comparaison des génomes de différentes espèces pour prédire des CRMs sont basées sur l'idée que les séquences de régulation sont plus fortement conservées que l'ADN non fonctionnel les entourant. Nous l'avons vu en 1.5.3, une proportion importante de CRMs ne satisfont pas à cette règle. Cette approche ne permet donc d'étudier que le sous-ensemble de CRMs qui a subi une forte pression de sélection depuis le dernier ancêtre commun aux espèces considérées et ne donne pas accès aux CRMs apparus récemment au sein d'une espèce.

- **Prédictions à partir de la contrainte évolutive seule**

L'alignement de séquences non-codantes orthologues fait apparaître des parties très conservées, avec peu de variations dans les séquences sous-jacentes, entourées de séquences accumulant les variations (fig. 1.22b). De telles séquences conservées sont alors interprétées comme ayant été sous sélection, les substitutions délétères ayant été rejetées au cours de l'évolution (Dermitzakis et al., 2005). Par analogie avec les empreintes à la DNAse I, on parle d'empreinte phylogénétique pour caractériser ces courtes séquences très conservées ($\sim$ 10bp), traces de la fixation putative d'un facteur de transcription. Ces empreintes s'avèrent être un indicateur fiable de fonctionnalité (Kheradpour et al., 2007) et, parce qu'elles ne reposent pas sur des modèles *a priori* de fixation, elles permettent de plus de trouver des motifs de régulation non connus (Xie et al., 2005). Au niveau de séquences plus longues ($\sim$ 100bp), la contrainte évolutive permet de détecter des CRMs entiers. Ainsi, comme nous l'avons vu en 1.5.3, l'utilisation de la conservation extrême permet d'atteindre 50% de taux validation (Pennacchio et al., 2006). Néanmoins, lorsque ces contraintes de conservation extrême (par exemple homme-*Fugu*) sont relâchées, le taux de validation tombe drastiquement, atteignant $\sim$ 5% (Attanasio et al., 2008), montrant la nécessité d'allier le critère de conservation à d'autres données (expression, ChIP...) pour améliorer la prédiction des CRMs.

- **Prédictions utilisant la phylogénie et des motifs connus**

Une approche pour améliorer les prédictions est de combiner les approches précédentes en utilisant à la fois le *clustering* en TFBS et la contrainte évolutive. À l'échelle du génome entier, cette approche permet de filtrer les résultats pour améliorer le signal de détection chez la Drosophile (Sinha et al., 2004). Du côté des mammifères, en utilisant les motifs de la base de données TRANSFAC et la conservation entre l'homme et la souris, Blanchette et al. (2006) ont créé une base de données de modules, PReMods, qui retrouve $\sim$ 17% de CRMs connus et recoupe 40% des fragments occupés par le co-activateur et marqueur de l'activité enhancer p300. D'autres méthodes se sont concentrées sur des types cellulaires bien définis. Par exemple, la recherche de sites conservés pour des motifs de TF des cellules sanguines connus (Donaldson et al., 2005) a permis de définir des CRMs dont 2 ont été testés et validés.

Certains efforts ont par ailleurs été menés pour sortir du cadre d'une conservation de séquence stricte en modélisant l'évolution d'un CRM fixé par un certain nombre de motifs connus. Par exemple, le modèle MorphMS (Sinha and He, 2007) cherche au sein d'un alignement de deux séquences orthologues des régions prédites par un modèle d'évolution dérivé d'un ensemble de motifs choisis par l'utilisateur. Une extension de cette approche incorpore le gain et la perte de sites de fixation, mais n'a cependant pas encore été appliquée à l'échelle du génome (Majoros and Ohler, 2010).

- **Approches utilisant la phylogénie pour générer des motifs *de novo***

De même que précédemment, tous les motifs ne sont pas connus et il peut être utile d'avoir recours à de l'apprentissage direct à partir de séquences fonctionnelles connues pour aider à la prédiction. Par exemple, l'algorithme ESPER cherche des patterns (TFBS, %GC, etc) surreprésentés dans des alignements multi-espèces de CRMs connus par rapport à des alignements d'ADN *a priori* non fonctionnel (Taylor et al., 2006). Cette méthode n'est pas restreinte à l'analyse de séquences conservées puisqu'elle peut potentiellement capturer des signatures de changements systématiques. La prédiction de régions de haut potentiel de régulation recouvre presque entièrement les prédictions de PReMods, et le test par transfection de ces régions à proximité de gènes exprimés dans les cellules érythroïdes et possédant un site pour un TF spécifique de l'érythroïde mène à un taux de validation de 50%. Une autre méthode consiste

à chercher des mots surreprésentés dans un ensemble d'apprentissage de CRMs connus puis à restreindre les prédictions aux régions conservées (Kantorovitz et al., 2009). Les prédictions réalisées ont toutes été validées chez la Drosophile (5/5) comme chez la souris (2/2).

1.6.3 Méthodes utilisant les marques épigénétiques et de ChIP-seq pour des TFs

- **Prédiction des promoteurs**

La méthode la plus fiable de prédiction d'un promoteur utilise le fait qu'il est toujours localisé au niveau d'un TSS, dont la position peut facilement être obtenue en alignant les séquences de l'ARN du gène correspondant sur le génome (Trinklein et al., 2003). Le taux de validation avec cette seule contrainte est très élevé : 91% ont une activité dans au moins un type cellulaire. Par ailleurs, la marque épigénétique H3K4me3 est aussi un indicateur des promoteurs actifs dans le type cellulaire étudié (Heintzman et al., 2007) (fig. 1.14).

- **Prédiction des enhancers**

La prédiction des enhancers à partir des marques épigénétiques, comme l'acétylation des histones (Roh et al., 2005), la méthylation H3K4me1 (Heintzman et al., 2009), ou encore la présence du co-activateur p300 (Visel et al., 2009a), est très efficace, avec une expression tissue-spécifique dans $\sim$ 80% des cas (Hardison and Taylor, 2012). Par exemple, ces différentes marques, présentes dans différents tissus, peuvent être utilisées comme autant d'entrées d'un modèle de Markov caché pour produire des prédictions fiables de CRMs tissu-spécifiques chez l'homme (Ernst et al., 2011).

En fait, les prédictions d'activité enhancer à partir de ces marques épigénétiques est plus fiable qu'en utilisant la fixation de facteurs de transcription tissu-spécifiques. Par exemple, sur 63 séquences ADN fixées *in vivo* par le facteur spécifique des cellules sanguines GATA1 chez la souris, seulement la moitié conduisent à une activité après transfection dans des cultures cellulaires (Cheng et al., 2008). Ces enhancers fonctionnels sont par ailleurs plus particulièrement associés à un site de fixation *conservé* pour GATA1, montrant à nouveau la nécessité de combiner les approches pour améliorer la détection. Un taux de validation similaire a été observé pour le facteur de différenciation myogénique MyoD, avec 40% de régions fixées ayant une activité après transfection en cellules.

L'utilisation de données de fixation pour plusieurs TFs à la fois semble cependant améliorer le pouvoir de prédiction. Ainsi, Tijssen et al. (2011) ont étudié la co-fixation de GATA1 avec 4 autres TFs hématopoïétiques dans des mégacaryocytes. En s'intéressant aux gènes à proximité de ces régions, ils en ont découvert plusieurs qui n'étaient pas précédemment connus comme étant important dans l'hématopoïèse. Leur fonction a été testée par knock-down, avec dans 8 cas sur les 9 testés une réduction de la production de globules rouges.

1.6.4 Validation expérimentale

Une méthode directe permettant de démontrer qu'un fragment d'ADN régule l'expression génétique consiste en une expérience de gain de fonction dans laquelle un plasmide contenant le CRM prédit à proximité d'un gène rapporteur est introduit par transfection *in vitro* en cellule, permettant un suivi quantitatif de l'activité, ou par transgenèse *in vivo* dans un organisme, auquel cas le suivi est plus qualitatif mais permet d'établir la spécificité spatio-temporelle (tissu et stade de développement) de l'élément de régulation (fig. 1.23). Ce type d'expérience montre que le CRM prédit est *suffisant* pour reproduire le motif génétique ob-

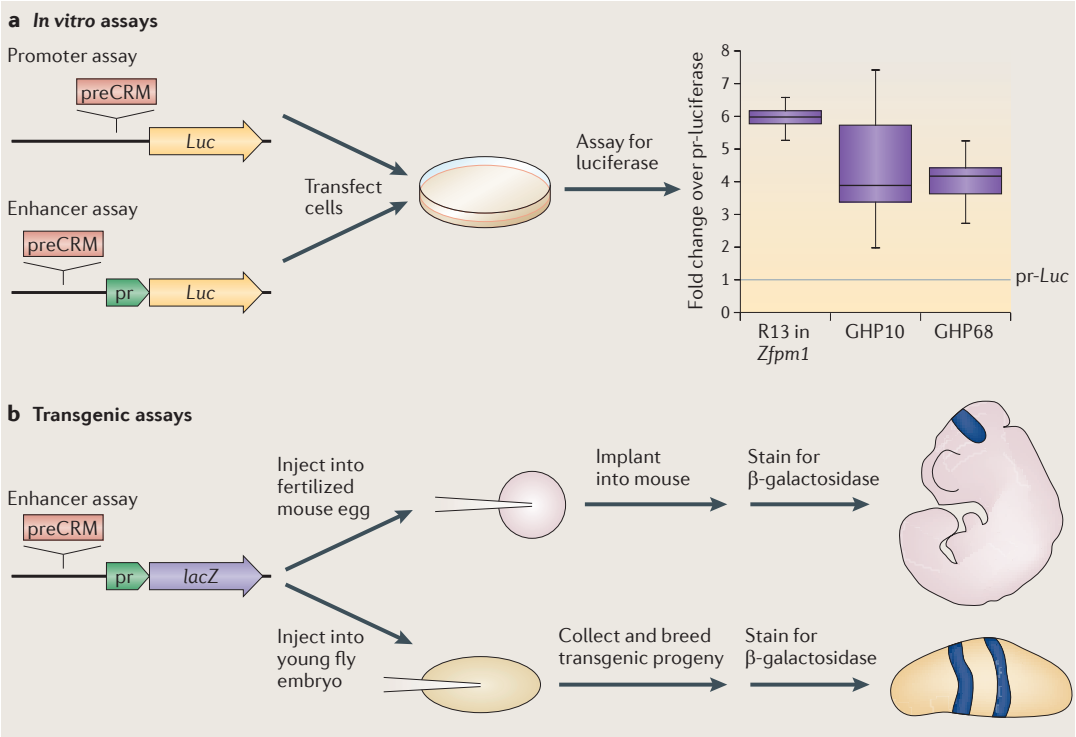


FIGURE 1.23 – **Méthodes de validation des CRMs par transfection et transgenèse.** Figure tirée de (Hartwell et al., 1999) présentant les méthodes *in vitro* et *in vivo* de validation des CRMs. La région dont on souhaite tester l'activité est insérée dans un plasmide codant pour un gène rapporteur qui est transféré dans une culture cellulaire (transfection, panel a) ou dans un organisme entier (transgenèse, panel b). Dans le cas du test d'un promoteur, le CRM est placé directement en amont du gène rapporteur (on utilise généralement la luciférase *Luc*), alors que dans le cas d'un enhancer, le CRM est placé en amont d'un promoteur minimal de faible activité. L'activité de la luciférase donne une information quantitative sur l'activité de la région testée (boîtes à moustache, panel a). Dans le cas d'une transgenèse, le gène rapporteur généralement utilisé est *lacZ* qui encode la β-galactosidase. La révélation par coloration permet de visualiser en bleu les tissus au sein lesquels l'enhancer est actif.

servé. De manière optimale, il faudrait aussi montrer par délétion ciblée de l'élément de régulation au sein du génome que ce dernier est *nécessaire* à l'expression du gène endogène.

1.6.5 Implication des CRMs dans les maladies humaines

Au cours des dernières décennies, de nombreuses mutations dans les régions codantes des gènes, impliquant des défauts structurels des protéines associées, ont pu être associées à des maladies génétiques. À l'inverse, le rôle des mutations affectant des régions non codantes n'a été que peu exploré, essentiellement du fait de la difficulté d'annoter ces régions correctement afin de définir celles qui pourraient avoir une fonction d'intérêt. Plusieurs études ont cependant pu montrer que des variations affectant des enhancers distaux pouvaient conduire à des pathologies (Visel et al., 2009b).

L'une de ces études concerne l'enhancer spécifique du membre de *Shh* (fig. 1.24). Cet enhancer, initialement décrit chez la souris, se situe à environ 1 Mb de distance de *Shh*, au sein

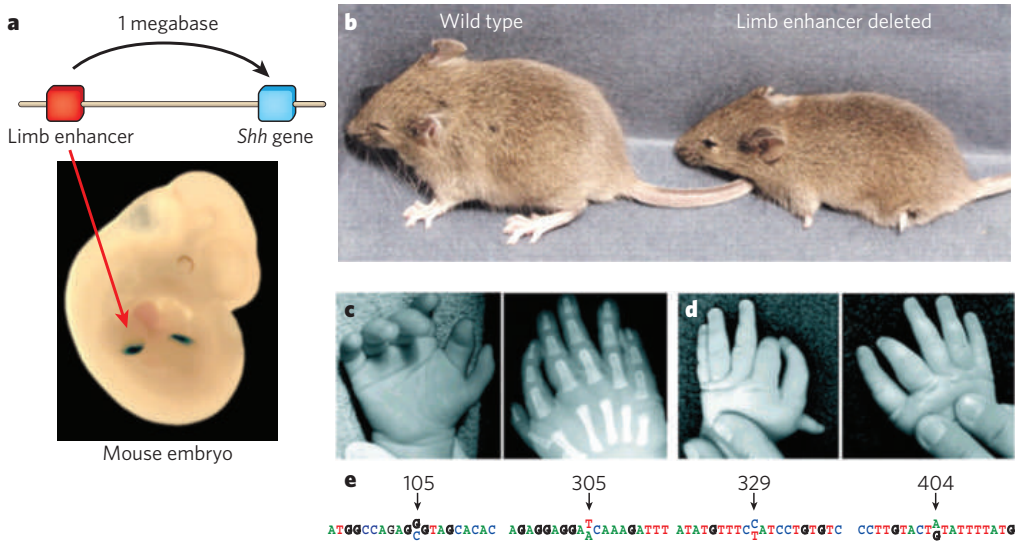


FIGURE 1.24 – **Impact physiologique de la délétion et de la mutation d'un enhancer.** Figure tirée de (Visel et al., 2009b). (a) L'enhancer « membre » de *Sonic hedgehog* (*Shh*) est situé à $\sim$ 1 Mb de son gène cible, au sein d'un intron d'un gène voisin. Par transgenèse il conduit l'expression d'un rapporteur dans la partie postérieure du bourgeon de membre chez l'embryon de souris (Lettice et al., 2003). (b) Les souris possédant une délétion ciblée de cet enhancer possèdent des membres tronqués, montrant son importance fonctionnelle au cours du développement (Sagai et al., 2005). (c-e) Différentes mutations ponctuelles dans la séquence enhancer orthologue chez l'homme résultent en une polydactilie pré axiale (présence de doigts ou orteils supplémentaires), montrant l'impact potentiel de variations dans des régions non-codantes (Lettice et al., 2003).

de l'intron d'un gène voisin. Le séquençage de cet enhancer chez plusieurs individus humains a permis d'associer une douzaine de variations mono-nucléotidiques à la polydactilis pré axiale, c'est-à-dire la présence de doigts ou d'orteils supplémentaires (Lettice et al., 2003). Des études supplémentaires chez la souris ont montré que les variations de séquences observées dans cet enhancer conduisent à une expression ectopique dans la partie antérieure du membre au cours du développement, ce qui est consistent avec la présence de doigts supplémentaires (Masuya et al., 2007). Par ailleurs, la délétion de l'enhancer orthologue de la souris entraîne la troncation des membres (Sagai et al., 2005).

Ainsi, ces résultats montrent l'importance de l'identification des enhancers pour permettre à des études de génétique humaine d'explorer le rôle potentiellement pathologique de mutations dans des régions non codantes fonctionnelles.

1.7 Bases de données

La biologie moderne est caractérisée par l'accumulation de données biologiques qu'il s'agit d'intégrer puis d'interpréter : on parle de biologie integrative. En particulier, depuis le séquençage du génome humain il y a maintenant plus de dix ans (Lander et al., 2001), le nombre de génome séquencés n'a cessé d'augmenter, tandis que dans le même temps le prix du séquençage diminuait drastiquement (fig. 1.25). Afin de permettre la gestion et l'utilisation de ces

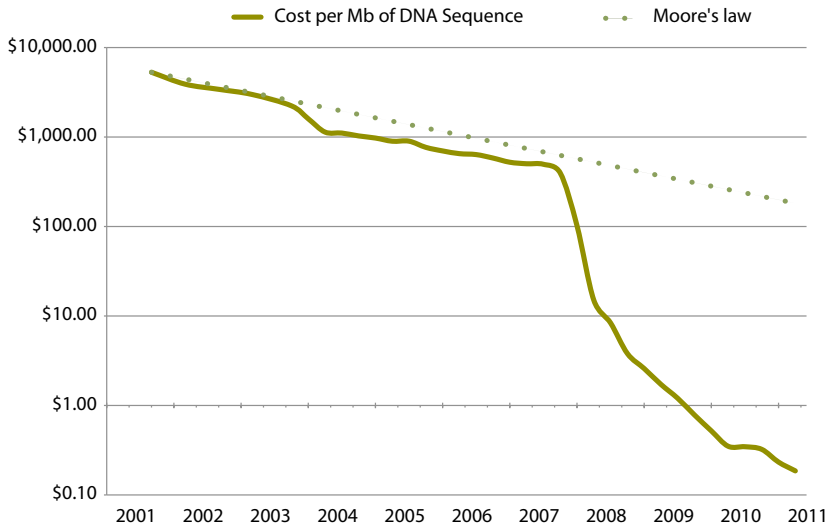


FIGURE 1.25 – **Évolution du coût de séquençage.**
Figure adaptée de (Sboner et al., 2011), montrant l'évolution du coût du séquençage d'1 Mb d'ADN au cours de la dernière décennie, comparé à une évolution de type « Loi de Moore » où le prix serait diminué de moitié tous les 18 mois.

données, de nombreux outils et bases de données ont été mis à disposition (Wasserman and Sandelin, 2004). Nous évoquons ici ceux qui nous paraissent essentiels du point de vue de la régulation en *cis*.

1.7.1 Obtention de données génomiques

Tout d'abord, les différents génomes séquencés sont à disposition sur des bases de données publiques d'où ils peuvent être téléchargés puis analysés en aval. Parmi les plus généralistes se trouvent la base de donnée de UCSC (UCSC Genome Browser, `http://genome.ucsc.edu`) et celle de l'EMBL (Ensembl, `http://www.ensembl.org`) [4].

Sont à disposition les génomes des différentes espèces séquencées pour les différents assemblages réalisés, des alignements des génomes de différentes espèces deux par deux (*pairwise alignments*) ou par groupes d'espèces (*multiple alignments*), ainsi qu'un certain nombre d'annotations essentielles à l'analyse de ces génomes : coordonnées des gènes (TSSs, exons, introns avec potentiellement différents transcrits alternatifs), miRNA ou lincRNA, ontologies associées, coordonnées des séquences répétitives (les *repeats*, en partie liés aux éléments transposables abordés en 1.5.3, et qui sont abondants dans les génomes vertébrés), différentes données ChIP-seq, indices de conservation [5]...

Au final, ces différentes données constituent une base de travail fiable et régulièrement mise à jour. Afin de faciliter leur obtention, il est possible d'utiliser le navigateur de tables de

4. Les données sont accessibles sur les pages de téléchargement, respectivement `http://hgdownload.cse.ucsc.edu/downloads.html` pour UCSC et `http://www.ensembl.org/info/data/ftp/index.html` pour Ensembl

5. Pour le cas de l'assemblage mm9 de la souris, ces annotations sont accessibles à l'adresse suivante : `http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/`

UCSC[6] ou la section BioMart d'Ensembl[7] .

Situé plus en amont, le projet Galaxy (http://galaxyproject.org) permet à l'utilisateur de récupérer des données depuis les différentes banques existantes, puis de leur faire subir divers traitements et analyses par divers outils de bioinformatique. Cet outil, qui peut être utilisé sur internet ou bien localement, a l'avantage de permettre la sauvegarde de plans de travail ou *workflows*, successions de commandes utilisées pour traiter une entrée donnée par différents outils stéréotypés et obtenir directement le résultat final, favorisant une approche conviviale orientée utilisateur.

En guise d'exemple, nous montrons en annexe A des statistiques obtenues aisément à partir d'annotations génétiques présentes sur UCSC et traitées avec Galaxy. Ces statistiques sont les distribution de tailles des régions intergéniques et introniques chez plusieurs espèces : la bactérie *Escherichia coli*, la levure *Saccharomyces cerevisiae*, le ver *Caenorhabditis elegans*, la mouche *Drosophila melanogaster*, la souris, le poulet et l'homme (fig. A.1).

1.7.2 Obtention de données sur les TFs

Nous l'avons vu, les données de fixation des TFs (ChIP-seq, ChIP-on-chip) peuvent être obtenues à partir du site UCSC Genome Browser. Ces données sont aussi généralement accessibles sur le site du NCBI (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) *via* un numéro d'accession donné lors de la publication des données.

De nombreux modèles de TFs ont déjà été bâtis préalablement à l'avènement des données haut-débit de type ChIP-seq, par exemple avec des données SELEX, et il existe des bases de données stockant les PWMs correspondantes : JAPSAR, base de donnée publique[8], et TRANSFAC, qui marche par abonnement[9]. Il est à noter que ces PWMs ayant souvent été construites à partir d'un faible nombre de sites de fixations et de données *in vitro*, elles peuvent être relativement inadaptées à l'analyse de données *in vivo*.

1.7.3 Outils de visualisation

Afin d'avoir une idée plus claire des événements de régulation qui se déroulent à un locus donné, il existe plusieurs outils de visualisation des annotations génomiques et épigénétiques, que ce soit sur le site du NCBI (http://www.ncbi.nlm.nih.gov/gene), sur Ensembl ou sur UCSC Genome Browser. Ce dernier possède notamment l'avantage qu'il est possible d'importer des données personnelles sous un grand nombre de formats, obtenues à partir de la littérature ou à partir de ses propres travaux. Ainsi, nous présentons en figure 1.26 quelques données de ChIP-seq pour des TFs musculaires et pour des marques épigénétiques, ainsi que des prédictions bioinformatiques de sites de fixation conservés pour les homéoprotéines Six réalisée par nos soins. La visualisation sur UCSC Genome Browser permet de rapidement déterminer le mode de régulation putatif du gène *Chrng* : fixation de Six et MyoD au niveau du promoteur et apparition de marques épigénétiques H3K4me1 et H3Ac sur les histones au cours de la différenciation de progéniteurs musculaires.

Par ailleurs, il existe un outil de visualisation complémentaire de ceux cités : le visualisateur de régions conservées au cours de l'évolution ECR Browser (http://ecrbrowser.dcode.

6. http://genome.ucsc.edu/cgi-bin/hgTables
7. http://www.ensembl.org/biomart/martview
8. http://jaspar.cgb.ki.se
9. http://www.gene-regulation.com/pub/databases.html

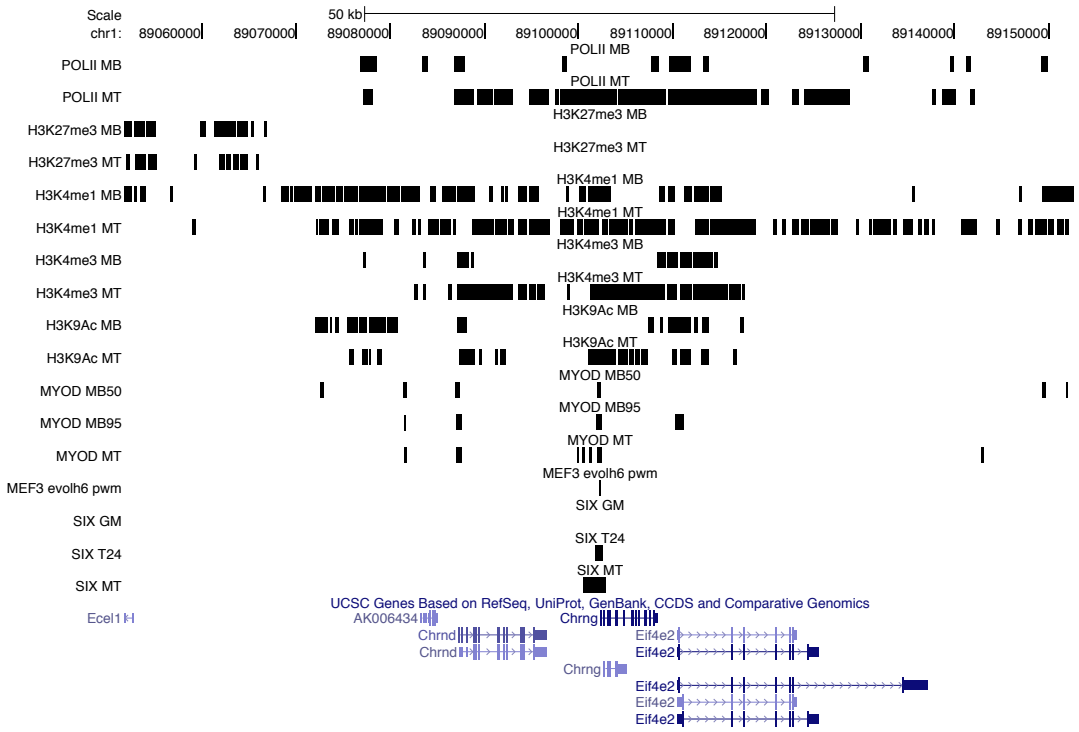


FIGURE 1.26 – **Visualisation de données ChIP-seq *via* le site UCSC.**
La visualisation de différentes données ChIP-seq et bioinformatiques (bandes noires) permet de mettre en perspective le cas de la régulation de *Chrng* (en bleu au bas de l'image) lors de la différenciation musculaire. Les données ChIP-seq sont issues de la littérature et les données bioinformatiques (MEF3) ont été obtenues au cours de cette thèse. Les rectangles noirs correspondent aux coordonnées des pics obtenus après avoir appliqué un seuil de filtrage du bruit. Les données de fixation de PolII ainsi que les données de méthylation et d'acétylation des histones (H3K4me3 et H3K9Ac, marques de l'activité transcriptionnelle, voir fig. 1.14), tirées de Asp et al. (2011), indiquent que le locus est transcrit lors de la formation de myotubes. Le TF MyoD est fixé au niveau du promoteur de *Chrng* au cours de la prolifération des myoblastes (MB50, 50% de confluence, et MB95, 95% de confluence) et au cours de la différenciation en myotubes MT (Cao et al., 2010). Le TF Six est co-fixé avec MyoD lors de la différenciation : à T24, 24h après différenciation, et à MT (Liu et al., 2010). De plus, des analyses bioinformatiques montre l'existence dans cette région d'un site de fixation MEF3 pour la protéine Six conservé chez les vertébrés, corroborant une liaison directe de l'ADN par Six. Prises ensemble, la simple visualisation de ces données suggèrent une régulation par Six et MyoD de *Chrng*.

org), intégrant de nombreux outils bioinformatiques (Loots and Ovcharenko, 2005). Ce navigateur permet de visualiser la conservation génomique d'un locus donné chez plusieurs espèces plus ou moins lointaines (par exemple souris, homme, vache, grenouille et poisson zèbre) afin de cibler l'étude de la régulation sur des régions extrêmement conservées. Il est ensuite possible d'analyser les séquences ultraconservées sélectionnées en utilisant les motifs de la base de donnée TRANSFAC *via* l'outil rVISTA (Loots and Ovcharenko, 2004). Un exemple d'utilisation de cet outil est donné par la découverte de plusieurs régions de régulation fonctionnelles de l'homéoprotéine Six1 possédant une extrême conservation (Sato et al., 2012).

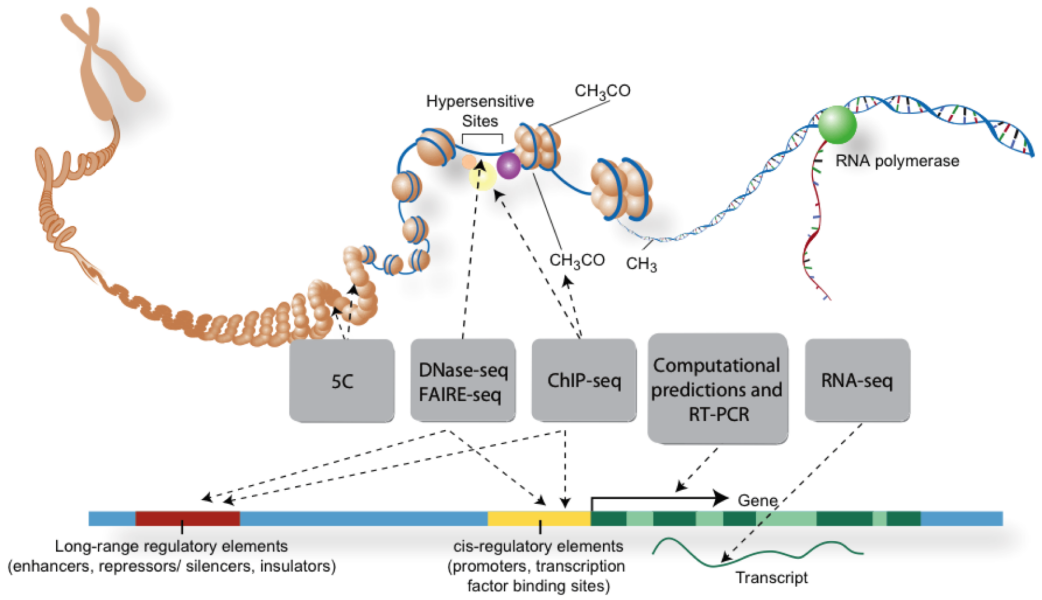


FIGURE 1.27 – **Les différentes données obtenues par le projet ENCODE.**
Figure tirée du site web du projet ENCODE (http://genome.ucsc.edu/ENCODE). Les équipes associées à ce projet utilisent de multiples expériences et méthodes pour identifier les éléments fonctionnels du génome. L'annotation des gènes se fait principalement par le séquençage de l'ARN provenant de différentes sources (tissu, stade de développement), la génomique comparative, des méthodes bioinformatiques ou encore par annotation manuelle. Les éléments de régulation sont quant à eux étudiés *via* des expériences d'hypersensibilité à la DNAseI, l'étude de la méthylation de l'ADN, ou encore les données ChIP-seq de divers facteurs de transcription et histones spécifiquement méthylées ou acétylées.

1.7.4 Le projet ENCODE

Le projet ENCODE (pour *Encyclopedia of DNA Elements*) est un consortium de groupes de recherche internationaux financés par le NHGRI (*National Human Genome Research Institute*) qui a vu le jour afin de systématiser les méthodes permettant l'annotation des génomes et de faciliter l'intégration des nombreuses données obtenues. Son but est de construire une liste exhaustive des éléments fonctionnels du génome humain, qu'ils agissent au niveau de l'ADN, de l'ARN ou des protéines, et des éléments de régulation qui contrôlent l'état cellulaire et l'activité des gènes. Les données sont mises à disposition du public gratuitement sur internet (http://genome.ucsc.edu/ENCODE/). À noter que des projets équivalents existent pour d'autres organismes, comme la souris (http://mouseencode.org), ou encore le ver *Caenorhabditis elegans* et la mouche *Drosophila melanogaster* (http://www.modencode.org).

Totalisant en septembre 2012 plus de 1600 expériences dans plus de 147 types cellulaires, les premières conclusions pointent vers une profusion d'événements de régulation, loin de l'idée d'ADN poubelle (*junk DNA*) : ainsi, 80% du génome est associé à un événement biochimique associé à de la formation d'ARN ou au remodelage de la chromatine, $\sim 400,000$ régions possèdent un état chromatinien caractéristique des enhancers et $\sim 70,000$ des promoteurs (ENCODE Project Consortium et al., 2012). Depuis mai 2013, les données ChIP-seq de

161 TFs couvrant 91 types cellulaires ont été mises à disposition sur UCSC Genome Browser [10].

10. http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeAwgTfbsUniform

Chapitre 2

Modèles de fixation des Facteurs de Transcription à l'ADN.

2.1 Observations de corrélations au sein des TFBS **50**
2.2 Modèles existants permettant de décrire la statistique des TFBS **51**
2.2.1 Modèle de référence sans corrélations : la PWM 51
2.2.2 Une PWM généralisée : le modèle GWM 52
2.2.3 Réseaux bayésiens 53
2.2.4 Modèles de mélange 53
2.3 Modèles de maximum d'entropie **54**
2.3.1 Pourquoi maximiser l'entropie ? 54
2.3.2 Maximisation de l'entropie sous contraintes 56
2.3.3 Application aux sites de fixation 57
2.4 Article **58**
2.5 Analyse thermodynamique des modèles **90**
2.5.1 Chaleur spécifique 90
2.5.2 Lien avec les valeurs des champs et des couplages 91
2.6 Conclusion et perspectives du chapitre 2 . . . . . **91**

Introduction du chapitre 2

Dans cette partie, nous nous intéressons à la description de l'interaction entre les facteurs de transcription et leurs sites de reconnaissance sur l'ADN. Pendant longtemps, la qualité de cette description a été limitée par la quantité de données disponibles. Ainsi, les expériences de type SELEX (voir 1.4.1), où des expériences de ChIP au cas par cas permettaient de récupérer de l'ordre de quelques dizaines de sites de fixation pour un TF d'intérêt. Or, le modèle PWM, qui est le modèle le plus simple (en terme de nombre de paramètres) que l'on puisse bâtir pour décrire l'interaction possède déjà plusieurs dizaines de paramètres – les fréquences des nucléotides à chaque position –.

Ces données ne permettaient donc pas d'explorer plus en avant des modèles plus complexes de fixation incluant par exemple des termes d'interaction entre nucléotides au sein des sites de fixation. Cependant, les avancées récentes en séquençage à haut débit ont permis l'obtention de données très grande échelle, que ce soit *in vivo* par ChIP-seq ou *in vitro* par HT-SELEX (voir 1.4). Le nombre de sites de fixation obtenus est de l'ordre de quelques milliers, ce qui permet de contraindre des modèles de fixation plus complexe que le modèle PWM.

En utilisant des données ChIP-seq pour un grand nombre de facteurs de transcription de la Drosophile et des vertébrés, nous avons contraint différents modèles de fixation incluant implicitement ou explicitement des interactions entre nucléotides. Nous les avons comparés sur leur capacité à décrire les statistiques de fixation TF-ADN observées *in vivo*. Nous présentons préalablement un survol des observations et modèles existant au sujet des corrélations dans les sites de fixations de facteurs de transcription.

2.1 Observations de corrélations au sein des TFBS

Différents travaux ont mis en exergue l'existence de corrélations entre nucléotides au sein des sites de fixation de TFs. Parce que limitées par la quantité de données alors possible à obtenir, les premières études de ce genre ont centré leur attention sur quelques corrélations importantes pour des cas particuliers. Ainsi, Man and Stormo (2001) ont observé que la protéine Mnt induit des corrélations entre les positions 16 et 17 de ses sites de reconnaissance *in vitro*. Ils ont mesuré expérimentalement la spécificité aux sites de liaisons contenant tous les variants possibles à ces deux positions. Ils ont ainsi observé que la mutation de la base consensus C en position 17 induisait un changement de préférence en position 16 de la base A vers la base C. Par ailleurs, Bulyk et al. (2002) ont montré que la protéine EGR1 induisait des corrélations au sein d'un triplet de nucléotides central de leur site de reconnaissance. La prise en compte de ces corrélations dans l'énergie de fixation permettait alors d'améliorer la description des données par rapport au modèle additif PWM.

À une plus grande échelle, Badis et al. (2009) ont utilisé des puces à ADN (technique PBM, cf 1.4.1) pour étudier la fixation *in vitro* de 104 TFs de la souris sur toutes les séquences d'ADN de 10 bp possibles. Pour chaque facteur, plusieurs centaines de séquences de fixation ont ainsi été obtenues. L'étude a révélé l'existence d'une multiplicité de motifs (PWMs) pour la plupart des TFs (seulement 15 étant mieux décrit par un motif unique). Certains motifs reconnaissent notamment des séquences à espacement variable pour lesquelles deux régions spécifiques du site sont séparées par un nombre variable de nucléotides. Enfin, les auteurs ont noté la présence de corrélations fortes dans 19 cas, celles-ci n'étant pas forcément limitées à des dinucléotides mais pouvant impliquer des trinucléotides. Plus récemment, Jolma et al. (2013) ont analysé par HT-SELEX plusieurs centaines de domaines de fixations à l'ADN de

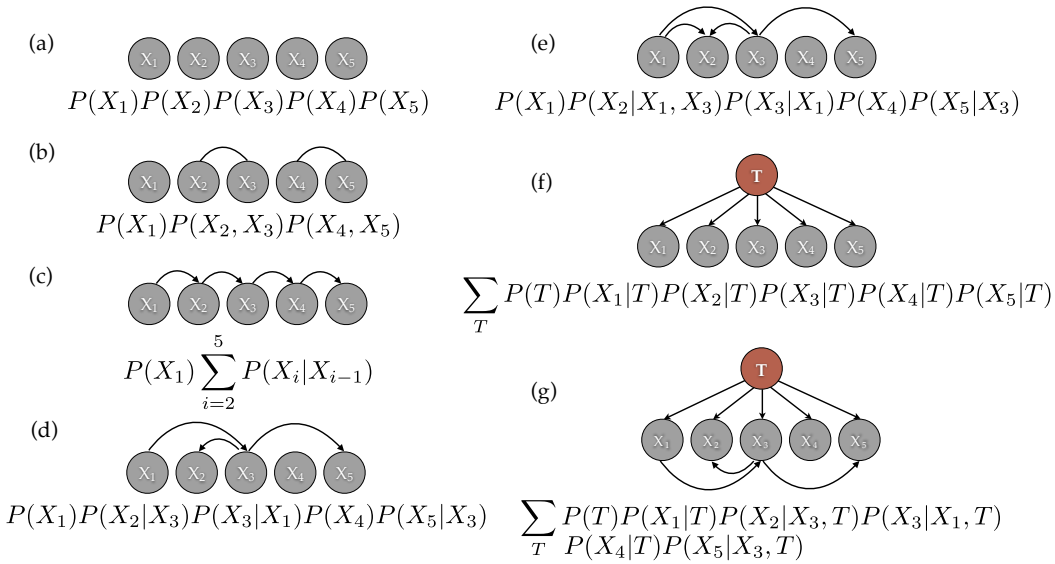


FIGURE 2.1 – **Différents modèles pour décrire les corrélations entre nucléotides dans les sites de fixation de facteurs de transcription.**
Figure adaptée de Barash et al. (2003) illustrant différents modèles de fixation sur un site de longueur 5. Pour chaque modèle, la structure du réseau de dépendances sous-jacent est représentée, ainsi que la distribution de probabilité $P(X_1, X_2, X_3, X_4, X_5)$ correspondante, où X_i est une variable aléatoire prenant la valeur (A, C, G ou T) du nucléotide à la position i. Les modèles représentés sont les suivants : (a) PWM (pas de corrélations), (b) GWM (corrélations mutuellement exclusives), (c) chaîne de Markov d'ordre 1 (corrélations entre plus proches voisins), (d) réseau bayésien en arbre (au plus un parent par nœud) ou (e) pas en arbre (le nœud 2 a deux parents), (f) mélange de PWMs et (e) mélange d'arbres à dépendances fixées.

TFs humains et de la souris, révélant aussi l'importance d'espacements variables et surtout des corrélations dinucléotidiques entre plus proches voisins.

2.2 Modèles existants permettant de décrire la statistique des TFBS

Différents modèles ont été proposés pour décrire ces corrélations (fig. 2.1). La méthode la plus directe consiste à partir du modèle PWM (fig. 2.1a) et à ajouter des corrélations mutuellement exclusives aux positions les plus corrélées (fig. 2.1b). D'autres méthodes utilisent des structures probabilistes de dépendances sous forme de chaînes de Markov (fig. 2.1c) ou plus généralement de réseau bayésien (fig. 2.1d-e). Enfin, une dernière méthode consiste à réaliser des mélanges de modèles afin de capturer des ensembles distincts de corrélations (fig. 2.1f-g).

2.2.1 Modèle de référence sans corrélations : la PWM

Nous l'avons vu, le modèle le plus simple (en termes de nombre de paramètres) décrivant l'interaction entre un TF et son site de reconnaissance sur l'ADN consiste à faire l'hypothèse que les nucléotides contribuent indépendamment à l'énergie de fixation. Cette hypothèse

conduit au modèle PWM (section 1.3.2 et fig.2.1a), qui s'écrit [11] :

$$P(X_1, \cdots, X_k) = \prod_{i=1}^{K} P(X_i) \tag{2.1}$$

où $P(X_i)$ est la probabilité marginale d'observer le nucléotide $X \in \{A, C, G, T\}$ à la position i. Un tel modèle possède $3K$ paramètres – 3 paramètres $P(X_i)$ par position, la normalisation des probabilités permettant de fixer le paramètre restant –. Pour une longueur de site typique $K = 10$, le modèle PWM contient 30 paramètres à contraindre, sachant qu'un « modèle » complet paramétrant la distribution jointe sans faire d'hypothèse comporterait $4^{10} - 1 \sim 10^6$ paramètres.

2.2.2 Une PWM généralisée : le modèle GWM

Une première méthode permettant de complexifier le modèle PWM consiste à intégrer explicitement des groupes mutuellement exclusif [12] de nucléotides corrélés au sein du modèle (fig. 2.1b). Une telle méthode fut d'abord employée par Benos et al. (2002) pour prendre en compte des corrélations préalablement définies entre nucléotides plus proches voisins. De manière plus générale, Zhou and Liu (2004) ont développé un modèle de matrice de poids généralisée (GWM pour *Generalized Weight Matrix*) qui prend en compte de manière systématique les corrélations permettant d'améliorer le modèle indépendant. Pour ce faire, les auteurs utilisent une méthode de Monte-Carlo par chaîne de Markov (MCMC) : des corrélations sont ajoutées ou enlevées au hasard au modèle et acceptées selon la règle de Metropolis-Hastings (Krauth, 2006). Cette acceptation est proportionnelle au facteur de Bayes, une quantité qui permet de comparer des modèles possédant des nombres de paramètres différent [13] . Ce facteur est défini par le rapport entre la probabilité de générer les données D (les séquences de fixation) avec un modèle M_1 de paramètres θ_1 plutôt qu'avec un autre modèle M_2 de paramètres θ_2 :

$$BF = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\theta_1, M_1)P(\theta_1|M_1)d\theta_1}{\int P(D|\theta_2, M_2)P(\theta_2|M_2)d\theta_2} \tag{2.2}$$

Le modèle final consiste en un ensemble de paramètres décrivant des positions indépendantes et des positions corrélées mutuellement exclusives . En analysant les données TRANSFAC, les auteurs ont noté que dans 25% des cas (22/95) le modèle GWM était significativement meilleur que le modèle PWM (facteur de Bayes supérieur à 6).

Cette méthode a par la suite été utilisée sur des données ChIP-seq pour 4 TFs mammifères – NRSF, STAT1, CTCF et ER – (Hu et al., 2010). En utilisant les 10% des pics les plus importants comme ensemble d'apprentissage et en se restreignant aux régions de 200bp centrées autour du sommet du pic ChIP, les auteurs ont réalisé un échantillonnage de Gibbs (Casella and George, 1992) pour obtenir les sites de fixation suivant les hypothèses que (1) chaque pic contient au plus un seul site de fixation (modèle ZOOPS pour *Zero or One Occurrences Per Sequence*), (2) la probabilité *a priori* d'avoir un site à une certaine position sur la séquence est plus forte autour du sommet du pic, et (3) les sites sont décrits par un modèle GWM. L'étude

11. Comme nous l'avons signalé en 1.3.2, le terme PWM (*Position Weight Matrix*) réfère en fait à la matrice des poids $\log(P(X_i)/\pi_{X_i})$ où π_{X_i} est une distribution neutre indépendante de la position (dite distribution *background*), par exemple calculée sur des régions intergéniques.

12. Les corrélations entre des couples de positions (i,j) et (j,k) ne peuvent être admises au sein du même modèle.

13. Sous certaines approximation, ce facteur peut se rapporter à une différence de valeurs du BIC, introduit dans l'article en 2.4.

a révélé l'existence de corrélations fortes limitées aux nucléotides plus proches voisins dans les quatre cas étudiés. Les nucléotides participant aux corrélations se situaient à des positions ayant un faible contenu en information dans le modèle PWM. Enfin, les auteurs ont noté la présence de plusieurs triplets de nucléotides voisins corrélés.

2.2.3 Réseaux bayésiens

Une généralisation du modèle GWM consiste à supprimer la condition d'exclusion mutuelle des paires de nucléotides corrélés en décrivant de manière plus générale le réseau de dépendance entre positions. Une telle description est possible en utilisant le langage des réseaux bayésiens. Les dépendances y sont représentées par un graphe orienté acyclique [14] G, dont les nœuds sont les variables X_i et les liens représentent les conditionnements d'une variable avec des variables parentes (fig. 2.1e). La probabilité jointe s'écrit :

$$P(X_1, \cdots, X_k) = \prod_{i=1}^{K} P(X_i | P_i^G) \tag{2.3}$$

où P_i^G est l'ensemble (pouvant être vide) des parents de X_i dans G. Le nombre de paramètres peut rapidement devenir grand : si l'on note N_i le nombre de parents de X_i, alors le nombre de paramètres du modèle est $3\sum_{i=1}^{K} 4^{N_i}$.

Lorsque les différents nœuds possèdent au plus un parent, on parle d'arbre bayésien (fig. 2.1d). Ce type de réseau bayésien généralise notamment le cas des chaînes de Markov d'ordre 1, où chaque nœud dépend du nœud précédent (fig. 2.1c). Le nombre de paramètres est alors restreint, puisqu'il est au plus de $3 \cdot 4K$.

L'avantage des arbres bayésiens est qu'il existe des algorithmes efficaces permettant de trouver la meilleure structure d'arbre (Friedman et al., 1997). De tels modèles d'arbres ont été utilisés pour décrire les données de 95 TFs de Transfac (Barash et al., 2003). Dans $\sim 25\%$ des cas (22/95), le modèle d'arbre bayésien s'avère significativement meilleur qu'un modèle PWM, ce qui est du même ordre de grandeur que pour le modèle GWM vu en 2.2.2.

2.2.4 Modèles de mélange

Dans les cas précédents, nous avons présenté des modèles capturant des dépendances « locales » entre paires de nucléotides. Néanmoins, il peut exister des dépendances plus largement réparties entre les positions, comme cela a déjà été observé empiriquement (Badis et al., 2009; Jolma et al., 2013). De telles corrélations à plus grande échelle peuvent être modélisées en supposant que le facteur de transcription possède plusieurs « modes » de fixation. Ceux-ci peuvent par exemple correspondre à différentes conformations de la protéine sur son site de fixation, chaque configuration possédant ses propres préférences de fixation. Ces modes sont décrits par une variable aléatoire T (le *type* de fixation) de probabilité $P(T)$. Il est ensuite possible de décrire la fixation au sein de chaque mode par l'un des modèles précédents.

- **Mélange de PWMs**

Le cas le plus naturel consiste à utiliser comme modèle de fixation le modèle PWM, c'est-à-dire que dans chaque mode il y a indépendance entre les positions. La probabilité d'observer

14. Un graphe orienté acyclique est un réseau dont les liens sont orientés et au sein duquel il n'est pas possible de revenir à son point de départ en suivant les flèches

un site est alors donnée par la somme sur les différents modes de fixation de la probabilité de fixer un site, conditionnée par la probabilité d'être dans ce mode :

$$P(X_1, \cdots, X_K) = \sum_{T=1}^{N} P(T) \prod_{i=1}^{K} P(X_i|T) \tag{2.4}$$

où N est le nombre de modes de fixation. Ce modèle a plusieurs avantages. D'abord, le nombre de paramètres reste linéaire en K : pour décrire $P(T)$ et les N PWMs il faut $N - 1 + 3KN$ paramètres. Ce nombre reste dont raisonnablement faible devant le nombre de paramètres requis pour complètement décrire les interactions à deux nucléotides, qui croît comme K^2. Ensuite, le modèle a une interprétation claire qui peut permettre de mettre en exergue un mécanisme biologique sous-jacent.

Ce type de modèle permet de dépasser le modèle PWM dans un nombre substantiel de cas. Ainsi, Barash et al. (2003) ont montré que $\sim$ 40% des TFs de Transfac (36/95) sont significativement mieux représentés par un mélange de 2 PWMs que par une seule PWM. En utilisant des données *in vitro* plus précises pour 104 TFs de la souris, Badis et al. (2009) ont montré que $\sim$ 85% (89/104) étaient mieux représenté par une combinaison de PWMs que par une PWM seule, plaidant pour un portée générale de l'existence de « motifs secondaires ».

- **Mélange d'arbres**

De la même manière que les PWMs, il est possible d'étendre les modèles d'arbres en réalisant un mélange d'arbres. Intuitivement, ceci permet de capturer des dépendances additionnelles en gardant un nombre de paramètres linéaire en fonction de la taille du motif. Un tel modèle semble posséder des performances comparables au mélange de PWM, et améliore la description des TFs de Transfac dans $\sim$ 40% des cas (35/95) (Barash et al., 2003).

2.3 Modèles de maximum d'entropie

2.3.1 Pourquoi maximiser l'entropie ?

Le concept d'entropie remonte aux prémisses de la physique statistique (Jaynes, 1978). Dans l'essence, il peut être compris de la manière suivante. Supposons qu'un système comporte K états distincts $\{s_1, \cdots, s_K\}$. Au cours du temps, le système explore les différents états (fig. 2.2). Au bout de N observations, chaque état a été observé un nombre n_i de fois. La question sous-jacente au calcul de l'entropie est la suivante : sans connaissance *a priori* sur le système, que puis-je dire de ces n_i ? Prenons l'exemple de la figure 2.2. On a certaines valeurs pour les n_i ($n_1 = 3$, $n_2 = 1$, etc.), et on aimerait savoir de combien de manières il est possible de réaliser un tel ensemble de valeurs. Notons ce nombre $\mathcal{N}(n_1, \cdots, n_K)$. Il est donné par la formule suivante :

$$\begin{aligned}\mathcal{N}(n_1, \cdots, n_K) &= \binom{N}{n_1}\binom{N-n_1}{n_2}\cdots\binom{N-\sum_{i=1}^{K-1} n_i}{n_K} \\ &= \frac{N!}{(N-n_1)!n_1!} \times \frac{(N-n_1)!}{(N-n_1-n_2)!n_2!} \times \cdots \times \frac{(N-\sum_{i=1}^{K-1} n_i)!}{0!n_K!}\end{aligned} \tag{2.5}$$

soit

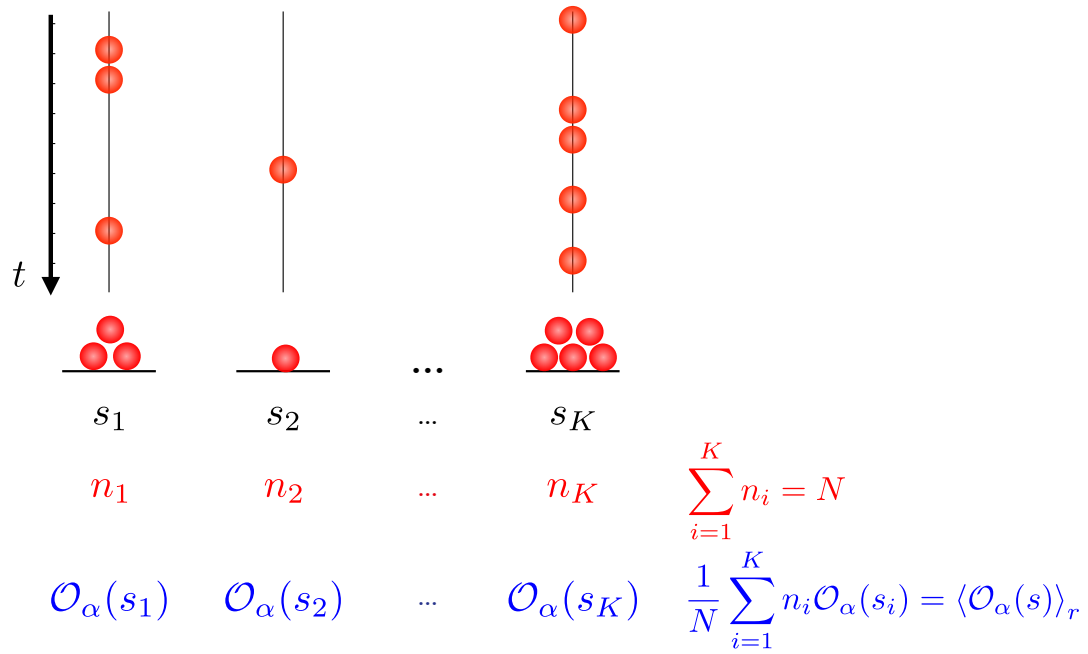


FIGURE 2.2 – **Illustration d'un système dont on veut maximiser l'entropie.**
Le système explore K états $\{s_1, \cdots, s_K\}$ au cours du temps. Au bout de N observations, un état s_i a été observé n_i fois. À chaque état correspond un certain nombre de quantités $\mathcal{O}_\alpha(s_i)$ dont seule la valeur moyenne empirique $\langle \mathcal{O}_\alpha(s) \rangle_r$ est connue de l'observateur (on parle d'« observable »).

$$\mathcal{N}(n_1, \cdots, n_K) = \frac{N!}{n_1! n_2! \cdots n_K!} \tag{2.6}$$

Il convient alors de s'intéresser au logarithme de cette quantité. En effet, dans le cas où les nombres d'observation sont grands $n_i \gg 1$, ceux-ci s'expriment simplement grâce à la formule de Stirling :

$$\log(n!) \xrightarrow[n\to\infty]{} n\log(n) - n \tag{2.7}$$

On peut alors écrire

$$\begin{aligned} \log \mathcal{N}(n_1, \cdots, n_K) &= N\log(N) - N - \sum_{i=1}^{K} (n_i \log(n_i) - n_i) \\ &= \sum_{i=1}^{K} n_i \log(\frac{N}{n_i}) \\ &= -N \sum_{i=1}^{K} \frac{n_i}{N} \log(\frac{n_i}{N}) \end{aligned} \tag{2.8}$$

On note l'apparition des probabilités empiriques $f(s_i) = n_i/N$ d'observer l'état s_i, qui tendent asymptotiquement (dans la limite « thermodynamique » $N \to \infty$) vers les « vraies »

probabilités $P(s_i)$. L'entropie est définie dans cette limite comme étant égale à $1/N \log \mathcal{N}(n_1, \cdots, n_K)$, soit

$$\boxed{S[P] = -\sum_{\{s\}} P(s) \log P(s)} \tag{2.9}$$

où $\{s\} = \{s_1, \cdots, s_K\}$ dénote l'ensemble des états accessibles. L'idée est alors la suivante : nous souhaitons savoir quels états le système a le plus probablement visité au cours des N transitions. Sans connaissance *a priori* sur le système, il est plus probable que les nombres $(n_1, \cdots, n_K)$ obtenus soient ceux qui sont réalisés le plus souvent, c'est-à-dire ceux qui maximisent la quantité $\mathcal{N}(n_1, \cdots, n_K)$ et donc au final l'entropie. Par ailleurs, les fluctuations relatives des quantités n_i sont de l'ordre de $1/\sqrt{n_i}$ (Sethna, 2006). Ainsi, la solution de maximum d'entropie domine largement les autres solutions possibles dans la limite thermodynamique.

2.3.2 Maximisation de l'entropie sous contraintes

Notons $\mathcal{O}_\alpha(s)$ une quantité attachée à s (fig. 2.2). En thermodynamique, une telle quantité correspond par exemple à l'énergie d'un état. L'observateur n'a lui accès qu'aux valeurs moyennes de telles quantités sous-jacentes. À l'état d'équilibre thermodynamique, l'échantillonnage des états est réalisé au sein de la distribution de probabilité de maximum d'entropie $P(s)$, et les valeurs moyennes calculées avec les fréquences empiriques $f(s)$ doivent donc être compatibles avec les valeurs moyennes calculées avec la distribution de probabilité $P(s)$:

$$\sum_{\{s\}} P(s)\mathcal{O}_\alpha(s) = \sum_{\{s\}} f(s)\mathcal{O}_\alpha(s) \tag{2.10}$$

Nous souhaiterions maintenant connaître la distribution $P(s)$ la moins biaisée (i.e de maximum d'entropie) qui satisfait les contraintes de l'éq. 2.10 imposées par l'observation des données (l'information que possède l'observateur). Ce problème revient à maximiser le Lagrangien suivant :

$$\mathcal{L} = -\sum_{\{s\}} P(s) \log P(s) + \lambda \left(\sum_{\{s\}} P(s) - 1 \right) + \sum_\alpha \beta_\alpha \sum_{\{s\}} \left(P(s) - f(s) \right) \mathcal{O}_\alpha(s) \tag{2.11}$$

où les paramètres λ et β_α sont les multiplicateurs de Lagrange correspondant respectivement à la contrainte de normalisation de la distribution de probabilité et aux informations qu'a l'observateur sur certaines valeurs moyennes du système (éq. 2.10). La maximisation de ce Lagrangien est obtenue en annulant la dérivée fonctionnelle par rapport à la distribution de probabilité $P(s)$:

$$\frac{\delta \mathcal{L}}{\delta P(s)} = 0 = -\ln P(s) - 1 + \lambda + \sum_\alpha \beta_\alpha \mathcal{O}_\alpha(s) \tag{2.12}$$

En utilisant la normalisation des probabilités, il est possible de trouver λ, et la solution se met finalement sous la forme

$$\boxed{P(s) = \frac{1}{\mathcal{Z}} e^{-\mathcal{H}(s)}} \tag{2.13}$$

où $\mathcal{H}$ est l'Hamiltonien du système :

$$\mathcal{H} = \sum_{\alpha} \beta_\alpha \mathcal{O}_\alpha(s) \tag{2.14}$$

et $\mathcal{Z}$ est la fonction de partition permettant la normalisation de la distribution $P(s)$:

$$\mathcal{Z} = \sum_{\{s\}} \exp[-\mathcal{H}(s)]. \tag{2.15}$$

Remarque

Il est possible de montrer que la maximisation de l'entropie, partant des contraintes de l'éq. 2.10 sur les valeurs moyennes pour en arriver à une forme exponentielle de la distribution de probabilité, est le contrepoint d'une maximisation de la vraisemblance partant d'une forme exponentielle pour en arriver aux mêmes contraintes sur les valeurs moyennes (Grendar Jr and Grendar, 2001; Jaynes, 1982).

2.3.3 Application aux sites de fixation

- **Corrélations à un point : le modèle PWM**

Dans le cas qui nous intéresse, un état s correspond à une séquence d'ADN appartenant à l'ensemble $\{s\}$ des sites de fixation d'un facteur de transcription. Considérons maintenant l'observable quantifiant la présence du nucléotide a à la position i d'un site :

$$\mathcal{O}_{i,a}(s) = \delta(s_i, a) \tag{2.16}$$

où δ est la fonction de Kronecker qui vaut 1 lorsque le nucléotide à la position i du site s_i vaut a et 0 sinon. De cette définition il suit que la valeur moyenne sur les fréquences empiriques

$$\sum_{\{s\}} f(s)\mathcal{O}_{i,a}(s) = f_{i,a} \tag{2.17}$$

se réduit à la fréquence du nucléotide a à la position i. Notons $h_i(a)$ le multiplicateur de Lagrange correspondant et $\mathcal{A} = \{A, C, G, T\}$. On trouve alors

$$\begin{aligned} \mathcal{H}(s) &= \sum_{i=1}^{L} \sum_{a \in \mathcal{A}} h_i(a)\delta(s_i, a) \\ &= \sum_{i=1}^{L} h_i(s_i) \end{aligned} \tag{2.18}$$

Les différentes positions étant indépendantes, la fonction de partition $\mathcal{Z}$ peut par ailleurs se scinder en différentes fonctions de partition par position : $\mathcal{Z} = \prod_{i=1}^{L} \mathcal{Z}_i$. On obtient au final

$$P(s) = \frac{1}{\mathcal{Z}} e^{-\sum_{i=1}^{L} h_i(s_i)} = \prod_{i=1}^{L} \frac{e^{-h_i(s_i)}}{\mathcal{Z}_i} \tag{2.19}$$

On retrouve le modèle PWM introduit dans l'éq. 2.13.

- **Corrélations à deux points : le modèle de Potts**

Il est maintenant relativement direct de complexifier le modèle en ajoutant l'observation des couples d'interaction au sein des sites de fixation :

$$\mathcal{O}_{i,a,j,b}(s) = \delta(s_i, a)\delta(s_j, b) \tag{2.20}$$

La corrélation à deux points entre le nucléotide a en position i et b en position j s'écrit donc

$$\sum_{\{s\}} f(s)\mathcal{O}_{i,a,j,b}(s) = f_{i,a,j,b} \tag{2.21}$$

où $f_{i,a,j,b}$ est la fréquence empirique d'observation de la paire de nucléotide (a, b) aux positions (i, j). Notons $J_{i,j}(a, b)$ le multiplicateur de Lagrange correspondant. L'Hamiltonien sous les contraintes imposées par les équations 2.17 et 2.21 s'écrit :

$$\begin{aligned}\mathcal{H}(s) &= \sum_{i=1}^{L}\sum_{a\in\mathcal{A}} h_i(a)\delta(s_i, a) + \sum_{i=1}^{L-1}\sum_{j>i}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{A}} J_{i,j}(a, b)\delta(s_i, a)\delta(s_j, b) \\ &= \sum_{i=1}^{L} h_i(s_i) + \sum_{i=1}^{L-1}\sum_{j>i} J_{i,j}(s_i, s_j)\end{aligned} \tag{2.22}$$

Le modèle de maximum d'entropie est finalement

$$\boxed{P(s) = \frac{1}{\mathcal{Z}} e^{-\sum_{i=1}^{L} h_i(s_i) - \sum_{i=1}^{L-1}\sum_{j>i} J_{i,j}(s_i, s_j)}} \tag{2.23}$$

On reconnaît le modèle de Potts inhomogène de champs magnétiques locaux h_i et de termes d'interaction $J_{i,j}$ couramment utilisé dans la description des verres de spins (Baxter, 2007).

2.4 Article

L'article qui suit décrit l'analyse de données de fixation *in vivo* à grande échelle pour plusieurs TFs drosophiles et mammifères. Différents modèles sont comparés, incluant ou non des dépendances : un modèle PWM, un modèle de mélange de PWMs, et un modèle de Potts.

Beyond position weight matrices: nucleotide correlations in transcription factor binding sites and their description

Marc Santolini, Thierry Mora, and Vincent Hakim
Laboratoire de Physique Statistique, CNRS, Université P. et M. Curie, Université D. Diderot, École Normale Supérieure, Paris, France.

The identification of transcription factor binding sites (TFBSs) on genomic DNA is of crucial importance for understanding and predicting regulatory elements in gene networks. TFBS motifs are commonly described by Position Weight Matrices (PWMs), in which each DNA base pair independently contributes to the transcription factor (TF) binding, despite mounting evidence of interdependence between base pairs positions. The recent availability of genome-wide data on TF-bound DNA regions offers the possibility to revisit this question in detail for TF binding *in vivo*. Here, we use available fly and mouse ChIPseq data, and show that the independent model generally does not reproduce the observed statistics of TFBS, generalizing previous observations. We further show that TFBS description and predictability can be systematically improved by taking into account pairwise correlations in the TFBS via the principle of maximum entropy. The resulting pairwise interaction model is formally equivalent to the disordered Potts models of statistical mechanics and it generalizes previous approaches to interdependent positions. Its structure allows for co-variation of two or more base pairs, as well as secondary motifs. Although models consisting of mixtures of PWMs also have this last feature, we show that pairwise interaction models outperform them. The significant pairwise interactions are found to be sparse and found dominantly between consecutive base pairs. Finally, the use of a pairwise interaction model for the identification of TFBSs is shown to give significantly different predictions than a model based on independent positions.

Author Summary

Transcription factors are proteins that bind on DNA to regulate several processes such as gene transcription or epigenetic modifications. Being able to predict the Transcription Factor Binding Sites (TFBSs) with accuracy on a genome-wide scale is one of the challenges of modern biology, as it allows for the bottom-up reconstruction of the gene regulatory networks. The description of the TFBSs has been to date mostly limited to a simple model, where the affinity of the protein for DNA, or binding energy, is the sum of independent contributions from uncorrelated amino-acids bound on base pairs. However, structural aspects are of prime importance in proteins and could imply appreciable correlations throughout the observed binding sequences. Using a statistical physics inspired description and high-throughput ChIPseq data for a variety of Drosophilae and mammals TFs, we show that such correlations exist and that accounting for their contribution greatly improves the predictability of genomic TFBSs.

Introduction

Gene regulatory networks are at the basis of our understanding of a cell state and of the dynamics of its response to environmental cues. Central effectors of this regulation are Transcription Factors (TF) that bind on short DNA regulatory sequences and interact with the transcription apparatus or with histone-modifying proteins to alter target gene expressions [1]. The determination of Transcription Factor Binding Sites (TFBSs) on a genome-wide scale is thus of importance and is the focus of many current experiments [2]. An important feature of TF in eukaryotes is that their binding specificity is moderate and that a given TF is found to bind a variety of different sequences *in vivo* [3]. The collection of binding sequences for a TF-DNA is widely described by a Position Weight Matrix (PWM) which simply gives the probability that a particular base pair stands at a given position in the TFBS. The PWM provides a full statistical description of the TFBS collection when there are no correlations between nucleotides at different positions. Provided that the TF concentration is far from saturation, the PWM description applies exactly at thermodynamic equilibrium in the simple case where the different nucleotides in the TFBS contribute independently to the TF-DNA interaction, such that the total binding energy is the mere sum of the individual contributions [4, 5].

Previous works have reported several cases of correlations between nucleotides at different positions in TFBSs [6–9]. A systematic *in vitro* study of 104 TFs using DNA microarrays revealed a rich picture of binding patterns [10], including the existence of multiple motifs, strong nucleotide position interdependence, and variable spacer motifs, where two small determining regions of the binding site are separated by a variable number of base pairs. Recently, the specificity of several hundred human and mouse DNA-binding domains was investigated using high-throughput SELEX. Correlations between nucleotides were found to be widespread among TFBSs and predominantly located between adjacent flanking bases in the TFBS [9]. The relevance of nucleotide correlations remains however debated [11].

On the modeling side, probabilistic models have been proposed to describe these correlations, either by explicitly identifying mutually exclusive groups of co-varying

nucleotide positions [7, 12, 13], or by assuming a specific and tractable probabilistic structure such as Bayesian networks or Markov chains [9, 14, 15]. However, the extent of nucleotide correlations in TFBSs *in vivo* remains to be assessed, and a systematic and general framework that accounts for the the rich landscape of observed TF binding behaviours is yet to be applied in this context. The recent breakthrough in the experimental acquisition of precise, genome-wide TF-bound DNA regions with the ChIPseq technology offers the opportunity to address these two important issues. Using a variety of ChIPseq experiments coming both from fly and mouse, we first show that the independent model generally does not reproduce well the observed TFBS statistics for a majority of TF. This calls for a refinement of the PWM description that accounts for interdependence between nucleotide positions.

The general problem of devising interaction parameters from observed state frequencies has been recently studied in different contexts where large amounts of data have become available. These include describing the probability of coinciding spikes [16, 17] or activation sequences [18, 19] in neural data, the statistics of protein sequences [20, 21], and even the flight directions of birds in large flocks [22]. Maximum-entropy models accounting for pairwise correlations in the least constrained way have been found to provide significant improvement over independent models. The PWM description of TF binding is equivalent to the maximum entropy solely constrained by nucleotide frequencies at each position. Thus, we propose, in the present paper, to refine this model by further constraining pairwise correlations between nucleotide positions. This corresponds to including effective pairwise interactions between nucleotides in an equilibrium thermodynamic model of TF-DNA interaction, as already proposed [23]. When enough data are available, the TFBS statistics and predictability are found to be significantly improved in this refined model. We consider, for comparison, a model that describes the statistics of TFBSs as a statistical mixture of PWMs [14] and generalizes previous proposals [24, 25]. This alternative model can directly capture some higher-order correlations between nucleotides but is found to be outperformed for all considered TF by the pairwise interaction model.

We further show that the pairwise interaction model accounts for the different PWMs appearing in the mixture model by studying its energy landscape: each basin of attraction of a metastable energy minimum in the pairwise interaction model is generally dominantly described by one PWM in the mixture model. Significant pairwise interactions between nucleotides are sparse and found dominantly between consecutive nucleotides, in general qualitative agreement with *in vitro* binding results [9]. The proposed model with pairwise interactions only requires a modest computational effort. When enough data are available, it should thus generally prove worth using the refined description of TFBS that it affords.

Results

The PWM model does not reproduce the TFBS statistics

We first tested how well the usual PWM model reproduced the observed TFBS statistics, *i.e.* how well the frequencies of different TFBSs were retrieved by using only single nucleotide frequencies. For this purpose, we used a collection of ChIPseq data available from the literature [26–28], both from *D. Melanogaster* and from mouse embryonic stem cells (ESC) and a myogenic cell line (C2C12). The TFBSs are short L-mers (we take here $L = 12$), which are determined in each few hundred nucleotides long ChIP-bound region with the help of a model of TF binding. One important consequence and specific features of these data, is that the TFBS collection is not independent of the model used to describe it. Thus, in order to self-consistently determine the collection of binding sites for a given TF from a collection of ChIPseq sequences, we iteratively refined the PWM together with the collection of TFBSs in the ChIPseq data (see Figure and *Methods*). This process ensured that the frequency of different nucleotides at a given position in the considered ensemble of binding sites was exactly accounted by the PWM. We then enquired whether the probability of the different binding sequences in the collection agreed with that predicted by the PWM, as would be the case if the probabilities of observing nucleotides at different positions were independent. Figure 2 displays the results for three different TFs, one from each of the three considered categories: Twi (Drosophila), Esrrb (mammals, ESC), and MyoD (mammals, C2C12). For each factor, the ten most frequent sequences in the TFBS collection are shown. For comparison, Figure 2 also displays the probabilities for these sequences as predicted by the PWM built from the TFBS collection. The independent PWM model strongly underestimates the probabilities of the most frequent sequences. Moreover, the PWM model does not correctly predict the frequency order of the sequences and attributes comparable probabilities to these different sequences, in contrast to their observed frequencies.

The relative entropy or Kullback-Leibler divergence (DKL) is a general way to measure the difference between two probability distributions [29]. In order to better quantify the differences between the observed binding sequence frequencies and the PWM frequencies, we computed the DKL between these distributions for all the considered TF, as shown in Figure 2D. For each transcription factor T, part of the differences comes from the finite number N(T) of its observed binding sites. The results are thus compared for each factor T to DKLs between the PWM probabilities and frequencies obtained for artificial sequence samples of size N(T) generated with the same PWM probabilities. For most TFs (22 out of 28), the difference between the observed binding sequence frequencies and the PWM frequencies is signifi-

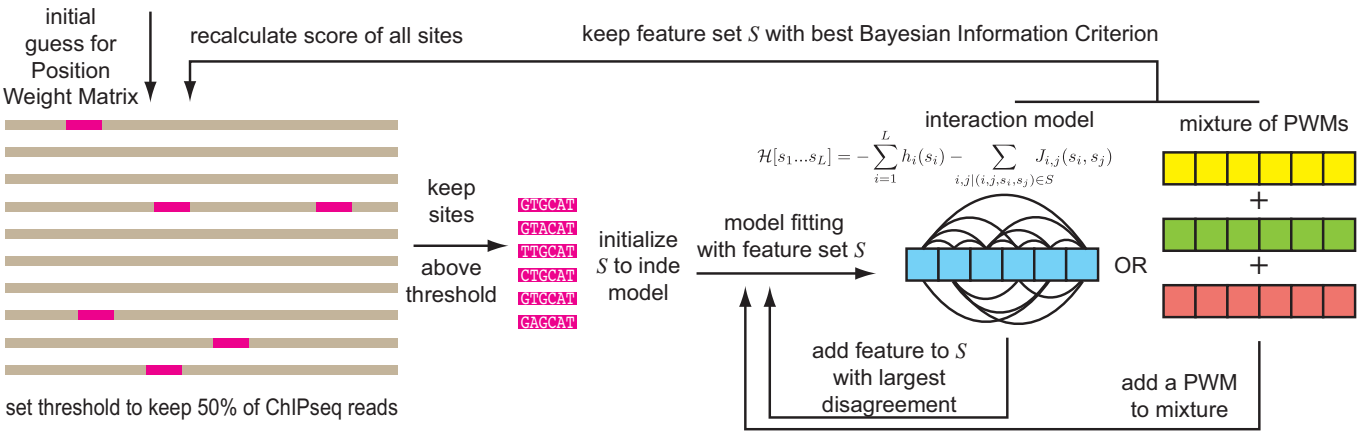


FIG. 1: **Workflow.** An initial Position Weight Matrix (PWM) is used to find a set of binding sites on ChIPseq data. Models are then learned using single-point frequencies (independent), two-point correlations (pairwise) or a mixture of independent models learned on sites clustered by K-Means (mixture) with increasing complexity, *i.e.* increasing number of features in the model. Finally the models with best Bayesian Information Criteria (BIC) are used to predict new sites until convergence to a stable set of sites.

cantly larger than expected from finite size sampling. In the following we focus on these 22 factors for which the PWM description of the TFBSs needs to be be refined. It can be noted that the 6 factors for which the PWM description appears satisfactory are predominantly those for which the smallest number of ChIP sequences is available (see Table 1 and Figure S1).

Pairwise interactions in the binding energy improve the TFBS description

The discrepancy between the observed statistics of TFBSs and the statistics predicted by the PWM model calls for a re-evaluation of the PWM main hypothesis, namely the independence of bound nucleotides. As recalled above, the inverse problem of devising interaction parameters from observed frequencies of "words" has been recently studied in different contexts. It has been proposed to include systematically pairwise correlations between the "letters" comprising the words to refine the independent letter description. In the case of a two-letter alphabet, the obtained model is equivalent to the classical Ising model of statistical mechanics[30]. In the present case, the 4-nucleotide alphabet (A,C,G,T) leads to a model equivalent to the so-called inhomogeneous Potts model [30] (hereafter called pairwise interaction model), a generalization of the Ising model to the case where spins assume q values and their fields and interaction parameters depend on the sites considered. In this analogy, nucleotides are spins with $q = 4$ colors.

In practice, the probability of observing a given word $(s_1...s_L)$ in the dataset is expressed as $P[s_1...s_L] = (1/\mathcal{Z})\exp(-\mathcal{H}[s_1...s_L])$, where $\mathcal{Z}$ is a normalization constant. $\mathcal{H}$ is formally equivalent to a Hamiltonian in the language of statistical mechanics, and reads:

$$\mathcal{H}[s_1...s_L] = -\sum_{i=1}^{L} h_i(s_i) - \sum_{i=1}^{L}\sum_{j<i} J_{i,j}(s_i,s_j), \quad s_i \in \{A,C,G,T\} \tag{1}$$

The "magnetic fields" h_i at each site i, along with the interaction parameters J_{ij} between nucleotides at positions i and j, are computed so as to reproduce the frequency of nucleotide usage at each position in the TFBS as well as the pairwise correlations between nucleotides at different positions (see *Methods*). In principle, the number of parameters in the model is sufficient to reproduce the observed values of all pairwise correlations between nucleotides. This however would result in over-fitting the finite-size data with an irrealistically large number of parameters. Therefore, to obtain the model parameters we instead maximized the likelihood that the data was generated by the model with a penalty proportional to the numbers of parameters involved, as provided by the Bayesian Information Criterion (BIC) [31]. Similarly to the procedure followed for the PWM, the pairwise interaction model and the collection of TFBSs for a given factor were iteratively refined together, as schematized in Figure .

Figure 3 shows the improvement in the description of TFBS statistics when using the final pairwise interaction model, for the three factors chosen for illustrative

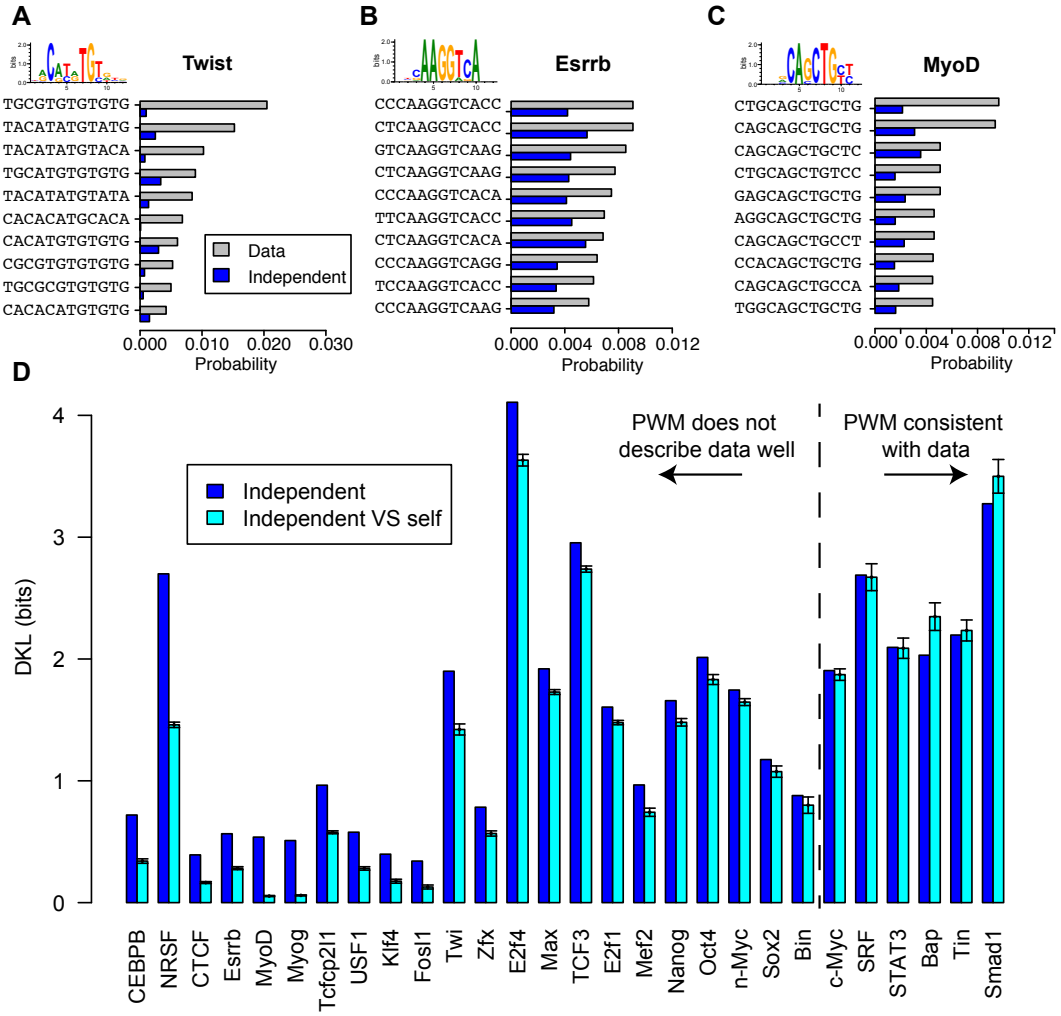


FIG. 2: **Observed TFBS frequencies are poorly predicted by a PWM model.** The observed frequencies of the most represented binding site sequences for the TF Twist (A), Esrrb (B) and MyoD (C) are shown (gray bars) as well as the probabilities of these sequences as predicted by the PWM model (blue bars). (D) Kullback-Leibler Divergence (DKL) between the observed probability distribution and the independent model distribution (blue). As a control we show the mean (cyan bars) along with two standard deviations of the DKL between the independent model and a finite sample drawn from it (see Methods). A discrepancy between the observed and predicted sequence probabilities is reported for 22 out of 28 factors.

purposes. Where the independent model failed at reproducing the strong amplitude and non-linear decrease in the frequencies of the most over-represented TFBSs, the pairwise interaction model provides a substantial improvement in reproducing the observed statistics. The improvement is most apparent when comparing the frequencies of the ten most observed TFBSs between the model and the ChIPseq data (Figure 3 A, C, E), and is further shown by the statistics of the full collection of TFBSs (Figure 3 B, D, F).

The pairwise model ranks binding sites differently from the PWM

Precise predictions of TFBSs are one important output of ChIPseq data. Moreover, they condition further validation experiments such as gel mobility shift assays or mutageneses. We therefore found it worth assessing the difference in TFBS predictions between pairwise and independent models.

First, we compared the set of ChIP sequences retrieved by the independent and pairwise models model at the cutoff of 50% TPR (True Positive Rate) used in the learning scheme, as shown in Figure 4A. The non overlapping set of ChIPseq sequences (*i.e.* sequences that were picked by one model but not by the other) was found to range

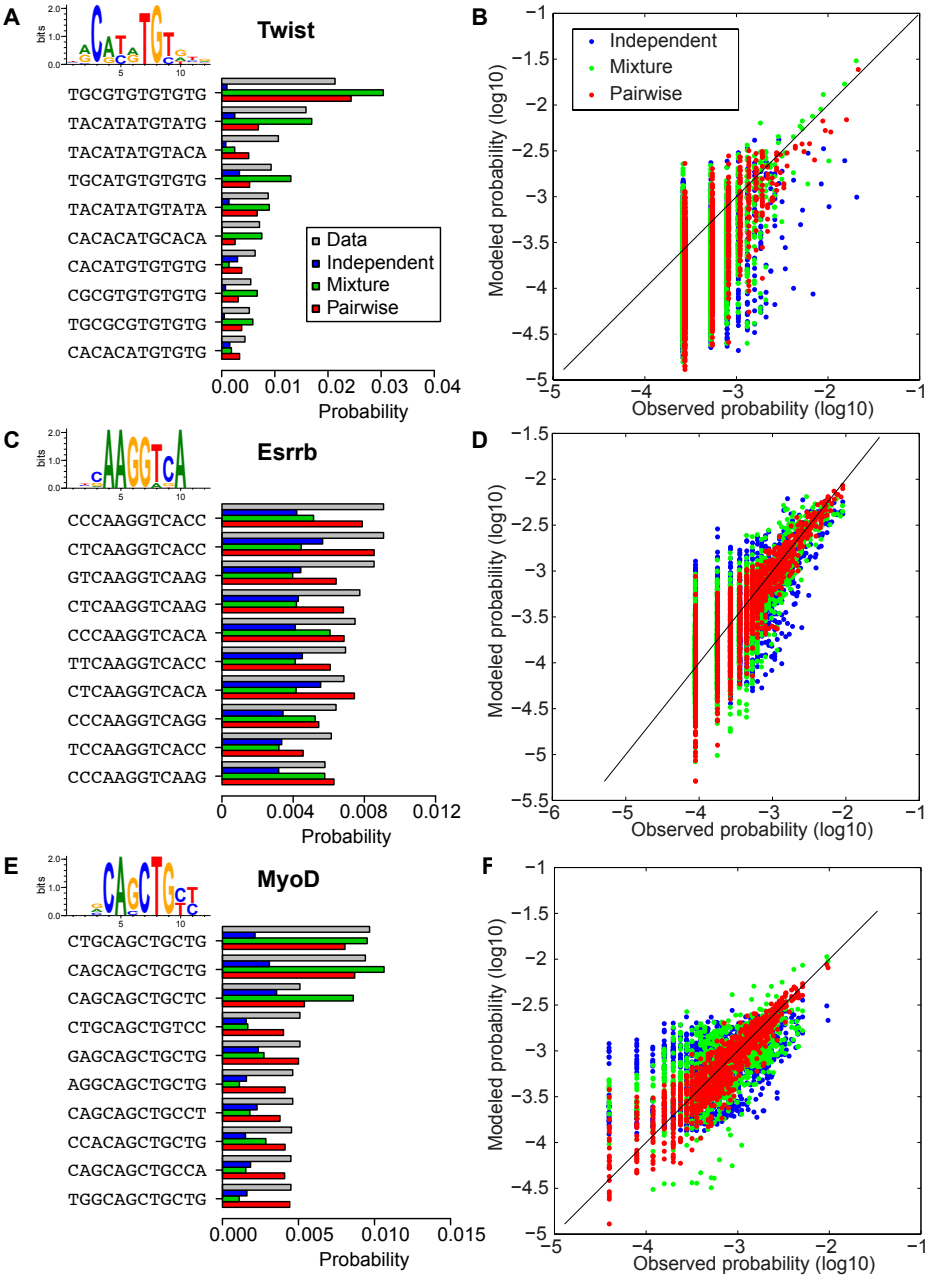


FIG. 3: **Models with correlations improve TFBS statistics prediction.** The observed frequencies (gray bars) of the most represented TFBSs for Twist (A), Esrrb (B) and MyoD (C) TFs, are shown together with the probabilities of these sequences predicted by the independent energy model (blue bars), the pairwise model taking into account interactions between nucleotides (red bars), and the K-means mixture model (green bars). (B,D,F) show the comparison between frequencies for all binding sequences and predicted sequence probabilities for the three models (same color code). The probability predictions of the pairwise model and to a lesser extent of the mixture model are in much better agreement with the observed frequencies than those of the PWM model.

from a few percent for TF like Esrrb, up to about 15 % for Twist. Thus, even when stemming from the same ChIPseq data, the two models can be learnt from significantly distinct set of sites.

Second, using the set of ChIPseq peaks on which the pairwise model was learned, we looked for the best predicted sites on each ChIPseq bound fragment using both the pairwise and PWM models (Figure 4B).

The overlap was found to be about 80% on average. The overlap between the sets comprising the two best TFBSs of each ChIPseq was also computed. This resulted in an overlap increase or decrease between the

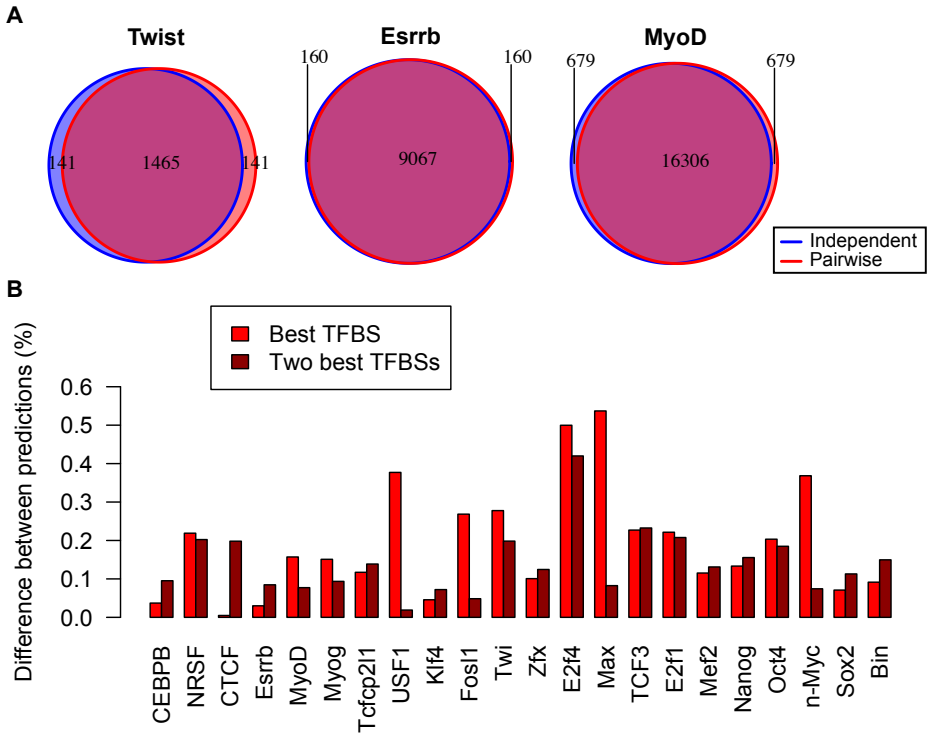


FIG. 4: **Overlap between predicted sites.** (A) Venn diagrams showing the overlap between the ChIP predicted by the independent (blue) and pairwise (red) models. (B) Difference (one minus the proportion of shared sites) between the best sites predicted by pairwise and PWM models on ChIPseq peaks (light red), and the same quantity when including the next best predicted sites on each peak (dark red). In several cases (*e.g.* Fosl1, Max, n-Myc, Srf, Stat3, Usf1), the difference between predicted sites is much smaller when the two best sites are considered, indicating that the pairwise model and the PWM model rank differently the two best sites in ChIP peaks with multiple bound sites.

prediction of the two models depending on the average of number of binding sites per retrieved ChIPseq fragment. In a few cases (*e.g* CTCF, Esrrb), the inclusion of the second best TFBS increased the difference between the two models. This generally happened when the ChIPseq fragments were retrieved with typically a single TFBS above threshold (*e.g.* for Essrb the TFBS specificity was fixed to retrieve 50% of 18453 ChIPseq and about 11000 fragments where found by the two models—see Table I). In these cases, the low specificity TFBSs tended to differ more between the two models than the very specific ones. In several other cases (*e.g* for Fosl1, Max, n-Myc, USF1), the inclusion of the second best predicted binding sites (Figure 4B) greatly increased the overlap between the two model predictions. This corresponded to cases for which the retrieved fragments contained on average two of more TFBSs about the specificity threshold (Table I). This showed that for these cases the prediction difference between the two models arose predominantly from a different ranking of the best TFBSs.

In conclusion, the TFBS predictions made by the two models can differ significantly both in the rank of ChIPseq fragments and in the rank of binding sites on these fragments.

Comparison with a PWM-mixture model

When described by a PWM, the binding energies of a TF for different nucleotides sequences form a simple energy well with a single minimum at a preferred consensus sequence. Some authors have instead analyzed the binding specificity of transcription factors by introducing multiple preferred sequences [24, 25]. A model of this type that naturally generalizes the PWM description consists of using multiple PWMs [14]. We found it interesting to investigate this approach based on a mixture of PWMs and compare it with the pairwise interaction model to get some insights into potentially important high-order correlations that would not be captured by the pairwise model. As precisely described in *Methods*, an initial mixture of K PWMs was generated by grouping into K clusters the TFBS data for a given TF. Similarly to the pairwise interactions, the number of clusters K was constrained, to avoid over-fitting, by penalizing the corresponding model score using the BIC. For a given TF, the PWM mixture and the collection of TFBSs in the ChIPSeq data were refined iteratively until convergence, usually reached after 10 iterations. The results are shown in Figure 5A for the three representative factors,

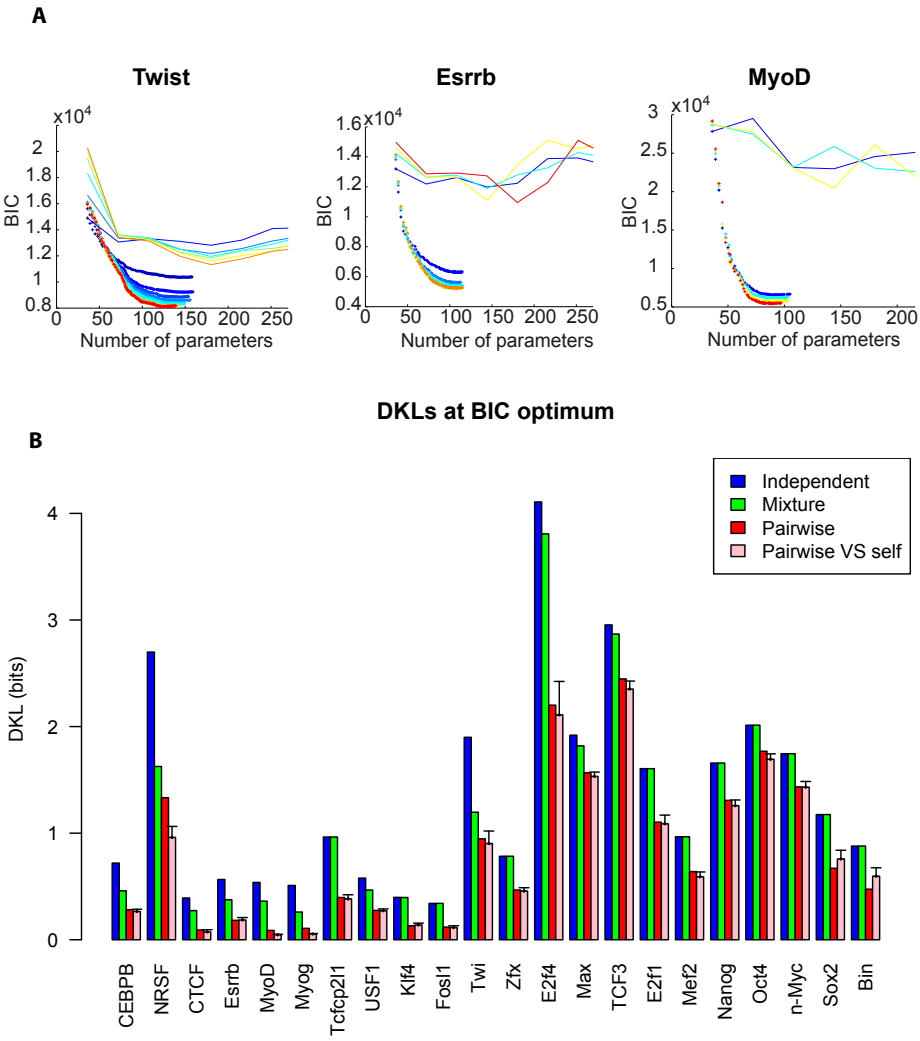


FIG. 5: **Model selection.** (A) Minimisation of the Bayesian information criterion (BIC, see *Methods*) is used to select the optimal number of model parameters and avoid over-fitting the training set. The evolution of the BIC is shown for the pairwise model (crosses) and the PWM-mixture model (lines). Colors from dark blue to red indicate the number of interations (see Fig.).
(B) Kullback-Leibler divergences (DKL) between the independent, K-means and pairwise distributions and the observed distribution for the different TFs, for the BIC optimal parameters. We also show the DKL of the pairwise model with a finite-size sample of sequences drawn from it (pink, see *Methods*). Error bars represent two standard deviations.

Twi, Esrrb and MyoD.

The best description of Twi ChIPSeq data is, for instance, provided by a mixture of 5 PWMs, which corresponds to 184 independent parameters. The mixture model yields a significant improvement when compared to the single-PWM model for Twi, and milder ones for Essrb and MyoD. In the three cases however, it proves inferior to the pairwise model.

More generally, Figure 5B shows the performances of the different models for all studied TFs using the Kullback-Leibler Divergence or DKL between the data distribution $P(s)$ and the models distributions $P_m(s)$. On the one hand, the mixture model improves the description of the binding data for 12 out of 27 TFs as compared to the single PWM model. The mixture model gives in particular strong improvements in the cases for which the binding sites have a palindromic structure (eg Twi, MyoD, Myog, Max, USF1). This feature often stems from the fact that the TF binds DNA as a dimer, which could give some concreteness to the mixture model: the recruitment of different partners by bHLH factors like MyoD or Myog could indeed lead to a mixture of TFs binding the same sites. On the other hand, the pairwise model clearly outperforms the other models in all cases studied.

As in the PWM case, the finite size of the datasets

leads us to expect fluctuations in the estimation of the DKL. In order to assess the magnitude of these finite-size fluctuations, we computed the average DKL between the best-fitting (pairwise) model and a finite-size artificial sample drawn from its own distribution, as shown in Figure 5B. Values of this DKL that are larger than the one obtained with the real dataset are indicative of overfitting, while the opposite case would suggest that the model is incomplete. In all cases, however, the DKL obtained with this control procedure was within error bars of the value computed with respect to the observed sample, with the exception of NRSF, MyoD, and Myog, as seen in Figure 5B. Thus, the pairwise model is generally the best possible model, insofar as the available dataset allows us to probe.

The metastable states of the pairwise interaction model

In order to more directly relate the pairwise interaction and the mixture models, it is useful to consider the energy landscape of the pairwise interaction model in the space of all possible TFBSs. By contrast with the simple, single-minimum energy well of the PWM model, the pairwise interaction model has multiple metastable energy minima. The energy landscape of the pairwise interaction model can thus be seen as a collection of energy wells, each centered on its metastable energy minimum. The span of the different energy wells in sequence space can be precisely defined as the basins of attraction of the different metastable minima in an energy minimizing procedure (see *Methods*). This allows one to associate each observed TFBS to a particular energy minimum. This defines basins of attraction that are used to build representative PWMs for each metastable minimum together with a weight—the number of sequences in the basin of attraction—for this energy minimum. We compared each metastable minimum to the PWMs of the mixture model, by calculating the DKL between the PWM computed from the sequences in its basin of attraction and the PWMs of the mixture model. This gave an effective distance which allowed us to associate each metastable state to the nearest PWM of the mixture model.

Using this procedure, we computed the set of PWMs and weights corresponding to the 27 considered TF pairwise interaction models. Examples are shown in Figure 6. In the case of Twi, the PWMs of the pairwise model ("metastable PWMs") can be clearly associated to the $K = 5$ PWMs of the mixture model. For MyoD, three of the 5 "metastable PWMs" can be clearly assigned to PWMs of the mixture model. The other two have a more spread out representation. The case of Esrrb is similar with one "metastable PWM" in clear correspondence with one PWM of the mixture model, and the other one less clearly so. The correspondence between the two models is shown in Figure S2 for the other TFs for which the mixture model uses more than a single PWM. This representation allows one to identify some features captured by the pairwise model. For example, in the case of Twist, most of the correlations are coming from the two nucleotides at the center of the motif, which take mainly 3 values among the 16 possible: CA,TG and TA. In the case of MyoD, the representation makes apparent the interdependencies between the two nucleotides following the core E-Box motif, and the restriction to the three main cases of CT, TC and TT.

Properties of the pairwise interactions

The computation of the interaction parameters allows an analysis of some of their properties. In particular, it is interesting to quantify their strengths and measure the typical distance between interacting nucleotides. We address these two questions in turn.

The concept of Direct Information was previously introduced to predict contacts between residues from large-scale correlation data in protein families [33]. We used it here to measure the strength of the pairwise interaction between two nucleotides. Using the previously generated interaction parameters from the pairwise model, we built the Normalized Direct Information (NDI), a quantity which varies from 0 for non-existing interactions, to 1 when interactions are so strong that knowing the amino acid identity at one position entirely determines the amino acid identity at the other position (see *Methods*). Heatmaps displaying the results for the representative Twist, Esrrb and MyoD factors are shown in Figure 7 and in Figures S3 for the other factors. An important observation is that the direct information between different nucleotides is rather weak—usually smaller than 10%—but substantially larger than the direct interaction between nucleotides in the surrounding background (1-3%, see Figure S4). It is interesting to note that such weak pairwise interactions give rise to a substantial improvement in the description of TFBS statistics, similarly to what was previously found in a different context [16]. The pairwise interactions are furthermore observed in Figure 7 to be concentrated on a small subset of all possible interactions. This can be made quantitative by computing the Participation Ratio of the interaction weights, an indicator of the fraction of pairwise interactions that accounts for the observed Direct Information (see Methods). Here, typical values of $10 - 20\%$ were found (Figure 7 and Table I), showing that the interactions tend to be concentrated on a few nucleotide pairs.

The interaction weights can also be used to measure the typical distance between interacting nucleotides. To that purpose, we computed the relative weight of the Direct Information as a function of the distance between nucleotides (see Methods). Figure 8 A shows box plots that summarize the results for the considered Drosophilae and mammalian TFs. Both plots show a clear bias towards nearest-neighbor interactions with a strong peak at $d = 1$, and a rapid decrease for $d \geq 2$. Finally, the

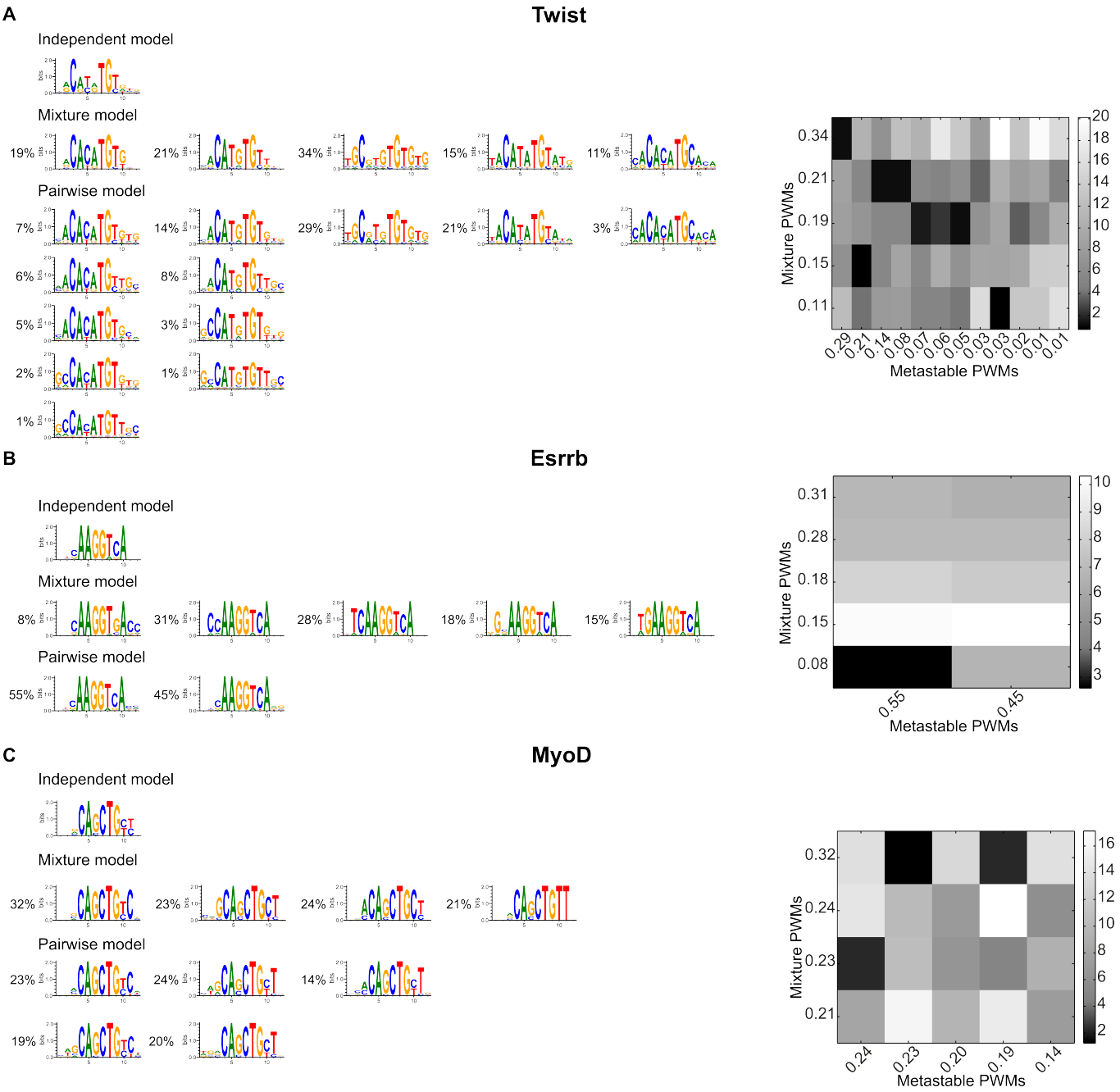


FIG. 6: **Metastable states.** The DNA sequence variety described by each model is illustrated using weblogos [32]. Shown are PWMs built from all sites, from the PWM-mixture model, and from the basins of attraction of the pairwise interaction model for Twist (A), Esrrb (B), and MyoD (C). The metastable PWMs are grouped under the mixture PWMs with smallest distance (measured by DKL, in bits). Heatmaps showing the DKLs between metastable PWMs and mixture PWMs are displayed on the right for each factor (minimal DKLs are in black). The proportions of sites used for each logo are also indicated and serve to denote the corresponding PWM.

dominant pair interactions are on average located in the flanking regions of the BS in clear anti-correlation with the most informative nucleotides which are on average in the central region (Figure 8 B). These observations for TF binding *in vivo* agree with similar ones made from a large recent analysis of TF binding *in vitro* [9]. The fact that for pair correlations to be important, nucleotide variation at a given location is required, may be one way to rationalize them.

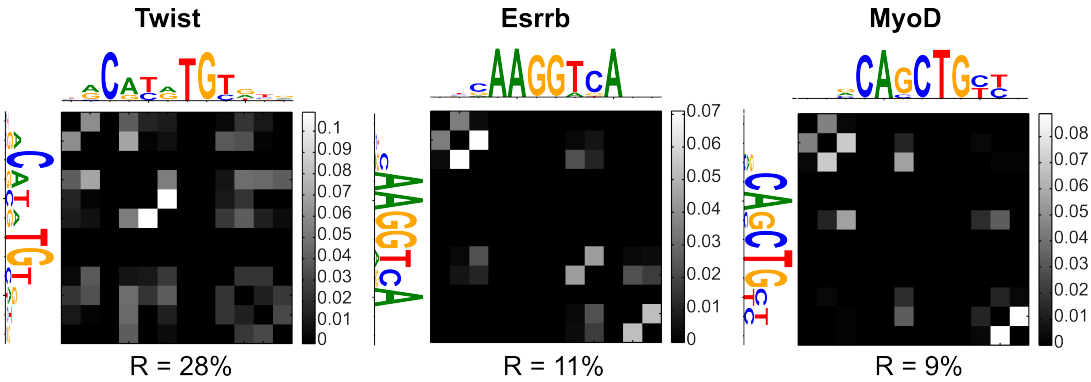


FIG. 7: **Nucleotide pair interactions.** Heat maps showing the values of the Normalized Direct Information between pairs of nucleotides. The matrix is symmetric by definition. PWMs are shown on the side for better visualization of the interacting nucleotides. The participation ratio R is indicated below each heat map.

TABLE I: **Participation Ratios**

Name	Part. Ratio
Bin	0.11
Mef2	0.19
Twi	0.28
E2f1	0.13
Esrrb	0.11
Klf4	0.16
Nanog	0.10
n-Myc	0.09
Oct4	0.24
Sox2	0.12
Tcfcp2l1	0.12
Zfx	0.10
CEBPB	0.05
CTCF	0.23
E2f4	0.14
Fosl1	0.09
Max	0.18
MyoD	0.09
Myog	0.09
NRSF	0.27
TCF3	0.19
USF1	0.07

Alternative representation of interactions by Hopfield patterns

Using a simple binary description of neurons, JJ Hopfield suggested, in a classic piece of work [34], that neural memories could be attractors corresponding to patterns arising from pair interactions between neurons. These interaction patterns can be computed in the present case. They offer an alternative way to analyze the patterns of correlation from the pair-interactions between positions, as already proposed in a mean-field context in [35]. Because the matrix of interactions J_{ij} is symmetric, it can be diagonalized in an orthonormal basis of eigenvectors ξ^k, the Hopfield patterns in the present case, with corresponding real eigenvalues λ_k. These orthonormal eigenvectors correspond to the Hopfield patterns in the present case. The Potts energy (Eq. (1)) for a binding sequence $s_1 \cdots s_L$ can be rewritten in terms of the Hopfield patterns as (see Methods):

$$\mathcal{H} = -\sum_i h_i(s_i) - \frac{1}{2}\sum_{k=1}^{4L} \lambda_k \left(\sum_{i=1}^{L} \xi_i^k(s_i) \right)^2 . \qquad (2)$$

Although here the presence of the diagonal h term prevents the patterns to be metastable energy states, they can still be useful to analyze the interaction matrix. This spectral decomposition of the interaction matrix is also similar in spirit to a principal component analysis, and even equivalent in the case of Gaussian variable. One can thus wonder how many patterns are needed to well approximate the full matrix of interactions J. To address this question, one can rank the eigenvalues λ_k in order of decreasing moduli and note J_p the restriction of the interaction matrix generated by the first p eigenvalues and their associated patterns. The full interaction matrix naturally corresponds to J_{48}. Approximate interaction matrices obtained by keeping different numbers of dominant patterns are shown in Figure 9 for the three considered representative factors. Pairs of successive patterns appear to provide the main interaction domains in this representation, as is particularly clear in the case of MyoD. One can see in Figure 9 that J_6 already closely approximates the full interaction matrix, a reflection in the present representation, that the important interactions are concentrated on a few links between pairs of nucleotides.

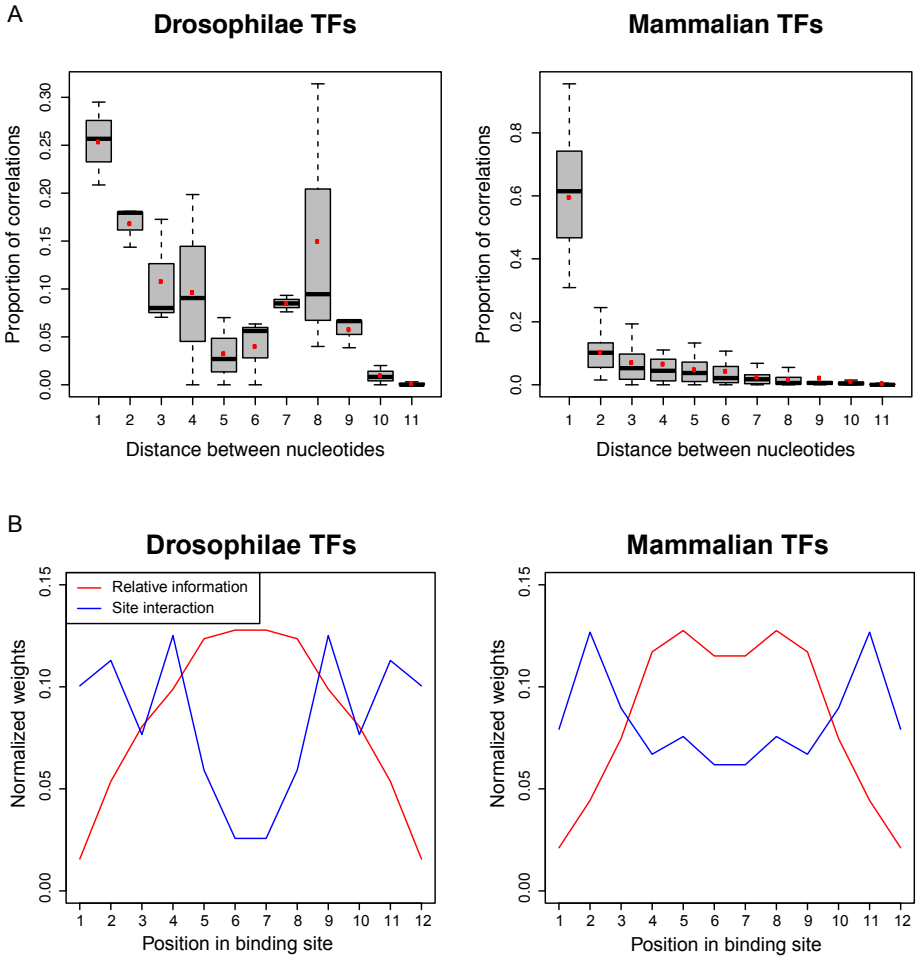


FIG. 8: **Properties of the pair interactions.** (A) Distances between interacting nucleotides. The box plots show the relative importance of the Normalized Direct Information as a function of the distance between interacting nucleotides. Red dots denote average values. (B) Sum of normalized direct informations in the TFBSs at a given position, averaged over all considered factors (blue line). The average site information content relative to background as a function of position is also shown (red line). In both quantities, the average over the two TFBS orientations has been taken.

Discussion

The availability of ChIPseq data for many TFs has led us to revisit the question of nucleotide correlations in TFBSs. In order to perform a fully consistent analysis of this type of data, we have developed a workflow in which the TFBS collection and the model describing them are simultaneously obtained and refined together. We have found that when a sufficiently large number of TFBSs is available, the PWM description does not account well for their statistics. The general presence of correlations that follows from this finding, agrees with previous reports for particular transcription factors [6, 8, 24] and with the conclusions of large scale *in vitro* TF binding studies [9, 10].

In order to refine the PWM description, we have analyzed a model with pairwise interactions [23], and a PWM mixture model [14]. Data overfitting is a concern for multi-parameter models and has been addressed by putting a penalty on the parameter number using the BIC. While the mixture-model improved in some cases the PWM description, especially for palindromic binding sites, a much more significant and general improvement was found with the pairwise interaction model. The success of the pairwise interaction model agrees with the results of its recent application (however, without the BIC) to high-throughput *in vitro* binding data [23]. It moreover shows that, at least in the case we considered, pairwise interactions are sufficient to account for higher-order correlations, and that an explicit description like the one provided by the PWM-mixture model is not necessary. For example, for Essrb, metastable states arising from nearest-neighbor interactions reproduce a triplet of flanking nucleotides with a variable spacer from the core motif (Figure S5).

Our detailed analysis of the obtained interaction mod-

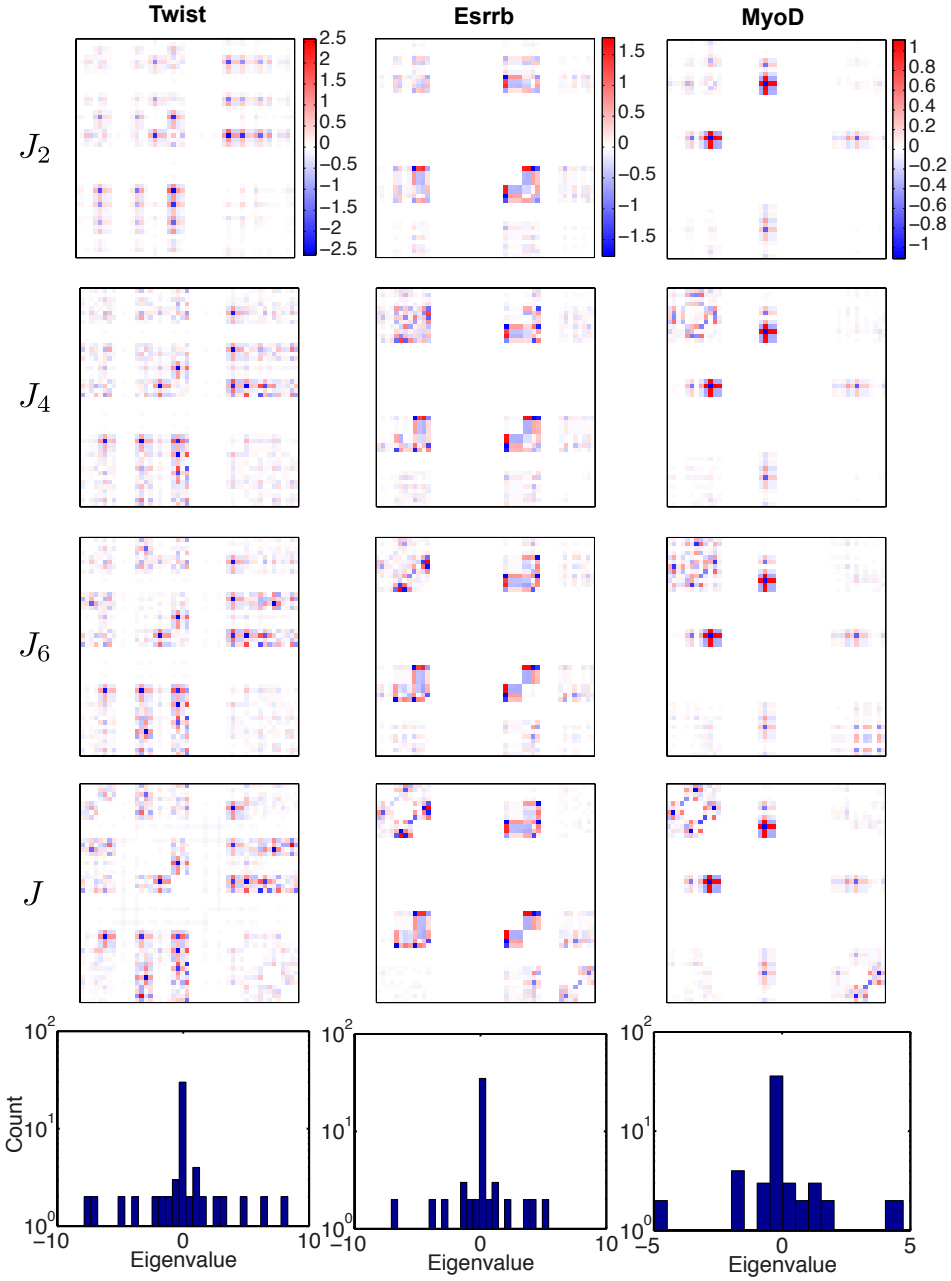


FIG. 9: **Representation of interactions by Hopfield patterns.** The full interaction matrix J is approximated by a matrix J_p built from the p Hopfield patterns with highest eigenvalue moduli. We show J_2, J_4, J_6 and the full matrix J in the basis (i, b) with $i = \{1, ..., 12\}$ and $b =$\{A,C,G,T\}. Color bars are shown on the first row for each factor. For MyoD, the correspondence between successive pairs of patterns and distinct interaction domains is seen particularly clearly. In all cases the full interaction matrix is already well approximated by J_6.

els for different TFs shows that the weights of pairwise interactions are generally weak. The most important are only about 10 % of the PWM weights, but significantly above the interaction weights in the surrounding background DNA (of the order 1-3% by the same measure). Nonetheless, collectively these interactions significantly improve the model description of the TF binding data as found in other examples [16].

We have here obtained the pairwise interaction models based on the principle of maximum entropy, constrained to account for the pair-correlations measured in the data. This approach has already been followed in a variety of biological contexts, from populations of spiking neurons [16, 17] to protein sequences [20] to bird flocks [22]. An interesting feature of these interaction models is their non-convexity, which allows for the existence of many lo-

cal maxima in the probability distribution of sequences, or local minima of energy. This was noted for repertoires of antibodies in a single individual [21], where many of these local states were observed and suggested as possible signatures of past infections. In a very different context, local probability maxima in the probability distribution of retinal spiking patterns was reported and linked to error-correcting properties of the visual system [36]. In the present case of TFBSs, these local mimima reflect the multiplicity of binding solutions and resemble the individual PWMs of the mixture model. Pairwise interaction models thus somehow incorporate models of multiple PWMs while outperforming them.

The previously considered case of protein sequences shares many similarities to the statistics of TFBSs, since correlations in protein sequences as in TFBSs reflect both structural and functional contraints. In proteins families, correlations are usually interpreted as resulting from the co-evolution of residues interacting with each other in the protein structure. These effects are hard to distinguish from phylogenic correlations or other observational biases. Nonetheless, the inference of interaction models from data was successfully used to predict physical contacts between amino-acids in the tertiary structure [37], and to aid molecular dynamics simulations in predicting protein structure [38–40]. In the case of TFBSs, comparison between *in vitro*[9, 10] and *in vivo* binding data may help to disentangle the different possible origins of the found correlations and seems worth pursuing. It appears similarly interesting to study how much of the found pair correlations can be explained on the basis of structural data. Finally, the role of nucleotide interaction in TFBS evolution [41] should be considered and could improve the reconstruction of TFBSs from multi-species comparison [42–44].

Independently of these future prospects, we have found that the TFBSs predicted from ChIP-seq data significantly depended on the model used to extract them. Since the pairwise interaction model and the developed workflow significantly improve TFBS description and require a modest computational effort, they should prove worthy tools in future data analyses.

Materials and Methods

Genome-wide data retrieval

We use both ChIP-on-chip data from *Drosophila Melanogaster* and ChIPseq data from *Mus Musculus*. Data was retrieved from the litterature [26, 27] and from ENCODE data accessible through the UCSC website `http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCaltechTfbs/`, for a total of 27 TFs. Among them, there are 5 developmental Drosophilae TFs: Bap, Bin, Mef2, Tin and Twi, 11 mammalian stem cells TFs: c-Myc, E2f1, Esrrb, Klf4, Nanog, n-Myc, Oct4, Sox2, Stat3, Tcfcp2l1, Zfx, and 11 factors involved in mammalian myogenesis: Cebpb, E2f4, Fosl1, Max, MyoD, Myog, Nrsf, Smad1, Srf, Tcf3, Usf1. Overall, there are between 678 and 38292 ChIP peaks, with average size 280bp. DNA sequences were masked for repeats using RepeatMasker [45].

Background models

It is important to discriminate the statistics of the motifs proper from that of the background DNA on which motifs are found. Besides particular nucleotides frequencies, the background DNA can exhibit significant nucleotide correlations, for instance arising from CpG depletion in mammalian genomes (Figure S4). For each ChIPseq data, we used, as background, all sites from both strands of the sequences. This serves to learn independent and pairwise background models which were used as reference models to score the corresponding TFBS models. The position information content in all plotted PWM logos is measured with respect to the nucleotide background frequencies (*i.e.* the independent background model)

Initial PWM refinement

Along with the ChIPseq data for the different factors, we also retrieved corresponding PWMs from the literature [26] or from TRANSFAC database [46]. These initial PWMs were refined according to the following protocol.

Given ChIPseq data (bound regions) for a given TF and an initial PWM of length L ($L = 12$ was taken for all computations in the present paper), we scanned both strands of each bound region and attributed to all observed L-mers a score defined as the ratio between the PWM and background models probabilities. A cutoff was set such that half of the bound regions had at least one predicted TFBS with a score above the cutoff, setting a True Positive Rate (TPR) of 50%. This heuristic criterium overcame the problem of False Positives among the ChIPseq peaks that might have polluted the data. This defined a training set of N L-mers with probability higher than the cutoff. Bound sites were again predicted using the same cutoff. This procedure was repeated until stabilization of the predicted sites to a fixed subset. This resulted in a refined PWM with its associated set of bound sites.

Independent model evaluation

The independent model consist of a matrix of single nucleotide probabilities of size $4 \times L$, where L is the width of the binding site. In a first approximation, the parameters appearing in the matrix can be estimated from a set of binding sites by computing the observed frequency $f_{b,i}$ of

TABLE II: **Information about the TFs**

Name	N_{chip}	$\Delta_{\text{inde-mixture}}$	$\Delta_{\text{inde-pairwise}}$	$\Delta_{\text{mixture-pairwise}}$	N_{inde}	N_{mixture}	N_{pairwise}
Bap	678	0	12	12	2205	2208	2117
Bin	1857	2	80	81	1300	1298	1228
Mef2	4545	0	161	161	3681	3681	3665
Tin	1791	0	40	40	1333	1333	1310
Twi	3211	182	141	128	3810	3862	3722
c-Myc	3038	0	95	95	2996	2996	2920
E2f1	17367	0	877	877	16625	16625	14915
Esrrb	18453	172	160	167	11243	11333	11275
Klf4	9404	0	97	97	5912	5912	5913
Nanog	8022	0	111	111	6196	6196	6224
n-Myc	6367	0	54	54	6981	6981	6954
Oct4	3147	0	74	74	3187	3187	3079
Smad1	907	0	24	24	690	690	667
Sox2	3523	0	95	95	2306	2306	2293
STAT3	2099	54	58	62	2308	2264	2231
Tcfcp2l1	22406	0	418	418	16691	16691	16649
Zfx	9152	0	203	203	6473	6473	6473
CEBPB	14500	399	337	334	8275	8322	8267
CTCF	32958	360	492	579	17087	17098	17060
E2f4	4132	248	590	517	4643	5146	3879
Fosl1	5981	0	90	90	5088	5088	5039
Max	8751	24	70	81	12531	12495	12386
MyoD	33969	717	679	665	25416	25430	25344
Myog	38292	1116	584	835	29520	29334	29647
NRSF	13756	639	672	488	13183	14363	13440
SRF	2370	1	34	35	2929	2928	2948
TCF3	9453	185	277	257	8528	8690	8775
USF1	8956	11	14	12	8628	8619	8625

For each TF, we show the number N_{chip} of ChIP sequences retrieved, the numbers $\Delta_{\text{inde-pairwise}}$, $\Delta_{\text{inde-mixture}}$, and $\Delta_{\text{pairwise-mixture}}$ of different ChIP sequences used for training between either two models, and the numbers N_{inde}, N_{mixture}, and N_{pairwise} of TFBSs used to learn each model.

nucleotide b at position i. However, this frequency fluctuates around the "true" probability due to finite sample size, and for example unobserved nucleotides could actually have a low probability of being observed provided that the number of observations be high enough. It is usual to correct for this effect by using the Bayesian pseudo-count approach stemming from Laplace's rule of succession [3]. The probability to observe nucleotide b at position i is given by:

$$p_{i,b} = \frac{n_{i,b} + \alpha_b}{N + \sum_b \alpha_b} \tag{3}$$

where $n_{i,b}$ is the number of observed b at position i, N is the total number of sites, and α_b's are the pseudo-counts, or prior probabilities to observe nucleotide b at position i. The pseudo-counts were all set to 1, however no significant effect was noted when changing this value, as expected from the large number of observations.

Kullback-Leibler divergence

The Kullback-Leibler divergence is a measure of distance between two probability distributions p and q of a variable s, and is defined as:

$$\text{DKL}(p\|q) = \sum_s p(s) \log \frac{p(s)}{q(s)}. \tag{4}$$

Throughout this paper, when a DKL is calculated between a finite sample and a model distribution, p corresponds to the sites frequencies in the sample, and q to the model distribution. When the DKL is calculated between a PWM of a basin of attraction of a metastable

state and a PWM from the mixture model, p is used for the former, and q for the latter.

Estimation of the fluctuations due to finite sampling: DKL *vs* self

To estimate whether the description of the data by a model (*e.g.* independent or pairwise) could be improved or was consistent with the finite number N of observed sequences, we computed the 'self' DKL between the distribution of a set of N sequences drawn from the model distribution and the model distribution itself. This procedure was repeated 100 times. TFs for which the independent model DKL was smaller than or within two standard deviations of the self DKL were discarded for later analysis.

Derivation of the pairwise interaction model

Information theory offers a principled way to determine the probabilities of a set of states given some measurable constraints. It consists in maximizing a functional known as the entropy[47, 48] over the set of possible probability distributions given the imposed constraints. Here, we wish to determine the probability $P(s)$ of a DNA sequence s of length L, in the set of TFBSs for a transcription factor, given the constraints that the probability distribution P retrieves the one- and two-point correlations observed in a set of bound DNA sequences. We denote by $\mathcal{A}$ the alphabet of possible nucleotides, $\mathcal{A} = \{A, C, G, T\}$ and by s_i the nucleotide at position i in the sequence s so that $s = s_1 \cdots s_L$. With these notations, the entropy with the considered constraints translates into the the following functional:

$$\begin{aligned}\mathcal{L} = -\sum_{\{s\}} P(s)\ln P(s) + \lambda\left(\sum_{\{s\}} P(s) - 1\right) + \sum_{i=1}^{L}\sum_{a\in\mathcal{A}} h_i(a)\left(\sum_{\{s\}} P(s)\delta(s_i,a) - P_i(a)\right) \\ + \sum_{i=1}^{L-1}\sum_{j>i}\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}} J_{i,j}(a,a')\left(\sum_{\{\sigma\}} P(a)\delta(s_i,a)\delta(s_j,s') - P_{i,j}(a,a')\right),\end{aligned} \tag{5}$$

where $P_i(a)$ (resp. $P_{i,j}(a,a')$) is the probability of having nucleotide a at position i (resp. nucleotides a and a' at position i and j) in the TFBS data set. The function δ denotes the Kronecker $\delta-$function defined by $\delta(a,a') = 1$ if $a = a'$,and 0 otherwise. The first term in Eq. (5) is the entropy of the probability distribution to be found and the other terms are the given constraints along with their Lagrangian multipliers. Maximization of the functional $\mathcal{L}$ is performed in a usual way by setting the functional derivative with respect to the probability distribution P to zero:

$$\frac{\delta\mathcal{L}}{\delta P(s)} = 0 = -\ln P(s) - 1 + \lambda + \sum_{i=1}^{L} h_i(s_i) + \sum_{i=1}^{L-1}\sum_{j>i} J_{i,j}(s_i,s_j). \tag{6}$$

Finally, using the constraint $\sum_{\{s\}} P(s) = 1$, one finds the probability distribution that maximizes entropy given the constraints that it reproduces the observed one- and two-point correlations:

$$P[s] = \exp[-\mathcal{H}(s)]/\mathcal{Z}, \tag{7}$$

where $\mathcal{H}(s)$ is the inhomogeneous Potts model Hamiltonian,

$$\begin{aligned}\mathcal{H}[s_1...s_L] = -\sum_{i=1}^{L} h_i(s_i) - \sum_{i=1}^{L}\sum_{j<i} J_{i,j}(s_i,s_j), \\ s_i \in \{A, C, G, T\}.\end{aligned} \tag{8}$$

The normalization constant $\mathcal{Z}$ is the partition function,

$$\mathcal{Z} = \sum_{\{s\}} \exp[-\mathcal{H}(s)]. \tag{9}$$

Gauge fixing

The probability distribution of sequences, as given by Eqs. (7, 8), is invariant under shifts of the local fields $h_i(a)$ and under transformations between the interaction terms $J_{i,j}(a,a')$ and the local fields. In order to uniquely determine $\mathcal{H}$, this arbitrariness needs to be taken care of by adding further conditions that uniquely fix the in-

teraction parameters, a process known as gauge fixation [20] that we detail here.

a. Local fields. Since it amounts to changing the reference energy and is cancelled by the normalization, the probability is invariant with respect to the following global shift of the $h_i(a)$

$$h_i(s_i) \to \tilde{h}_i(s_i) = h_i(s_i) + \varepsilon_i. \tag{10}$$

We choose to fix this invariance by minimizing the square norm $S_i = \sum_{a\in\mathcal{A}} \tilde{h}_i(a)^2$ of local field terms with respect to the gauge degree of freedom. The corresponding gauge-fixing condition reads

$$\sum_{a\in\mathcal{A}} \tilde{h}_i(a) = 0. \tag{11}$$

This condition can be imposed on any set of fields h_i by using the symmetry (10) and redefining the fields as follows,

$$h_i(s_i) \to h_i(s_i) - \frac{1}{4}\sum_{a\in\mathcal{A}} h_i(a). \tag{12}$$

b. Interaction terms. Another invariance stems from the fact that contributions can be shifted between local fields and interaction energies. Namely, the following change of variables does not affect the probability:

$$J_{ij}(s_i,s_j) \to \tilde{J}_{ij}(s_i,s_j) = J_{ij}(s_i,s_j)+\psi_i(s_i)+\phi_j(s_j)+C_{i,j}, \tag{13}$$

since the local fields ψ_i and ϕ_j can be redistributed in h and the constant $C_{i,j}$ gives an energy reference for the interacting nucleotides that is cancelled by the normalization process. Again, a gauge condition is obtained by minimizing the square norm $S_{i,j} = \sum_{a,a'\in\mathcal{A}}[\tilde{J}_{ij}(a,a')]^2$ of interaction terms with respect to the gauge degrees of freedom. This yields the conditions:

$$\sum_{a\in\mathcal{A}} \tilde{J}_{i,j}(a,a') = \sum_{a'\in\mathcal{A}} \tilde{J}_{i,j}(a,a') = 0. \tag{14}$$

These can be imposed on a set a of J_{ij} parameters by redefining them as follows:

$$\begin{aligned} J_{ij}(s_i,s_j) \to & J_{ij}(s_i,s_j) + \frac{1}{16}\sum_{a,a'\in\mathcal{A}} J_{i,j}(a,a') \\ & - \frac{1}{4}\sum_{a\in\mathcal{A}} J_{i,j}(a,s_j) - \frac{1}{4}\sum_{a\in\mathcal{A}} J_{i,j}(s_i,a). \end{aligned} \tag{15}$$

Determination of the pairwise interaction model from the data

The parameters of the inhomogeneous Potts model in Eq. (8), giving the energy of an observed sequence of length L, must be computed from the data. The parameters h and J represent the energy contributions respectively coming from individual nucleotides and from their interactions. The PWM model is the particular case where all the interaction parameters vanish: $J_{i,j}(a,a') = 0$.

To build the model, we start from the PWM description, characterized by the set of initial $h_i(a) = \log p_{i,a}$ and the interaction parameters J's set to zero. We add one interaction parameter $J_{i,j}(a,a')$ at a time, corresponding to the pair of nucleotides whose pairwise distribution predicted by the model differs most from data, as estimated by a binomial p-value. We then fit the augmented model to data, use this model to select a new set binding sites from the reads, and repeat the whole procedure. In each of these steps, fitting is performed by a gradient descent algorithm:

$$J \;\to\; J + \epsilon\left[c_2^{\text{data}} - c_2^{\text{model}}\right], \tag{16}$$

$$h \;\to\; h + \epsilon\left[c_1^{\text{data}} - c_1^{\text{model}}\right], \tag{17}$$

where c_1 and c_2 are matrices of size $4\times L$ and $4L\times 4L$ respectively corresponding to the single- and two-point frequencies, and superscripts denote whether the matrices are computed from the data or from the model distribution. This algorithm converges to the set of parameters $(\{\tilde{h}_i\},\tilde{J}_{i,j})$ that match all single marginals and the pairwise marginals of interest. The number of interaction parameters that are being added is controlled by the Bayesian Information Criterion, or BIC (Figure 5). The BIC computes the opposite log-likelihood and adds a penalty proportional to the number of parameters involved. This adverts the over-fitting of a finite dataset with an extravagant number of parameters. The procedure is iterated until minimization of the BIC, yielding the best pairwise model with the full set of parameters $(\{h_i(a)\},\{J_{i,j}(a,a')\})$. As in the case of the PWM model, we score each sequence using the ratio between the TF and background pairwise models and impose a score cutoff so as to select a set of bound sites yielding 50% TPR, on which a new pairwise model is learned. This process is iterated until convergence to a stable set of bound sites.

BIC computation

Consider a sample $X = (X_1,\ldots,X_N)$ of N TFBSs drawn from an unknown distribution function f we wish to estimate. To this extent, several models $\{M_1,\ldots,M_m\}$ are proposed, each model M_i having a density g_{M_i} with parameter θ_i of dimension K_i. It is straightforward to see that, as K_i increases, the fit to the observed sample as measured by the likelihood function $g_{M_i}(X|\theta_i)$ increases as well, the limiting case being when f is estimated as the sample distribution. However, such an estimator is inappropriate to account for new, yet unobserved TFBSs, *i.e.* it is not predictive. Such a case where the number of parameters used to estimate a distribution becomes of the order of the size of the sample is known as overfitting. The BIC allows to overcome overfitting by penalizing high dimension parameters. Using

Bayes Rule, and a uniform a priori distribution on the models, we have

$$P(M_i|X) \propto P(X|M_i). \tag{18}$$

That is, the probability of the model given the data can be inferred from the probability that the data is generated by the model. The latter is obtained by marginalizing the joint distribution of the data and the parameters over the space of parameters Θ:

$$P(X|M_i) = \int_\Theta P(X,\theta|M_i)d\theta = \int_\Theta g_{M_i}(X|\theta)P(\theta|M_i)d\theta. \tag{19}$$

For a unidimensional parameter θ, the likelihood $g_{M_i}(X|\theta)$ is maximized at some particular $\hat{\theta}_i$ with an uncertainty (or width) proportional to $1/\sqrt{N}$ in the limit of large N. Assuming a broad prior, then for large N the integral is dominated by the likelihood which is concentrated around its maximum. One can then approximate the integral by the area of the region of height the maximum likelihood and of width $1/\sqrt{N}$, that is $g_{M_i}(X,\hat{\theta}_i)/\sqrt{N}$. This result can be retrieved analytically using the method of steepest descent. For a number K_i of parameters, one gets a total volume $g_{M_i}(X,\hat{\theta}_i)/N^{K_i/2}$ [31]. Taking the logarithm yields the BIC condition:

$$BIC_i = -2\log(P(X|M_i)) \simeq -2\log(g_{M_i}(X,\hat{\theta}_i)) + K_i\log(N). \tag{20}$$

In the present case, the sample X is the set of observed TFBSs and the model M_i determines the probability $P_{M_i}(s)$ of belonging to X,

$$\log(g_{M_i}(X,\hat{\theta}_i)) = \sum_{s\in X}\log[P_{M_i(\hat{\theta}_i)}((s)]. \tag{21}$$

The interpretation of Eq. (20) is clear: adding new parameters improves the fit, but also adds new sources of uncertainty about these parameters due to the finite size of the data. This uncertainty disappears as $N \to \infty$, since the log-likelihood scales with N while the correction scales with $\log(N)$.

Finally, Eq. (20) is a functional over models, the chosen model M_{BIC} is the one that minimizes it,

$$M_{BIC} = \underset{M_i}{\operatorname{argmin}}\ BIC_i. \tag{22}$$

PWM mixture model

We investigated an approach based on a mixture of PWMs. For that purpose, we used a comparable setup as for the pairwise model. However, instead of adding correlations to a given PWM, new PWMs were added to a mixture model. More precisely, a mixture of K PWMs, with $1 \leq K \leq 10$, was generated by using a K-means algorithm with a Hamming distance metrics on the initial set of bound sites. This resulted in K clusters, each comprising n_k sites among the initial N sites. A PWM was generated on each of these clusters, with probability distribution $\mathcal{P}_k$. The mixture model of order K was then defined as [31]:

$$\mathcal{P}[s] = \sum_{k=1}^{K} p_k\mathcal{P}_k[s], \tag{23}$$

where $p_k = n_k/N$ is the cluster weight. Because a PWM has $3\times L$ degrees of freedom (L of them being constrained by the summation of nucleotide probabilities to one) and there are $K-1$ free weight parameters, the number of parameters corresponding to a mixture of order K is $3LK + (K-1)$. As previously, the model showing minimal BIC score was used for sites detection, a new set of PWMs and weights p_k was generated by clustering the set of detected sites and the procedure was iterated until convergence to a stable set of sites.

Metastable minima of the pairwise interaction model and their basins of attractions

We defined the basins of attraction of a pairwise interaction model energy landscape, in the following fashion. Let s be a site with energy $\mathcal{H}(s)$. We looked for the nucleotides that could be changed to minimize $\mathcal{H}(s)$. If such nucleotides existed, one of them was chosen at random, and its value was updated. One local minimum of the energy landscape, or metastable state, was reached when no such nucleotide could be found. The basin of attraction of a metastable state was then defined as the ensemble of sites that fell to this metastable state when their energy was minimized following the above procedure. We computed metastable states and their basins of attraction for the final set of bound sites obtained with the best pairwise model. A PWM was learned on each basin of attraction, leading to a set of representative PWMs, with different weights representing different proportions of bound sites in their basins.

Computation of the Direct Information

We wanted to build a quantity based solely on direct interactions $J_{i,j}$ between nucleotides, discarding indirect interactions. To this end, we used the interaction parameters obtained from the pairwise model to build the direct dinucleotide probability function:

$$P^d_{i,j}(a,a') = e^{\tilde{h}_i(a)+\tilde{h}_j(a')+J_{i,j}(a,a')}/\mathcal{Z}_{i,j}, \tag{24}$$

where

$$\mathcal{Z}_{i,j} = \sum_{a,a'} e^{\tilde{h}_i(a)+\tilde{h}_j(a')+J_{i,j}(a,a')}.$$

The 8 effective fields $\tilde{h}_i$ and $\tilde{h}_j$ were fully determined by the constraints that the direct probability function matches the observed one-point frequencies:

$$\begin{aligned} \sum_{a'} P^d_{i,j}(a,a') &= P_i(a), \qquad a' \in \{A,C,G,T\}, \\ \sum_{a} P^d_{i,j}(a,a') &= P_j(a'), \qquad a \in \{A,C,G,T\}. \end{aligned} \tag{25}$$

The normalization of the probabilities $\sum_a P_i(a) = 1$, served to reduce this system to 6 equations. The fields $\tilde{h}_i(a)$, which are determined up to a constant, were fixed by the gauge condition that they vanished for the nucleotide A, $\tilde{h}(A) = 0$. The system was solved using the Levenberg-Marquadt algorithm with $\lambda = 0.005$.

The Direct Information [37] was then defined as:

$$DI_{i,j} = \sum_{a,a'} P^d_{i,j}(a,a') \log_2 \left(\frac{P^d_{i,j}(a,a')}{P_i(a)P_j(a')} \right). \tag{26}$$

As there is no upper bound for this direct information, we built a normalized version of the direct information:

$$NDI_{i,j} = \frac{DI_{i,j}}{\sqrt{S_i S_j}}, \tag{27}$$

where S_i denotes the entropy at position i. Note that $S_i = DI_{i,i}$ so that $NDI_{i,i} = 1$ for this maximally correlated case. On the contrary, independent nucleotides give $NDI_{i,j} = DI_{i,j} = 0$.

Participation Ratio

For each TF, an interaction weight was defined for each pair of nucleotides as

$$w_{i,j} = NDI_{i,j} / \sum_{i \neq j} NDI_{i,j}. \tag{28}$$

Self-interactions have no meaning here and were attributed $w_{i,i} = 0$. Let us note $N = L(L-1)$ the number of possible interactions. Using our weight, one writes the Participation Ratio as:

$$R = \frac{1}{N \sum_{i \neq j} w^2_{i,j}}. \tag{29}$$

The interpretation is simple: if all weights are equal, $w_{i,j} = 1/N$ and $R = 1$, that is all possible interactions are represented. Conversely, if only one interaction accounts in the distribution budget, then $R = 1/N$, meaning that only one of all possible interactions is represented.

Distance between interactions

The previously defined interaction weights were averaged over all possible pairs of nucleotides at a given distance d of one another, yielding the distance distribution:

$$P(|i-j| = d) = \mathcal{Z}^{-1} \frac{1}{N-d} \sum_{|i-j|=d} w_{i,j}, \tag{30}$$

where

$$\mathcal{Z} = \sum_{d=1}^{N-1} \frac{1}{N-d} \sum_{|i-j|=d} w_{i,j} \tag{31}$$

is a normalization factor. Note that we introduced a correction $1/(N-d)$ to account for finite-size effects, namely the fact that randomly distributed interactions will lead to an overrepresentation of nearest neighbours interactions just because these are more numerous.

Interaction matrix and Hopfield patterns

In the Hamiltonian shown in (1), only $16L(L-1)/2$ terms appear in the interaction budget: indeed, we forbid self-interations (already accounted for by the local field h) and do not count the interactions twice. However, we can straightforwardly extend the interaction matrix to a full symmetric matrix $\hat{J}_{(i,a),(j,b)}$ of size $(4L)^2$, with $4L$-valued indices $(i,a), i \in \{1, \cdots, L\}, a \in \mathcal{A}$. The matrix $\hat{J}$ is such that for $i > j$, $\hat{J}_{(i,a)(j,b)} = J_{i,j}(a,b)$ with furthermore $\hat{J}_{(i,a),(i,b)} = 0$ and $\hat{J}_{(i,a),(j,b)} = \hat{J}_{(j,b),(i,a)}$. The energy of a sequence s can then be written with these notations

$$\sum_{1 \leq i < j \leq L} J_{i,j}(s_i, s_j) = \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \hat{J}_{(i,s_i),(j,s_j)} = v(s)^\dagger \hat{J} v(s), \tag{32}$$

where in the last equality the † sign denotes vector transposition and we have introduced the $4L$ vector $v(s)$ associated to sequence s, $v(s)_{i,a} = 1$ if $a = s_i$ and $v(s)_{i,a} = 0$ otherwise.

Since the matrix $\hat{J}$ is symmetric, it can be diagonalized in an orthonormal basis of eigenvectors ξ^k, $k = 1, \cdots, L$ with real eigenvalues λ_k,

$$\hat{J} = \sum_k \lambda_k \xi^k \, \xi^{k\dagger}. \tag{33}$$

Denoting by $\xi^k_{(i,a)}$ the coordinates of the k-th eigenvector then, one can rewrite Eq. (32) as

$$\sum_{1 \leq i < j \leq L} J_{i,j}(s_i, s_j) = \frac{1}{2} \sum_{k=1}^{4L} \lambda_k \left(\sum_{i=1}^{L} \xi^k_{(i,s_i)} \right)^2. \tag{34}$$

Finally, the full Hamiltonian is given by:

$$\mathcal{H} = -\sum_i h_i(s_i) - \frac{1}{2} \sum_{k=1}^{4L} \lambda_k \left(\sum_{i=1}^{L} \xi^k_{(i,s_i)} \right)^2. \tag{35}$$

Acknowledgments

We wish to thank PY Bourguignon and I Grosse for stimulating discussions at a preliminary stage of this work.

[1] F. Spitz and E. E. Furlong, Nat. Rev. Genet. **13**, 613 (2012).
[2] J. A. Stamatoyannopoulos, Genome Res. **22**, 1602 (2012).
[3] W. W. Wasserman and A. Sandelin, Nat. Rev. Genet. **5**, 276 (2004).
[4] O. G. Berg and P. H. von Hippel, J Mol Biol **193**, 723 (1987).
[5] G. D. Stormo and D. S. Fields, Trends Biochem Sci **23**, 109 (1998).
[6] T. K. Man and G. D. Stormo, Nucleic Acids Res **29**, 2471 (2001).
[7] P. V. Benos, M. L. Bulyk, and G. D. Stormo, Nucleic Acids Res **30**, 4442 (2002).
[8] M. L. Bulyk, P. L. F. Johnson, and G. M. Church, Nucleic Acids Res **30**, 1255 (2002).
[9] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, et al., Cell **152**, 327 (2013).
[10] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, et al., Science **324**, 1720 (2009), URL `http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19443739&retmode=ref&cmd=prlinks`.
[11] Y. Zhao and G. D. Stormo, Nat. Biotechnol. **29**, 480 (2011).
[12] Q. Zhou and J. S. Liu, Bioinformatics **20**, 909 (2004).
[13] M. Hu, J. Yu, J. M. G. Taylor, A. M. Chinnaiyan, and Z. S. Qin, Nucleic Acids Res **38**, 2154 (2010).
[14] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan, in *Proceedings of the seventh annual international conference on Research in computational molecular biology* (ACM, 2003), pp. 28–37.
[15] E. Sharon, S. Lubliner, and E. Segal, PLoS Comput Biol **4**, e1000154 (2008).
[16] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006), URL `http://www.nature.com/nature/journal/v440/n7087/full/nature04701.html`.
[17] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky, J Neurosci **26**, 8254 (2006), URL `http://www.jneurosci.org/cgi/content/full/26/32/8254`.
[18] Y. Ikegaya, G. Aaron, R. Cossart, D. Aronov, I. Lampl, D. Ferster, and R. Yuste, Science **304**, 559 (2004).
[19] A. Roxin, V. Hakim, and N. Brunel, J. Neurosci. **28**, 10734 (2008).
[20] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc Natl Acad Sci USA **106**, 67 (2009), URL `http://www.pnas.org/content/106/1/67.long`.
[21] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proc Natl Acad Sci USA **107**, 5405 (2010), URL `http://www.pnas.org/content/107/12/5405`.
[22] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Proc Natl Acad Sci USA **109**, 4786 (2012).
[23] Y. Zhao, S. Ruan, M. Pandey, and G. D. Stormo, Genetics **191**, 781 (2012).
[24] Y. Cao, Z. Yao, D. Sarkar, M. Lawrence, G. J. Sanchez, M. H. Parker, K. L. MacQuarrie, J. Davison, M. T. Morgan, W. L. Ruzzo, et al., Dev Cell **18**, 662 (2010).
[25] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, Mol Cell **38**, 576 (2010).
[26] R. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. Furlong, Nature **462**, 65 (2009).
[27] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, et al., Cell **133**, 1106 (2008).
[28] I. Dunham and et al., Nature **489**, 57 (2012).
[29] T. Cover and J. Thomas, *Elements of information theory* (Wiley-interscience, 2006).
[30] R. Baxter, *Exactly solved models in statistical mechanics* (Dover Publications, 2008).
[31] C. Bishop et al., *Pattern recognition and machine learning* (Springer New York, 2006).
[32] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, Genome Res **14**, 1188 (2004).
[33] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc Natl Acad Sci USA **106**, 67 (2009).
[34] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[35] S. Cocco, R. Monasson, and V. Sessak, Phys Rev E Stat Nonlin Soft Matter Phys **83**, 051123 (2011).
[36] G. Tkacik, E. Schneidman, M. J. B. II, and W. Bialek, arXiv **q-bio.NC** (2006), 4 pages, 3 figures, q-bio/0611072v1, URL `http://arxiv.org/abs/q-bio/0611072v1`.
[37] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc Natl Acad Sci USA **108**, E1293 (2011).
[38] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, Proc Natl Acad Sci USA **109**, 10340 (2012).
[39] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, Cell **149**, 1607 (2012).
[40] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, PLoS ONE **6**, e28766 (2011).
[41] M. Lassig, BMC Bioinformatics **8 Suppl 6**, S7 (2007).
[42] A. Moses, D. Chiang, D. Pollard, V. Iyer, and M. Eisen, Genome biology **5**, R98 (2004).
[43] R. Siddharthan, E. Siggia, and E. van Nimwegen, PLoS Comput Biol **1**, e67 (2005).
[44] H. Rouault, K. Mazouni, L. Couturier, V. Hakim, and F. Schweisguth, Proc Natl Acad Sci U S A **107**, 14615 (2010).
[45] Z. Bao and S. R. Eddy, Genome Res **12**, 1269 (2002).
[46] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and

F. Schacherer, Nucleic Acids Res **28**, 316 (2000).
[47] E. Jaynes, Physical review **108**, 171 (1957).
[48] C. Shannon, Bell Syst Tech J **27**, 623 (1948).

Supporting Figures

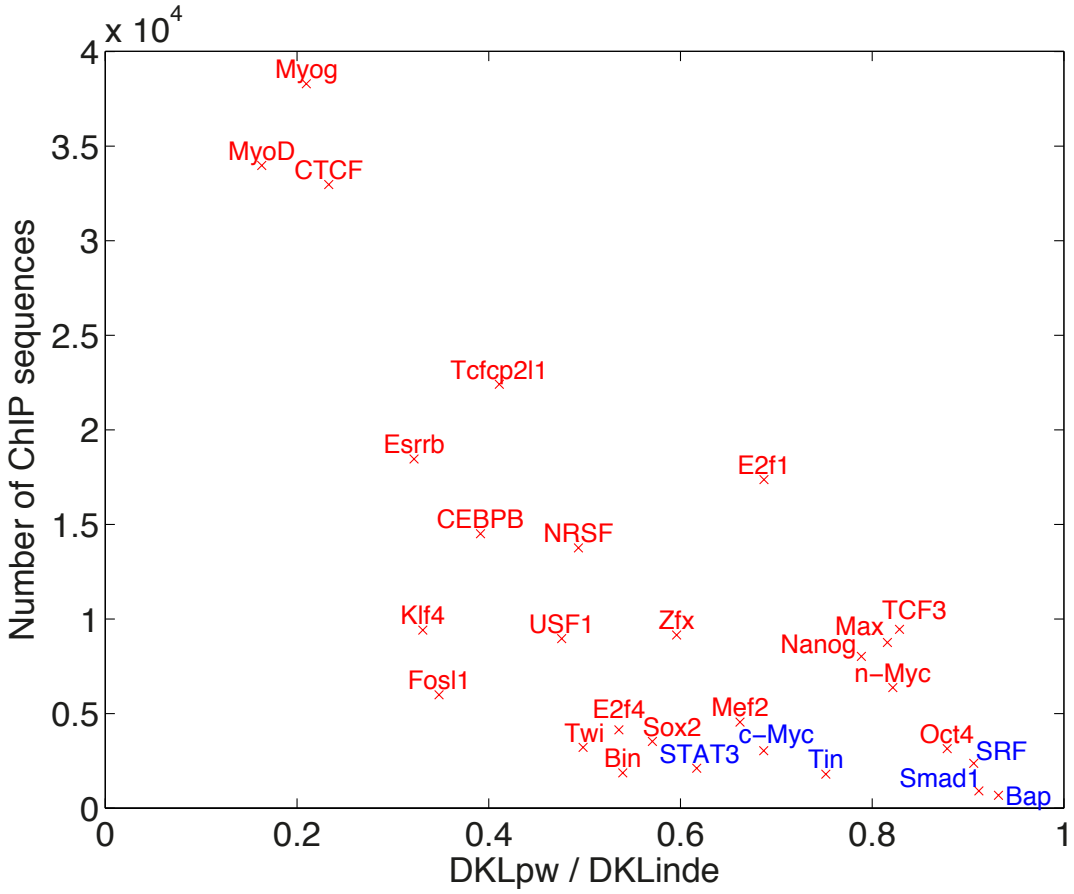
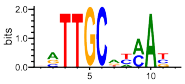


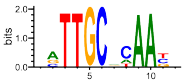
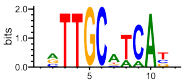
FIG. S1: **Dependence of the fit on the number of ChIP sequences**. For each TF, the number of available ChIP sequences is plotted *vs.* the improvement in the description of its TFBS statistics, provided by the he pairwise model as compared to the PWM independent model. The latter is quantified by the ratio of DKL between the respective model probability distributions and the experimental ones provided by the ChIP data, $\mathrm{DKL_{pw}}/\mathrm{DKL_{inde}}$. The improvement afforded by the pairwise model is clearly seen to be correlated to the number of ChIP sequences available.

CEBPB

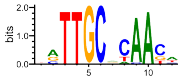
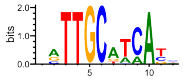
Independent model

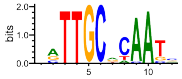
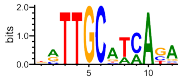


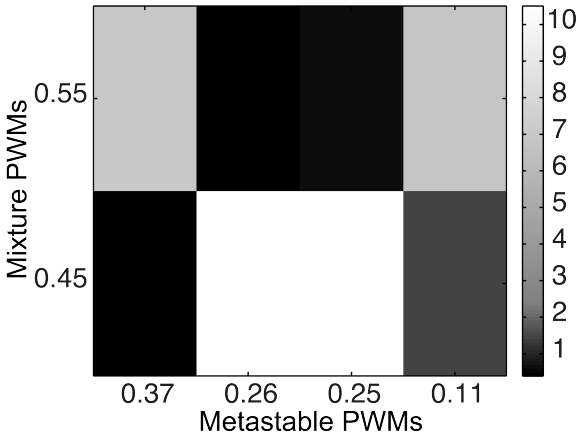
Mixture model

55%  45% 

Pairwise model

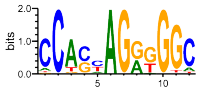
26%  37% 

25%  11% 

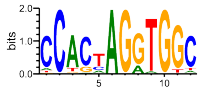
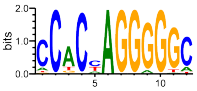
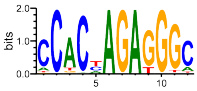
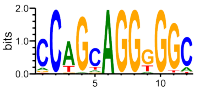
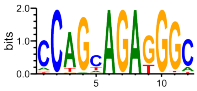


CTCF

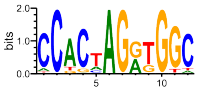
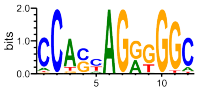
Independent model

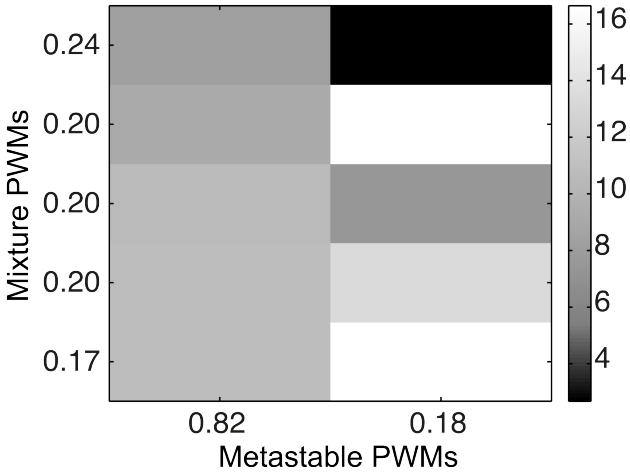


Mixture model

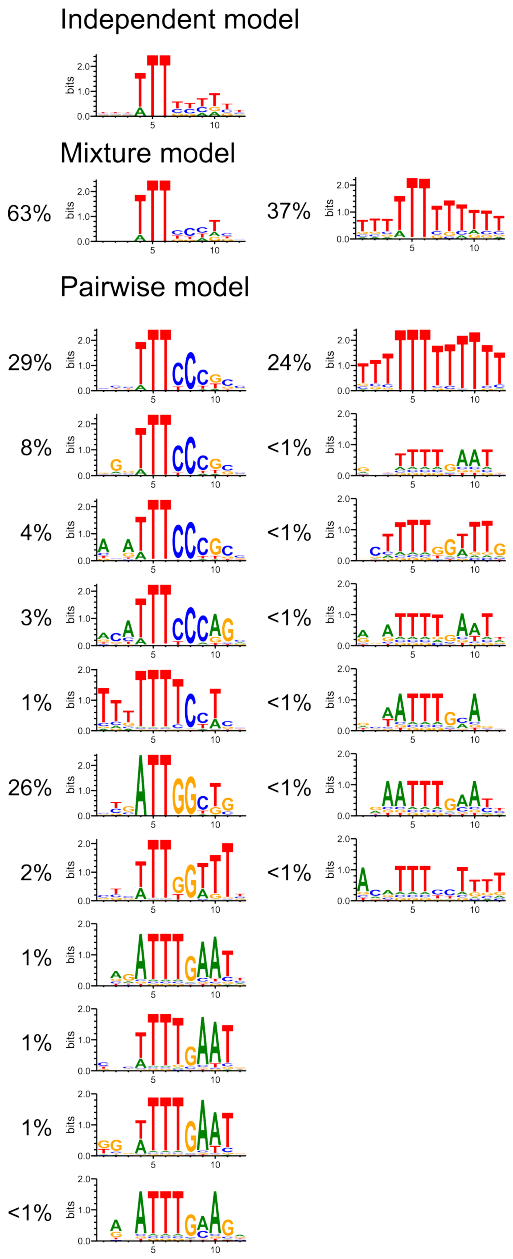
24%  20%  20%  20%  17% 

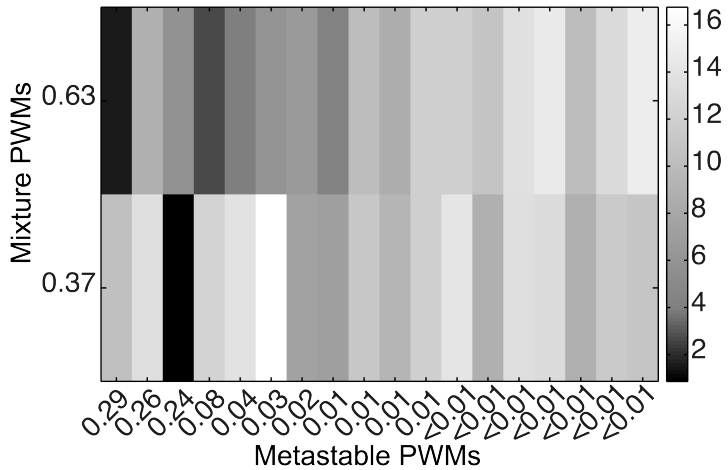
Pairwise model

18%  82% 

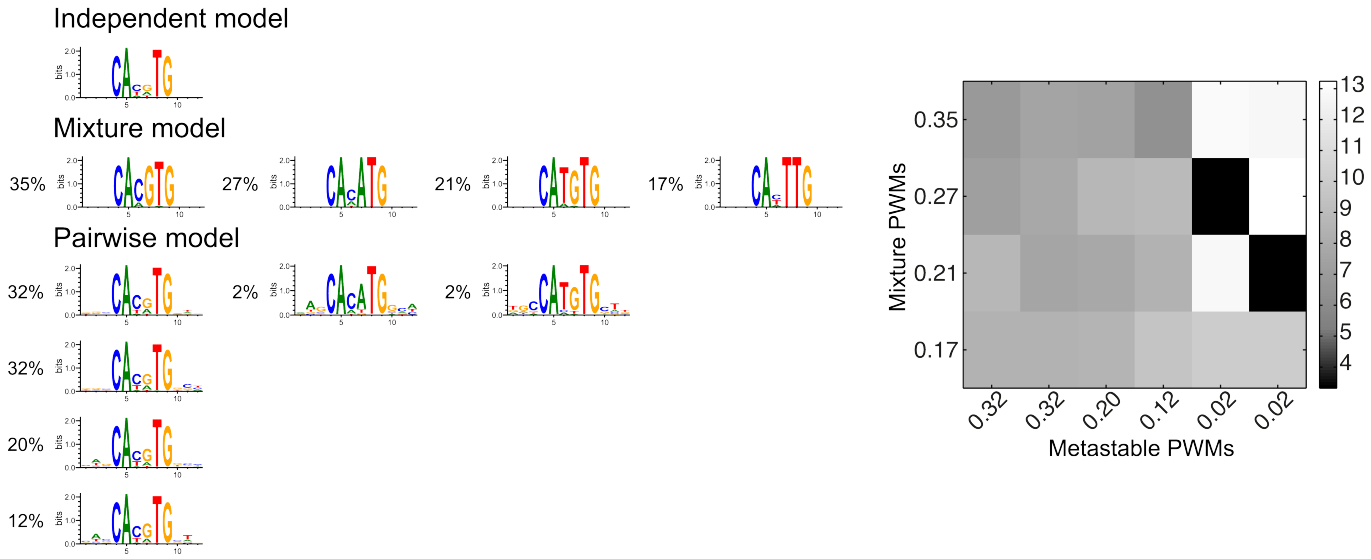


E2f4

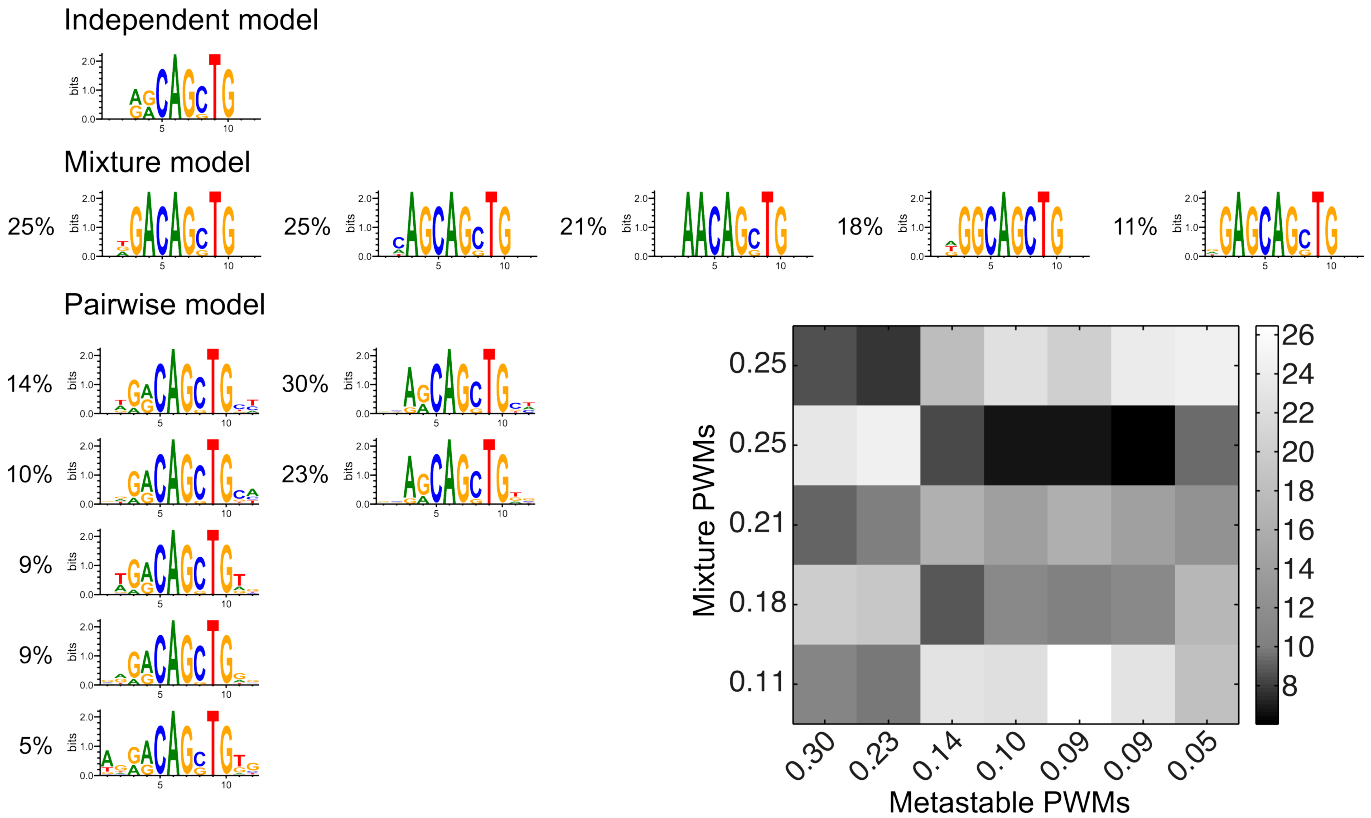

Independent model
Mixture model
63%
37%
Pairwise model
29%
24%
8%
<1%
4%
<1%
3%
<1%
1%
<1%
26%
<1%
2%
<1%
1%
1%
1%
<1%


Mixture PWMs
0.63
0.37
Metastable PWMs
0.29
0.26
0.24
0.08
0.04
0.03
0.02
0.01
0.01
0.01
0.01
<0.01
<0.01
<0.01
<0.01
<0.01
<0.01
16
14
12
10
8
6
4
2

Max

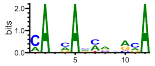

Independent model
Mixture model
35%
27%
21%
17%
Pairwise model
32%
2%
2%
32%
20%
12%
Mixture PWMs
0.35
0.27
0.21
0.17
0.32
0.32
0.20
0.12
0.02
0.02
Metastable PWMs
13
12
11
10
9
8
7
6
5
4

Myog

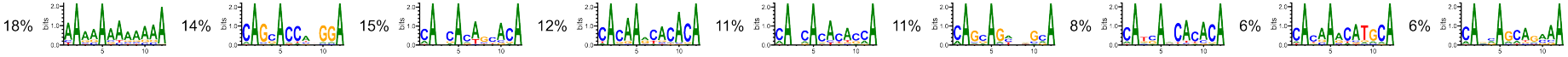

Independent model
Mixture model
25%
25%
21%
18%
11%
Pairwise model
14%
30%
10%
23%
9%
9%
5%
Mixture PWMs
0.25
0.25
0.21
0.18
0.11
0.30
0.23
0.14
0.10
0.09
0.09
0.05
Metastable PWMs
26
24
22
20
18
16
14
12
10
8

NRSF

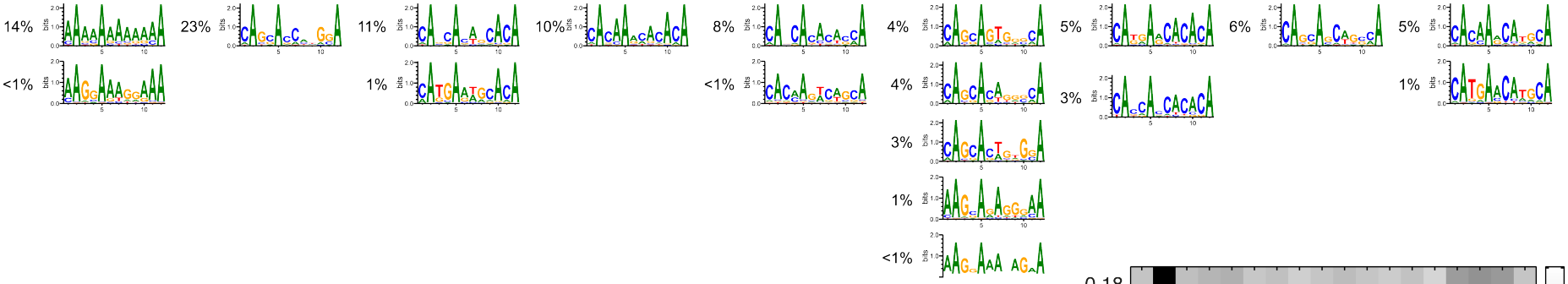
Independent model

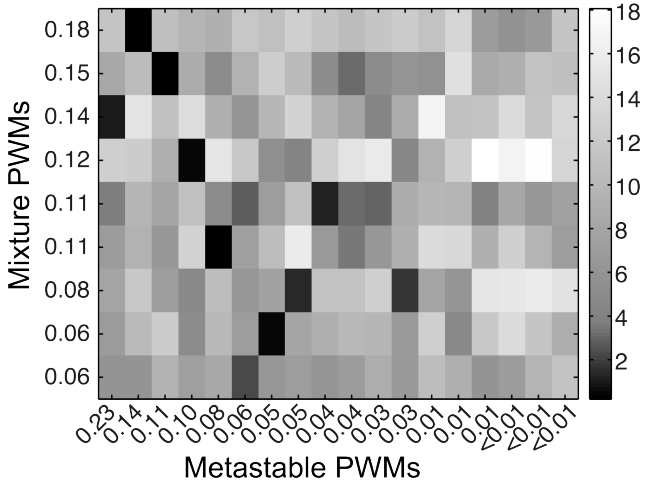


Mixture model

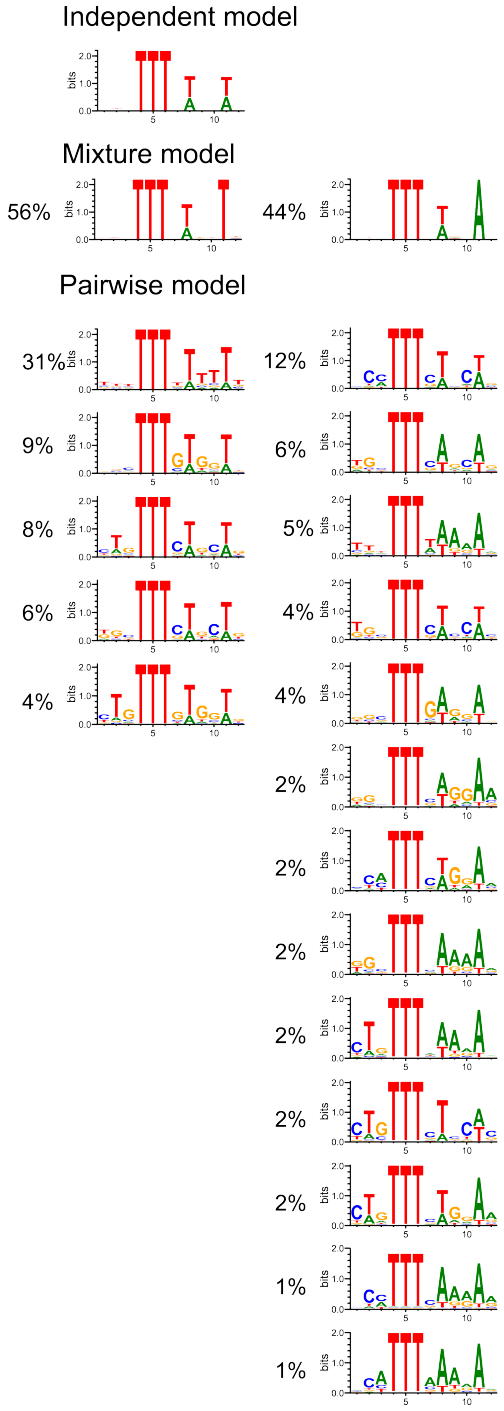


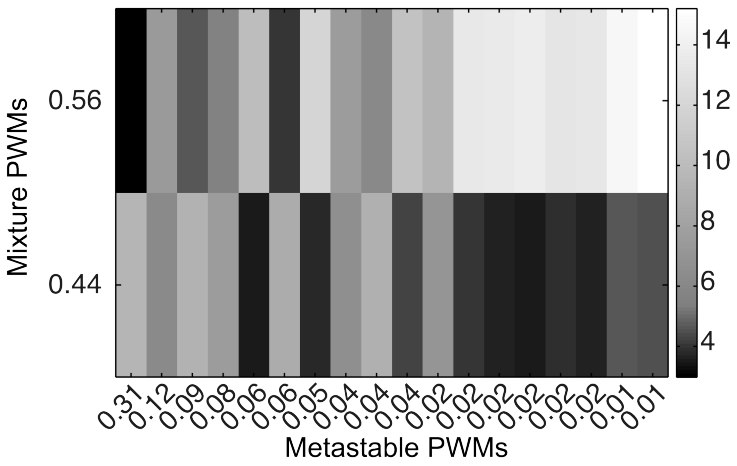
Pairwise model





Tcf3


Independent model
Mixture model
56%
44%
Pairwise model
31%
12%
9%
6%
8%
5%
6%
4%
4%
4%
2%
2%
2%
2%
2%
2%
1%
1%


Mixture PWMs
0.56
0.44
0.31
0.12
0.09
0.08
0.06
0.06
0.05
0.04
0.04
0.04
0.02
0.02
0.02
0.02
0.02
0.02
0.01
0.01
Metastable PWMs
14
12
10
8
6
4

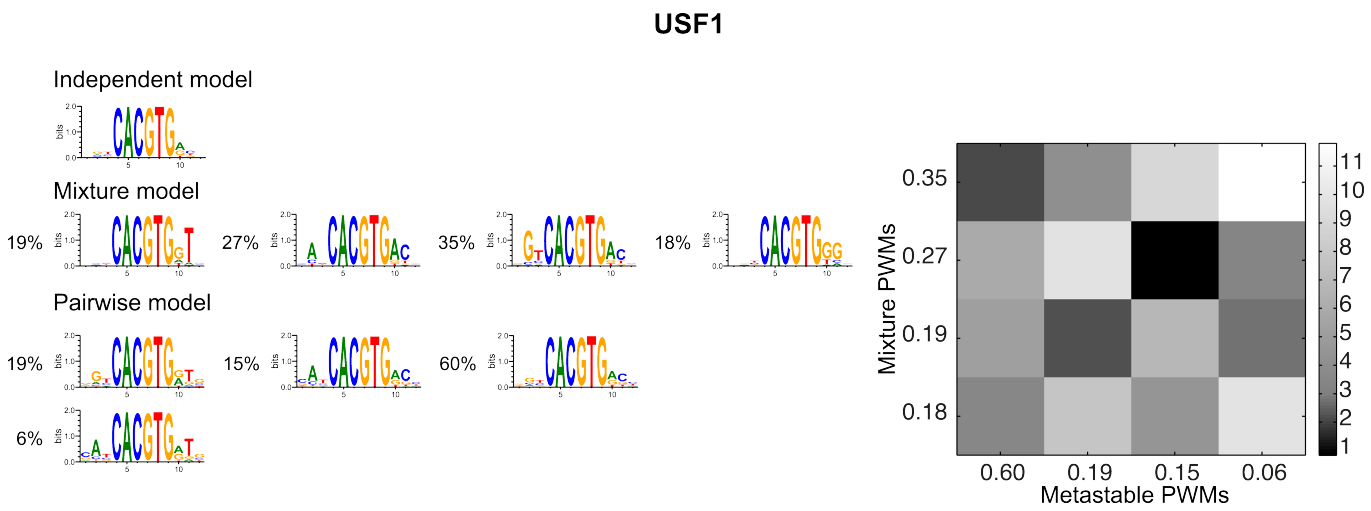
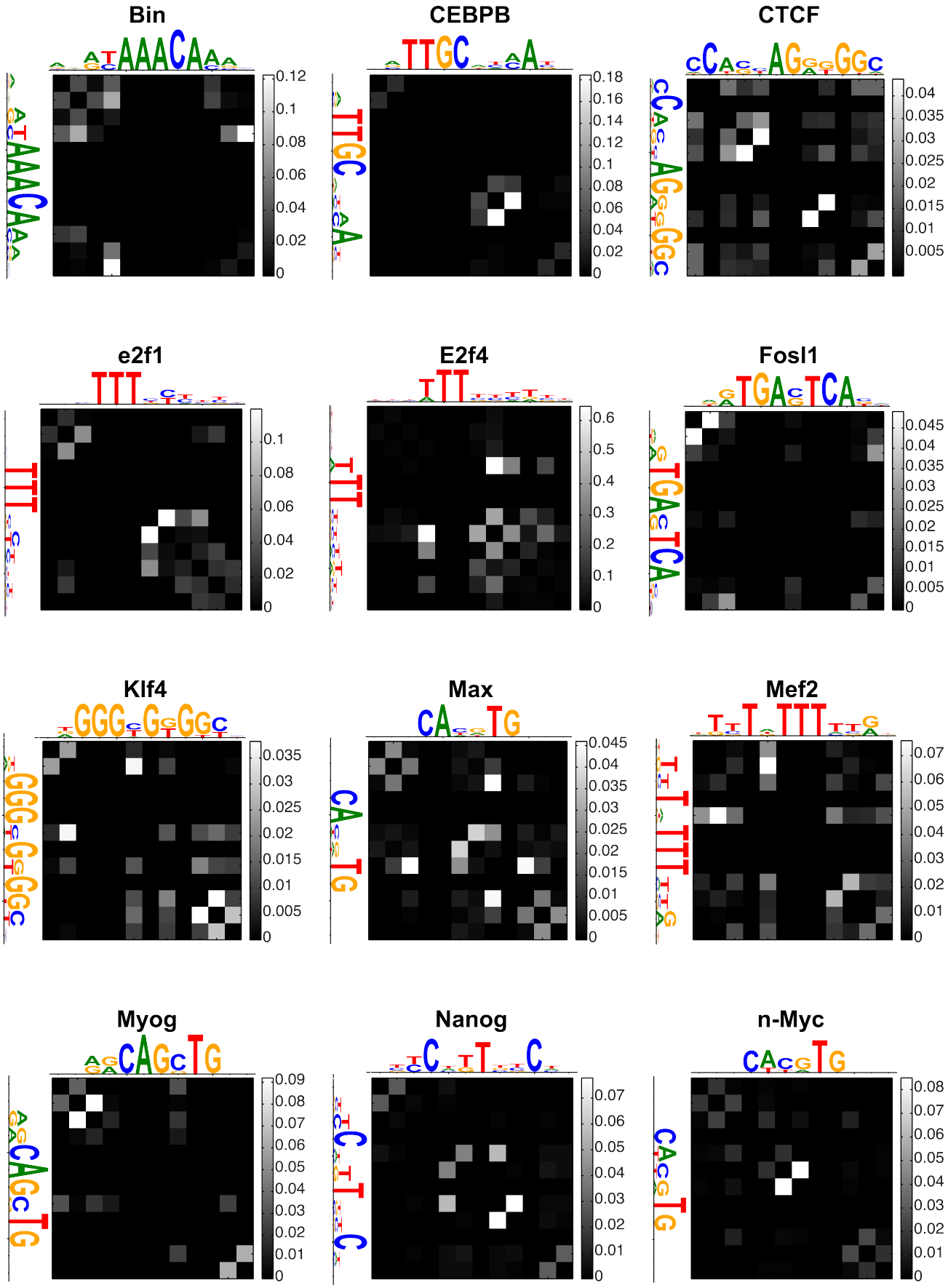


FIG. S2: Same as Figure 6 of the main text for all considered factors described by a mixture model with two or more PWMs.


Bin
0.12
0.1
0.08
0.06
0.04
0.02
0
CEBPB
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
CTCF
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
e2f1
0.1
0.08
0.06
0.04
0.02
0
E2f4
0.6
0.5
0.4
0.3
0.2
0.1
0
Fosl1
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Klf4
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Max
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Mef2
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Myog
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Nanog
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
n-Myc
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

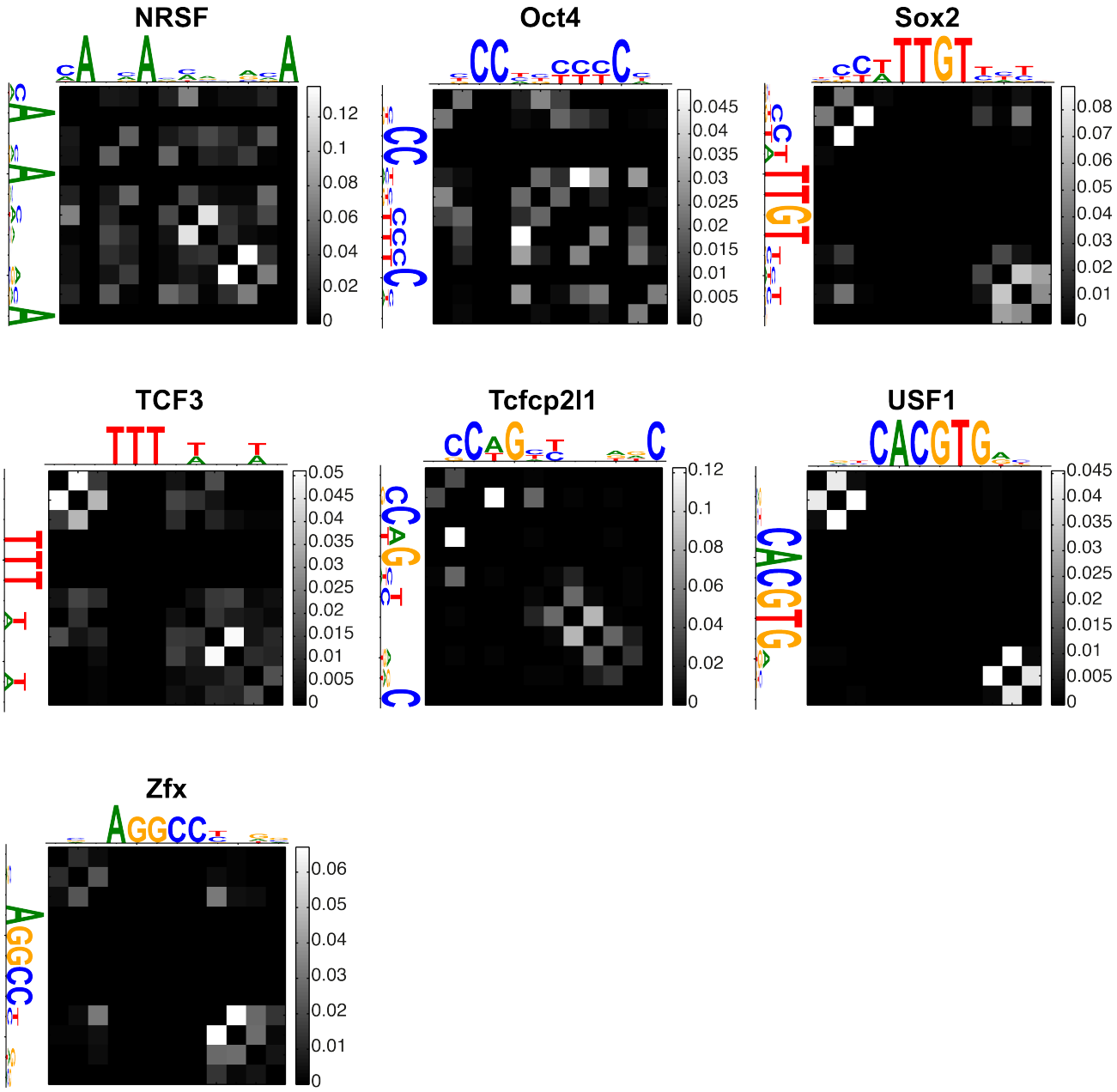


FIG. S3: Same as Figure 7 of the main text for the other considered factors.

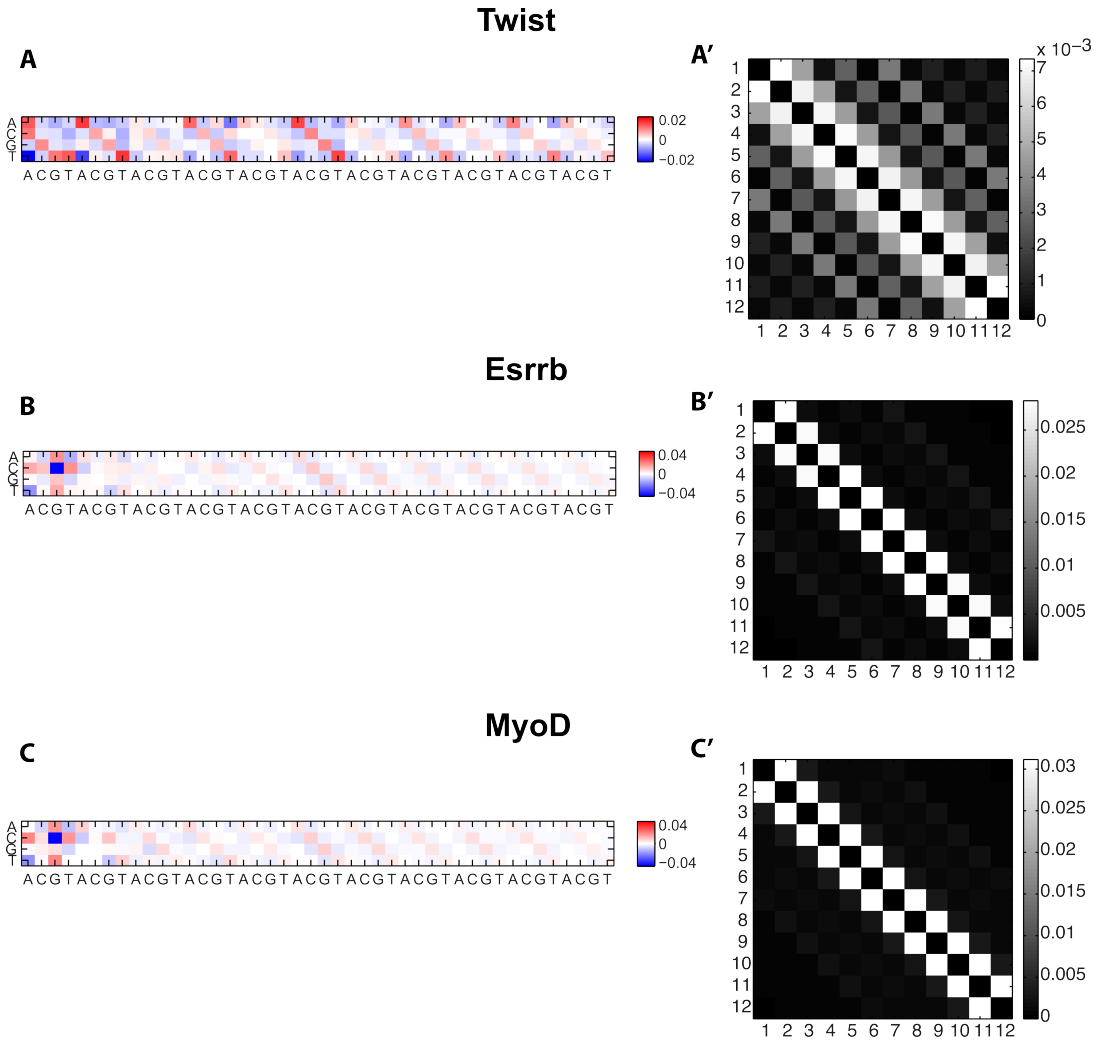


FIG. S4: **Background correlations** (A,B,C) Heat maps showing the correlations between nucleotides in the ChIP data of the 3 factors from the main text. Because of translation invariance, we only show the correlations between a nucleotide (rows) and the next nearest (first four columns) to farthest (last four columns) nucleotides, using the binding site length of $L = 12$. We see in the Drosophila data the appreciable presence of repeated sequences (of type AA, TT, CC, and GG). In the mammalian data sets, we observe the known CpG depletion. (A',B',C') Heat maps showing the values of the Normalized Direct Information between pairs of nucleotides.

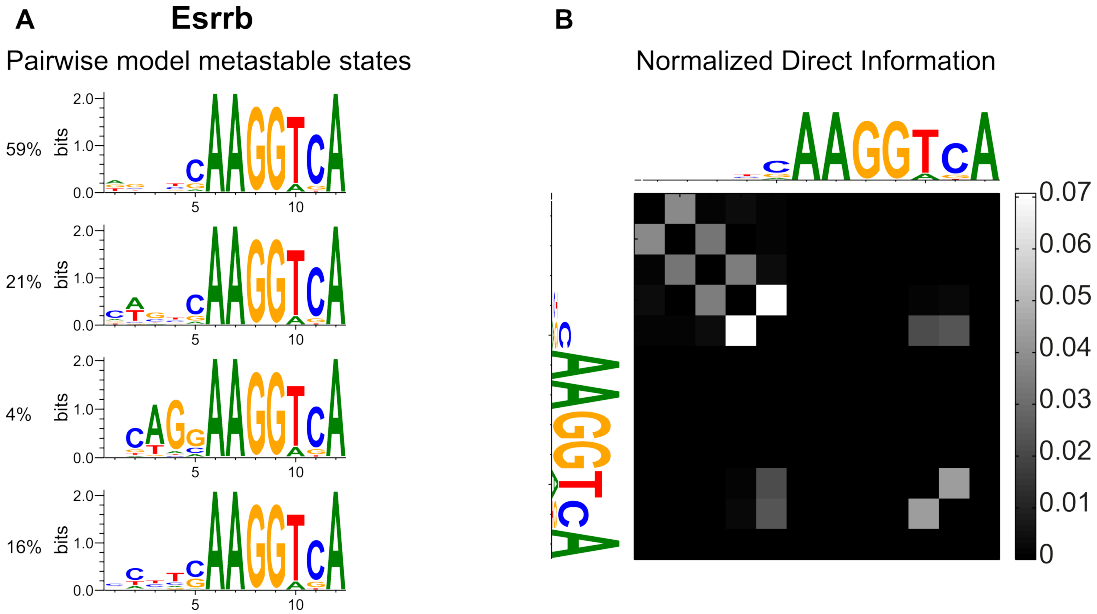


FIG. S5: **Variable spacer length** We learned a pairwise model for Esrrb including the 4 flanking nucleotides on the left of the main motif. (A) The metastable states of this model show a feature not captured in the main text where binding sites are defined symmetrically around the center of mass of the information content: namely a 'CAG' trinucleotide with variable spacer length from the main motif. This feature is apparent in the first 3 logos shown here. (B) The contribution of this trinucleotidic interaction to the Direct Information is captured through strong direct links between the 4 flanking nucleotides, showing that the pairwise model is implicitly able to capture higher order correlations. Logos from the PWM model are surrounding the heatmap for clarity.

2.5 Analyse thermodynamique des modèles

2.5.1 Chaleur spécifique

En plus des résultats présentés dans l'article, nous nous sommes intéressés à une quantité classique de la thermodynamique : la chaleur spécifique ou capacité calorifique. Considérons un modèle décrit par la statistique de Boltzmann à la température inverse $\beta = 1/T$ (on omet la constante de Boltzmann k en l'intégrant à l'énergie) :

$$P(s) = \frac{1}{\mathcal{Z}} e^{-\beta E(s)} \tag{2.24}$$

Le cas de l'équation 2.23 correspond au cas particulier $\beta = 1$. Nous voulons voir comment l'amplification ou la diminution globale de l'écart entre les énergies affecte la possibilité du système d'explorer les différents états possibles. À température nulle ($T \to 0$ ou $\beta \to \infty$), le système reste dans le niveau fondamental de minimum d'énergie et de probabilité 1, alors qu'à des températures non nulles le système à l'énergie E_0 transite vers un état d'énergie supérieure E_1 avec une probabilité $\propto \exp\left(-\beta(E_1 - E_0)\right)$. Lorsqu'un paysage énergétique est composé de plusieurs puits d'énergie séparés par des barrières énergétiques importantes, on s'attend à avoir une (ou plusieurs) températures critiques à partir desquelles de fortes différences d'énergie deviennent franchissables. L'énergie moyenne peut alors être significativement affectée, sautant soudainement à une nouvelle valeur du fait du poids des nouveaux états explorés.

La chaleur spécifique permet de caractériser ces sauts soudains d'énergie moyenne lors de la variation de la température, caractéristiques des transitions de phase. Elle mesure simplement la variation de l'énergie moyenne lors d'une variation de température :

$$C(T) = \frac{d\langle E \rangle}{dT} \tag{2.25}$$

où

$$\langle E \rangle = \sum_{\{s\}} E(s) \frac{e^{-\beta E(s)}}{\mathcal{Z}} \tag{2.26}$$

Cette chaleur spécifique peut par ailleurs s'écrire sous une forme plus pratique :

$$\begin{aligned} \frac{d\langle E \rangle}{dT} &= -\beta^2 \frac{d\langle E \rangle}{d\beta} \\ &= -\beta^2 \left[\sum_{\{s\}} E(s) \left(-E(s) e^{-\beta E(s)} \right) \frac{1}{\mathcal{Z}} + \sum_{\{s\}} E(s) e^{-\beta(E(s))} \left(-\frac{d\mathcal{Z}}{d\beta} \frac{1}{\mathcal{Z}^2} \right) \right] \\ &= \beta^2 \left[\langle E^2 \rangle - \langle E \rangle^2 \right] \end{aligned} \tag{2.27}$$

Ainsi, la chaleur spécifique $C(T)$ est directement accessible en regardant les corrélations de l'énergie sur l'ensemble des états du système, ce qui peut se calculer simplement à partir des modèles de fixation. Nous avons calculé la variation de $C(T)$ en fonction de la température pour les modèles indépendant et avec dépendances obtenus en 2.4 pour les différents TFs étudiés. Une température fictive est introduite dans les modèles en multipliant les énergies par β, afin de se placer dans le cadre de l'équation 2.24. Les résultats sont montrés en figure 2.3

(modèle indépendant en bleu, modèle de Potts en rouge). On observe pour la plupart des facteurs l'existence de deux pics de chaleur spécifique pour des températures de l'ordre de $T \sim 10^{-1}$ et $T \sim 5$ (par exemple, $T = 0.4$ et $T = 2.8$ dans le cas du modèle indépendant de Twist). Il y a de légères variations entre les deux modèles : notamment, le premier pic semble renforcé par le modèle de Potts dans plusieurs cas (par exemple, E2f4, NRSF, TCF3 ou Twist). Néanmoins, le nombre de pics (ou de transitions de phases) reste le même.

2.5.2 Lien avec les valeurs des champs et des couplages

Afin de comprendre l'existence des pics de chaleur spécifique et les énergies (températures) associées, il faut revenir aux modèles d'énergie. Lorsque l'on regarde l'histogramme des valeurs absolues de h_i obtenues dans les modèles indépendant des différents TFs étudiés, on trouve plusieurs valeurs typiques autour de 10^{-4}, 1 et 10 (fig. 2.4A). Celles-ci peuvent s'expliquer de la manière suivante. Dans le modèle indépendant, les champs sont simplement le logarithme naturel de la probabilité d'observer un nucléotide a à une position i donnée $h_i(a) = -\log P_i(a)$ (la jauge est choisie telle que $\mathcal{Z}_\rangle = 1$). En valeur absolue, les champs h_i proches de 0 ($h_i \sim 10^{-4} - 10^{-3}$) correspondent aux nucléotides très conservés (toujours observés), les valeurs autour de 1 correspondent à des nucléotides dégénérés (i.e également observés : $|\log(1/4)| \sim 1.4$) et les valeurs autour de 10 correspondent aux nucléotides qui ne sont jamais observés, au pseudocount près (pour un pseudocount de 1 et 10^4 séquences, $|\log(10^{-4})| \sim 9.2$). On peut maintenant mieux comprendre les pics de chaleur spécifique. À température nulle, seuls les sites consensus sont accessibles. Lorsque la température se rapproche de 1, les nucléotides dégénérés d'énergie $h_i \sim 1$ deviennent accessibles, augmentant significativement la valeur de l'énergie moyenne (premier pic). Puis, lorsque la température se rapproche de 10, les nucléotides non observés d'énergie $h_i \sim 10$ deviennent à leur tour accessibles, augmentant à nouveau l'énergie moyenne (deuxième pic).

Dans le cas du modèle de Potts (fig. 2.4B), les champs h_i prennent des valeurs proches de celles obtenues avec le modèle indépendant. Par ailleurs, les interactions $J_{i,j}$ sont réparties autour d'un mode centré autour de $J_{i,j} \sim 0.5$, ce qui correspond l'échelle d'énergie du premier pic. Ainsi, le renforcement du premier pic de chaleur spécifique par rapport au cas indépendant observé pour plusieurs TFs de la figure 2.3 peut s'expliquer par l'effet des termes d'interaction $J_{i,j}$.

2.6 Conclusion et perspectives du chapitre 2

Nous avons analysé les dépendances au sein des sites de fixation liés *in vivo* pour différents facteurs de transcription Drosophiles et mammifères. Nous avons comparé les performances d'un modèle PWM, d'un modèle de mélange de PWMs, et d'un modèle de Potts, en utilisant un critère bayésien (BIC) pénalisant les modèles à grand nombre de paramètres. Nous avons exhibé l'existence de corrélations faibles dont la prise en compte permet de significativement améliorer la description des données, le modèle de Potts étant significativement supérieur aux deux autres modèles dans la plupart des cas (22/28). Les interactions ont été étudiées systématiquement, montrant notamment une prépondérance des interactions entre plus proches voisins. Nous avons établi une correspondance entre les PWMs du modèle de mélange et les PWMs décrivant les états métastables du paysage énergétique généré par le modèle de Potts. Enfin, nous avons montré que les corrélations pouvaient être groupées en patterns de Hopfield ou « mémoires », et qu'un petit nombre était suffisant à reconstruire le paysage d'interactions.

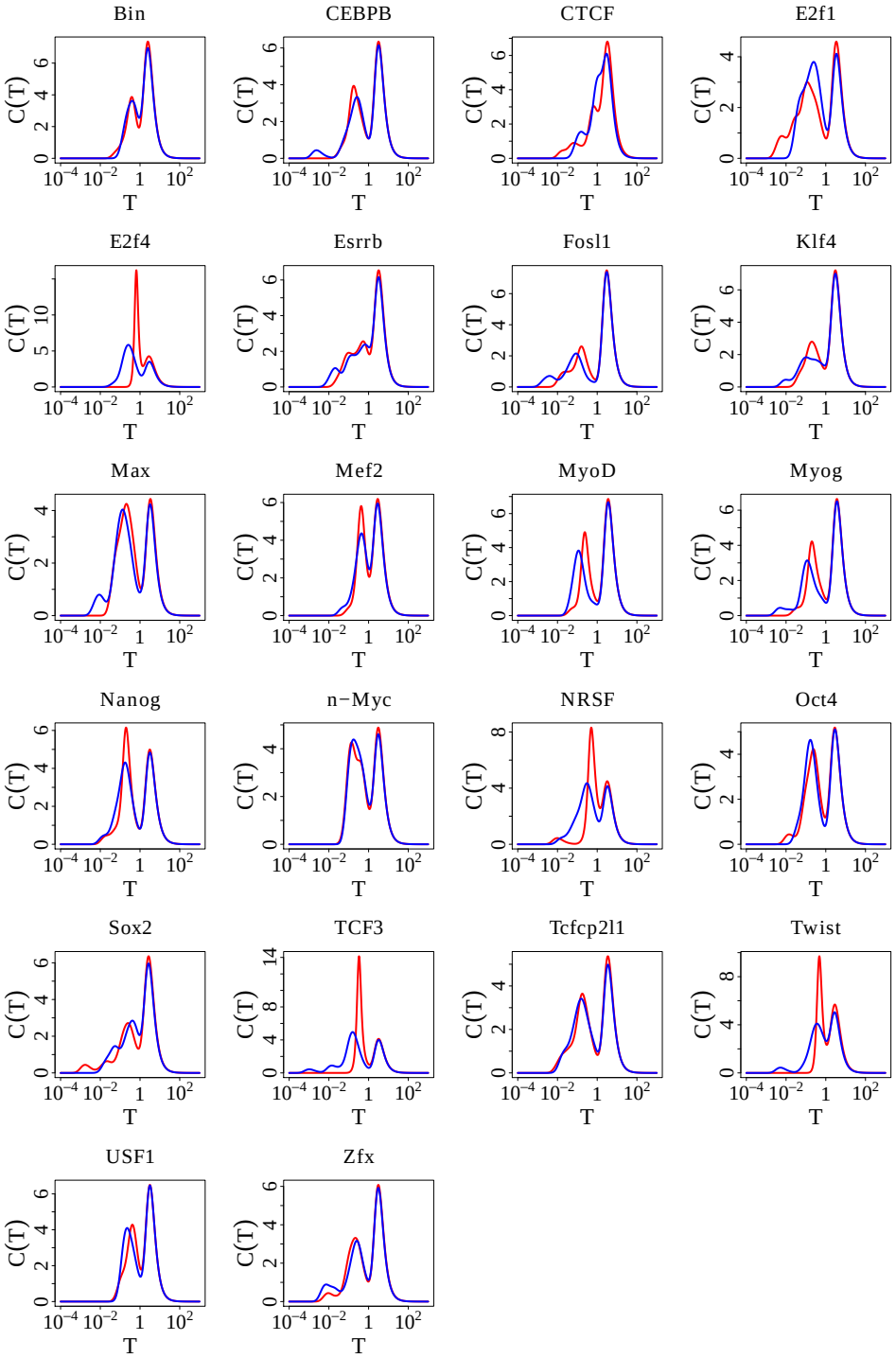


FIGURE 2.3 – **Chaleur spécifique pour différents TFs.**
La chaleur spécifique (l'équivalent de la capacité calorifique en thermodynamique) $C(T) = d\langle E\rangle/dT$ est tracée en fonction de la température kT (échelle logarithmique) pour les différents TFs considérés. Le modèle indépendant (bleu) et le modèle de Potts avec interactions (rouge) sont comparables dans la plupart des cas.

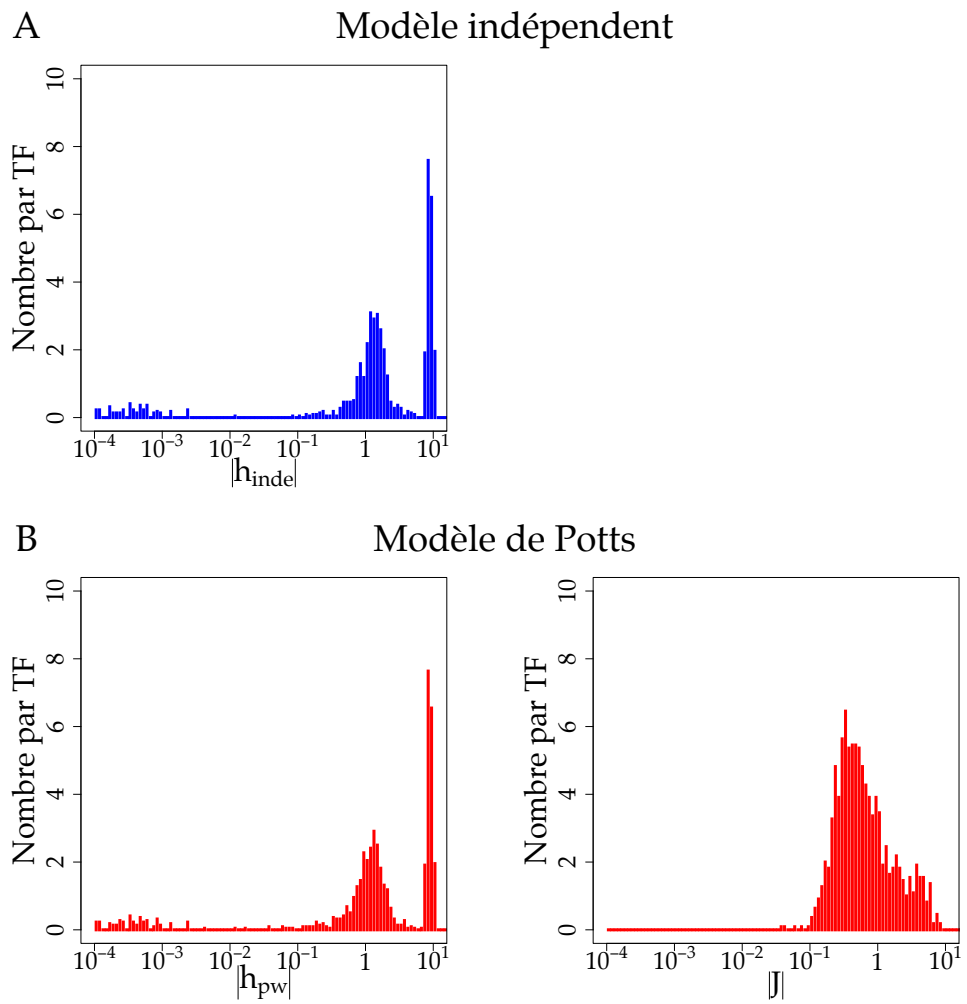


FIGURE 2.4 – **Histogrammes des valeurs des champs h et couplages J.**
Histogrammes réalisés à partir des valeurs obtenues pour l'ensemble des TFs. Les champs et les couplages sont montrés en valeur absolue sur une échelle logarithmique d'espacement 0.05, et les valeurs nulles ne sont pas représentées. (A) Champs h_{inde} dans le modèle indépendant. (B) Champs h_{pw} et couplages J dans le modèle de Potts.

Une perspective intéressante de ce travail serait de conduire la même analyse sur des données grande échelle obtenues *in vitro* par la méthode HT-SELEX (Jolma et al., 2013). Notamment, certains des facteurs que nous avons étudiés *in vivo* sont représentés dans ces données, et il serait intéressant de voir les différences entre les modèles obtenus. Notamment, retrouve-t-on les mêmes corrélations ? Peut-on exhiber des spécificités de la fixation *in vivo*, où l'on s'attend à avoir des effets provenant de diverses sources (fixation de nucléosomes, superposition de sites de fixations, . . .) ? Ces questions feront certainement l'objet d'un prochain travail.

Chapitre 3

Imogene : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle

3.1 Quelques approches existantes pour la recherche de motifs et de modules de régulation 96

3.1.1 MEME : une approche *de novo* par Espérance-Maximisation 96

3.1.2 STUBB : une méthode utilisant les corrélations entre sites de fixation et la phylogénie 99

3.1.3 MONKEY : vers des modèles phylogénétiques plus complexes 100

3.1.4 Approches sans motifs ou *motif-blind* 103

3.1.5 Autres méthodes utilisant des collections d'oligonucléotides 104

3.2 Article 105

3.3 Calcul de la moyenne de la postérieure par une méthode MCMC 132

3.3.1 Principe de l'algorithme de Metropolis-Hastings 132

3.3.2 Application au calcul de la postérieure 133

3.3.3 Illustration sur un exemple 134

3.4 Conclusion et perspectives du chapitre 3 139

Introduction du chapitre 3

Dans le chapitre 2, nous avons vu comment décrire l'interaction TF-ADN lorsque des sites de fixation sont connus. Dans ce chapitre, nous adoptons une démarche plus générale. Nous connaissons l'activité de régulation d'un certain nombre de CRMs, et nous souhaitons savoir quels TFs s'y fixent (recherche de motifs), et si le génome contient d'autres CRMs avec la même activité (recherche de modules). Un algorithme permettant précisément de réaliser ces étapes a été développé précédemment par Hervé Rouault et appliqué au cas de la différenciation des organes sensoriels de la Drosophile (Rouault et al., 2010). Cet algorithme se distingue des précédents par le fait qu'il n'utilise pas de motifs connus en entrée mais les génère purement *de novo*, et par son utilisation systématique de l'information provenant de la conservation chez d'autres espèces grâce à des modèles d'évolution, le rendant notamment adapté au cas où les CRMs connus sont en petit nombre. Nous présentons ici Imogene, l'extension de cet algorithme au cas des mammifères, ainsi que son utilisation comme outil de classification de CRMs associés à différentes régulations.

Avant de rentrer dans le détail d'Imogene, nous présentons les méthodes existantes de recherche de motifs dans des CRMs. Le problème général est le suivant : étant données des CRMs conduisant à une même régulation (l'ensemble d'apprentissage), peut-on construire des modèles de sites de fixation qui « expliquent » cette co-régulation, c'est-à-dire qui prédisent l'existence de sites sur les CRMs mais pas sur des séquences ne participant pas à la co-régulation ?

3.1 Quelques approches existantes pour la recherche de motifs et de modules de régulation

Nous avons déjà introduit différentes méthodes de prédiction de motifs et modules en introduction (section 1.6). Ici nous décrivons plus en avant certaines de ces méthodes que nous jugeons utiles à la mise en perspective d'Imogene, soit par leur approche de génération *de novo* de motifs, soit par leur utilisation de la conservation chez d'autres espèces et de modèles d'évolution pour la prendre en compte, soit par le fait qu'elles développent des statistiques appropriées à l'étude de petits échantillons de CRMs. Pour une revue plus exhaustive, le lecteur intéressé pourra se référer à Wasserman and Sandelin (2004) et Aerts (2012).

3.1.1 MEME : une approche *de novo* par Espérance-Maximisation

L'une des premières approches pour la prédiction *de novo* de motifs à partir de séquences a été celle de MEME (Bailey and Elkan, 1994), un algorithme basé sur la méthode d'Espérance-Maximisation ou EM (*Expectation Maximization*) utilisée précédemment dans ce cadre par Lawrence and Reilly (1990). Cet algorithme utilise une approche générative pour décrire les processus probabilistes qui ont permis la génération des séquences CRMs, ce qui permet d'écrire la probabilité qu'une séquence soit générée par un motif, et inversement de trouver le meilleur motif décrivant des séquences données. L'approche est illustrée en figure 3.1.

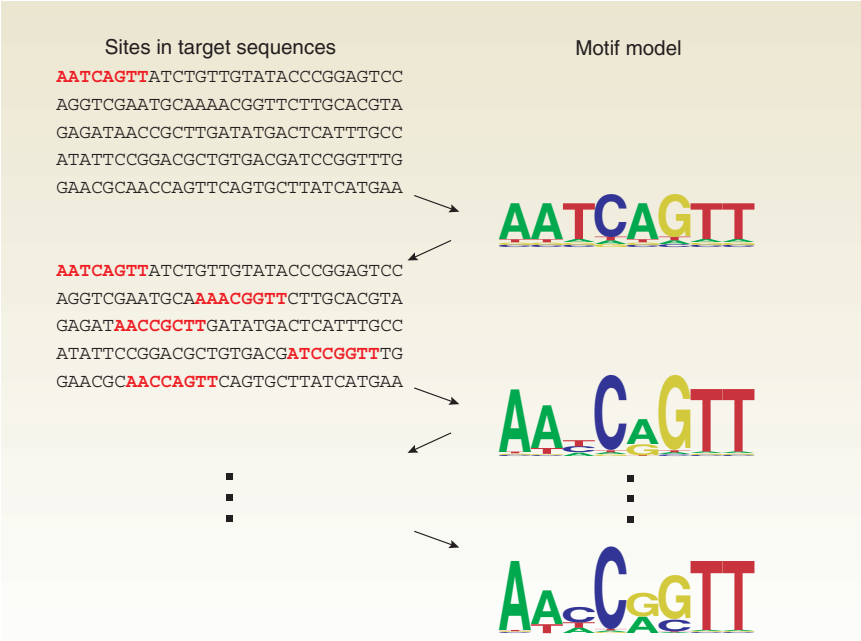


FIGURE 3.1 – **Illustration de l'approche Espérance-Maximisation.**
Figure tirée de D'haeseleer (2006) décrivant l'approche EM. Un premier modèle de motif est construit à partir d'un site initial. Ce modèle permet de pondérer l'ensemble des sites sur les séquences (étape E). En rouge sont montrés les meilleurs sites pour chaque séquence. En utilisant les poids des sites, il est possible de construire un modèle de vraisemblance maximale (étape M). La méthode originale de Lawrence and Reilly (1990) fait l'hypothèse qu'il y a exactement un site de fixation par séquence, condition qui est relâchée par MEME.

- **Vraisemblance d'une séquence**

Notons $S = \{S_1, \cdots, S_L\}$ une séquence[15] de taille L. Supposons qu'il y a exactement un site de régulation par CRM. C'est l'approche de Lawrence and Reilly (1990), et cette condition est relâchée par MEME, qui autorise l'utilisateur à préciser un nombre moyen de sites par séquence. La probabilité que la séquence possède un site de taille K à la position i étant donné le modèle de motif $\mathcal{M}$ s'écrit

$$P(S|i, \mathcal{M}) = P_0(S_{1,i-1}) \times P(S_{i,i+K-1}|\mathcal{M}) \times P_0(S_{K,L}) \tag{3.1}$$

où $S_{i,j}$ dénote la séquence entre les positions i et j incluses, $P(S_{i,i+K-1}|\mathcal{M})$ est la probabilité de générer la séquence de taille K débutant à la position i avec le modèle $\mathcal{M}$ (voir section 2.2), et $P_0(s)$ est la probabilité dite *background* de générer la séquence s étant donné un modèle génératif neutre, généralement pris comme étant une chaîne de Markov $\mathcal{P}_k$ d'ordre k petit (0 à 2) :

$$P_0(S_{i,j}) = \prod_{l=i}^{j} \mathcal{P}_k(S_l|S_{l-k,l-1}) \tag{3.2}$$

15. On concatène les deux brins d'ADN dans cette séquence. On suppose en effet qu'ils participent équiprobablement à la fixation. La séquence génomique double brins est donc de longueur $L/2$.

L'équation 3.1 décrit donc la probabilité de générer la séquence S avec le modèle *background*, sauf à la position i où un site est généré avec le modèle de fixation $\mathcal{M}$. La probabilité de générer la séquence s'obtient finalement en sommant sur les positions pondérées par la probabilité *a priori* $P(i)$ que le site soit à la position i :

$$P(S|\mathcal{M}) = \sum_{i=1}^{L-K+1} P(i)P(S|i,\mathcal{M}) \tag{3.3}$$

Cette probabilité est généralement prise uniforme, mais on peut y incorporer certaines informations, comme le nombre de séquences alignées (*reads*) d'une expérience de ChIP-seq.

- **Apprentissage du modèle**

Maintenant que nous savons exprimer la vraisemblance d'une séquence régulée par le motif $\mathcal{M}$, nous pouvons apprendre le meilleur modèle possible l'ayant générée : c'est la maximisation de la vraisemblance. Soit un ensemble de séquences $\mathcal{S}$ constitué de M séquences co-régulées $S[1], \cdots, S[M]$. Ces séquences étant supposées indépendantes, la vraisemblance que ces données soient générées par un modèle $\mathcal{M}$ est le produit sur les séquences de la quantité $P(S[m]|\mathcal{M})$. Il est plus utile dans ce cas de regarder la log-vraisemblance, s'écrivant alors comme une somme :

$$l(\mathcal{S}|\mathcal{M}) = \sum_{m=1}^{M} \log P(S[m]|\mathcal{M}) \tag{3.4}$$

Nous désirons obtenir le modèle $\mathcal{M}$ maximisant cette quantité [16]. Nous ne connaissons pas les positions exactes des sites, qui sont des « variables cachées » et il n'existe pas de méthode d'estimation simple permettant de résoudre ce problème. C'est à ce stade qu'intervient la méthode Espérance-Maximisation (EM) (Dempster et al., 1977). L'algorithme EM est une méthode itérative qui part d'un modèle initial $\mathcal{M}^0$ permettant de calculer les poids des positions dans les séquences (étape E d'espérance), puis estime le meilleur modèle $\mathcal{M}^1$ étant données ces poids (étape M de maximisation). L'itération a lieu jusqu'à convergence vers un maximum local.

Notons $\mathcal{M}^t$ le modèle à l'itération t. La probabilité qu'un site à la position i dans la séquence $S[i]$ soit un site de fixation s'écrit $P(i|S[m],\mathcal{M}^t)$. On définit la log-vraisemblance moyenne d'un modèle $\mathcal{M}$ à l'itération t par :

$$Q(\mathcal{M}|\mathcal{M}_t,\mathcal{S}) = \sum_m \sum_i P(i|S[m],\mathcal{M}^t) \log P(\mathcal{S},i|\mathcal{M}) \tag{3.5}$$

Le modèle suivant $\mathcal{M}^{t+1}$ est celui qui maximise cette quantité :

$$\mathcal{M}^{t+1} = \underset{\mathcal{M}}{\operatorname{argmax}} Q(\mathcal{M}|\mathcal{M}_t,\mathcal{S}) \tag{3.6}$$

L'équation 3.5 se scinde en une partie qui dépend de $\mathcal{M}$ et une partie *background* qui n'en dépend pas (eq. 3.1), que l'on peut dont ignorer pour ce qui est de la maximisation. Ainsi, le modèle $\mathcal{M}$ maximise la quantité suivante :

$$Q(\mathcal{M}|\mathcal{M}_t,\mathcal{S}) = \sum_m \sum_i P(i|S[m],\mathcal{M}^t) \log P(S_{i,i+K-1}|\mathcal{M}) \tag{3.7}$$

16. La distribution *background* étant fixée (par exemple la chaîne de Markov peut être apprise sur un grand nombre de séquences intergéniques non codantes).

Chaque K-mer $S_{i,i+K-1}$ des séquences de $\mathcal{S}$ est donc pris en compte dans l'apprentissage en proportion de la croyance courante $P(i|S[m], \mathcal{M}^t)$ que c'est un site de fixation.

Pour résumer, on a deux étapes :
- étape E : utiliser $\mathcal{M}^t$ pour attribuer un poids à chaque K-mer des séquences
- étape M : apprendre $\mathcal{M}^{t+1}$ qui a la plus grande vraisemblance de générer les données pondérées par $\mathcal{M}^t$.

Reste le problème de choisir un modèle initial adéquat pour être sûrs de converger vers un maximum de vraisemblance global et pas juste local. MEME adopte pour cela une approche semi-exhaustive. Les différents K-mers des séquences d'apprentissage sont successivement utilisés pour générer un modèle initial. L'algorithme EM est itéré une fois. Le modèle de plus grande (log-)vraisemblance est finalement gardé come motif initial pour une itération complète.

3.1.2 STUBB : une méthode utilisant les corrélations entre sites de fixation et la phylogénie

L'algorithme STUBB (Sinha et al., 2003) décrit les séquences par un modèle de Markov caché (HMM pour *Hidden Markov Model*) et les motifs par des PWMs. Il est basé sur l'algorithme Ahab (Rajewsky et al., 2002) – lui-même basé sur l'algorithme MobyDick (Bussemaker et al., 2000) –, qui peut être vu comme une extension de MEME au cas où les séquences contiennent plusieurs sites de fixations pour différents motifs. Notamment, ces méthodes ont l'intérêt tout comme MEME de ne pas avoir de seuil arbitraire pour définir un site car elles moyennent sur toutes les positions de sites (ou segmentations) possibles de la séquence. On parle de modèles thermodynamiques (voir 1.6.1). La différence entre STUBB et Ahab est qu'il introduit deux informations supplémentaires : les corrélations entre motifs et la phylogénie. Enfin, contrairement à MEME, l'algorithme utilise comme condition initiale un ensemble de motifs connus pour être impliqués dans la co-régulation étudiée.

- **Description du modèle HMM**

Nous décrivons d'abord l'algorithme Ahab (Rajewsky et al., 2002). Notons W l'ensemble des motifs initiaux. Le modèle HMM à l'ordre 0 (HMM0) utilisé décrit la génération d'une séquence S de la manière suivante. La séquence est initialement de taille nulle. Le processus choisit un motif $w_i \in W$ avec une probabilité p_i ou le motif *background* w_b (une PWM de longueur 1) avec une probabilité $1 - \sum_i p_i$. Une fois le motif w choisi, une séquence est échantillonnée à partir de la PWM de w et est ajoutée à la séquence S. Le processus est itéré jusqu'à ce que la séquence générée atteigne une taille L. La séquence de motifs choisis au cours de la procédure définit une segmentation T. La probabilité que la séquence observée soit générée par ce processus de paramètres $\theta = \{w_i, p_i\}$ est

$$P(S|\theta) = \sum_T P(T|\theta) P(S|T, \theta) \tag{3.8}$$

et peut être calculée par programmation dynamique (algorithme forward-backward). Le score d'une séquence est obtenu en comparant cette probabilité et la probabilité $P(S|\theta_b)$ que la séquence soit générée uniquement par le modèle *background* :

$$F(S) = \underset{\theta}{\operatorname{argmax}} \log \left(\frac{P(S|\theta)}{P(S, \theta_b)} \right) \tag{3.9}$$

Le paramètre θ (c'est-à-dire les p_i) qui maximise le membre de droite est obtenu grâce à un algorithme de type EM (Sinha et al., 2003).

- **Ajout des corrélations entre motifs**

Des informations sur les corrélations entre motifs sont introduites dans θ sous la forme de probabilités de transition p_{ij} que le motif choisi lors de la génération de la séquence soit w_j lorsque le premier motif précédent non-*background* est w_i. Parce que le nombre de paramètres devient grand, seules les corrélations importantes (dépassant un seuil fixé) sont ajoutées.

- **Incorporation de l'information phylogénétique**

Enfin, STUBB utilise l'information provenant de la conservation de la séquence chez d'autres espèces. Les séquences des différentes espèces sont d'abord alignées, puis la probabilité de générer l'alignement est calculée à l'aide d'un modèle phylogénétique. Ce modèle permet de prendre en compte le fait que les séquences homologues sont corrélées du fait qu'elles dérivent d'un ancêtre commun. Dans le cas de Stubb, les espèces sont supposées liées par un topologie en étoile, c'est-à-dire que les espèces partagent un seul ancêtre commun. Le modèle d'évolution suppose que les différentes bases de la séquence évoluent indépendamment, mutent à la même fréquence, et que la probabilité de fixation d'une mutation $b \to b'$ à la position i est proportionnelle au poids $w_{i,b'}$ de la PWM du nucléotide b' à cette position. Ce modèle est identique au modèle *Felsenstein* que nous introduisons dans l'article (voir section 3.2) et qui est inspiré du modèle neutre de Felsenstein (1981) – les probabilités neutres étant remplacées par les fréquences PWM –. La probabilité $P(\sigma|w)$ de générer l'alignement σ de séquences s avec le motif w de taille L_w s'écrit alors :

$$P(\sigma|w) = \prod_{i=1}^{L_w} \left[\sum_b w_{i,b} \prod_{s \in \sigma} (q_s \delta_{b,s_i} + (1 - q_s) w_{i,s_i}) \right] \tag{3.10}$$

où w_{i,s_i} est la probabilité de générer le nucléotide s_i à la position i pour le motif w, $\delta_{x,y} = 1$ si $x = y$ et 0 sinon, et $q_s = e^{-\lambda t_s}$ est la probabilité de conserver un nucléotide au cours de l'évolution, qui est une fonction du taux de mutation neutre λ et du temps d'évolution t_s entre l'ancêtre commun et l'espèce s. En résumé, pour chaque position i, un nucléotide b est généré chez l'ancêtre commun avec une probabilité $w_{i,b}$, puis ce nucléotide est soit conservé chez l'espèce s avec une probabilité q_s ou bien il mute avec une probabilité $1 - q_s$, et une nouvelle base est sélectionnée selon les poids définis par w_i. Pour des espèces proches, $q \sim 1$ et le fait d'observer des bases différentes à des positions homologues diminue fortement $P(\sigma|w)$, même si leur fréquence *a priori* donnée par w est identique : le modèle donne alors naturellement plus de poids aux séquences relativement conservées. Pour des espèces lointaines, $q \sim 0$ et tout se passe comme si les séquences de σ étaient indépendantes.

3.1.3 MONKEY : vers des modèles phylogénétiques plus complexes

Dans la section précédente, le modèle phylogénétique utilisé par Stubb est relativement simple et la probabilité de fixation n'a pas de base claire. L'algorithme MONKEY (Moses et al., 2004) propose d'utiliser des modèles d'évolution plus complexes pour détecter les sites conservés à partir de motifs connus.

Les motifs w sont décrits par le modèle PWM et le *background* par les fréquences des nucléotides π_b. Le score d'un site est défini par la fraction des probabilités de générer la séquence s par l'un ou l'autre des modèles (*log-likelihood ratio* ou LLR) :

$$LLR(s) = \log \frac{P(s|w)}{P(s|\pi)} = \sum_i \log \frac{w_{i,b(i)}}{\pi_b} \tag{3.11}$$

Le but est de généraliser ce score au cas d'un alignement σ de séquences s, en réalisant son calcul sur l'ancêtre commun des séquences. Cet ancêtre commun, ainsi que tous les ancêtres communs intermédiaires pour une topologie d'arbre T quelconque, ne sont pas observés, et il faut donc sommer sur tous leurs états (nucléotides) possibles étant donnés l'alignement observé σ, l'arbre T et le modèle d'évolution. La solution générale de ce problème a été donnée par Felsenstein (1981). Notamment, nous pouvons nous concentrer sur le cas à deux espèces, le cas général procédant par récurrence à partir de ce cas simple. Nous pouvons aussi nous concentrer sur une position i donnée, puisque les positions sont indépendantes.

Considérons un alignement de deux nucléotides s_1 et s_2. On définit un nouveau score comme étant le LLR comparant l'hypothèse que s_1 et s_2 représentent un site conservé pour un motif w et l'hypothèse que les bases ont été tirées dans le *background* :

$$LLR_{\text{cons}}(s_1, s_2) = \log \frac{P(s_1, s_2|w, T, R_w)}{P(s_1, s_2|\pi, T, R_{\text{back}})} \tag{3.12}$$

où R_w et R_{back} sont des matrices de taux de transition décrivant les processus de substitution au cours de l'évolution pour le cas d'un site de fixation du motif w et pour le *background*, et interviennent dans l'écriture de la probabilité de transition de la base b à la base b' :

$$p_{b \to b'} = \left(e^{R_M d}\right)_{b,b'} \tag{3.13}$$

où R_M est la matrice de taux de transition du modèle M et d le temps évolutif. Dans le cas simple à deux espèces, l'arbre T est en étoile, avec des distances d_1 et d_2 de l'ancêtre commun aux espèces contenant les bases s_1 et s_2. Par ailleurs, les espèces évoluent indépendamment depuis leur séparation. Notant b la base sur l'ancêtre commun, on a finalement :

$$\begin{aligned} P(s_1, s_2|M, T, R_M) &= \sum_b P(s_1|b, d_1, R_M) P(s_2|b, d_2, R_M) P(b|M) \\ &= \sum_b \left(e^{R_M d_1}\right)_{b,s_1} \left(e^{R_M d_2}\right)_{b,s_2} P(b|M) \end{aligned} \tag{3.14}$$

Le calcul sur un nombre quelconque d'espèces se fait récursivement (Felsenstein, 1981), jusqu'à la racine de l'arbre où $P(b|M)$ vaut $w_{i,b}$ pour le motif et π_b pour le *background*.

Plusieurs modèles peuvent être choisis pour les matrices de transition. Dans le cas du *background*, il peut être décrit par des modèles neutres. Par exemple le modèle de Felsenstein (Felsenstein, 1981) est le modèle le plus simple dont la distribution d'équilibre redonne les fréquences du *background*. Le modèle HKY (Hasegawa et al., 1985) est une variante qui inclut le fait que les mutations entre bases de même nature chimique (purine ou pyrimidine), appelées transitions, sont 2 fois plus fréquentes que les autres mutations, appelées transversions. Ce dernier modèle est utilisé dans MONKEY pour décrire le *background*. Pour ce qui est du motif, les taux de transition dépendent de la position au sein du site : par exemple les bases dégénérées mutent plus vite que les bases très conservées (Moses et al., 2003). Pour prendre cette variation en compte, une possibilité est de modifier le modèle neutre en utilisant à la place des fréquences *background* les fréquences données par la PWM : c'est ce qui est fait dans

Stubb avec le modèle Felsenstein (éq.3.10). Dans MONKEY, les auteurs utilisent un modèle plus complexe, appelé modèle Halpern-Bruno ou HB, préalablement introduit pour étudier l'évolution des régions codantes (Halpern and Bruno, 1998). Dans ce modèle, le taux de substitution $R(i)_{a,b}$ de la base a vers la base b en position i est à une constante près le produit de la probabilité de mutation neutre de a vers b (indépendante de la position) par une probabilité de fixation $f^i_{a,b}$ (dépendante de la position) :

$$R(i)_{a,b} \propto Q_{a,b} f^i_{a,b} \tag{3.15}$$

où $Q = R_{\text{back}}$ est la matrice de transition du modèle *background*. La probabilité de fixation peut être obtenue en utilisant un résultat classique de génétique des populations (Kimura, 1962) :

$$f^i_{a,b} \simeq \frac{2s}{1 - e^{-2Ns}} \tag{3.16}$$

$$f^i_{b,a} \simeq \frac{-2s}{1 - e^{2Ns}} \tag{3.17}$$

où N est la taille effective de la population (le facteur 2 vient du fait que la population est diploïde), s est la valeur adaptative relative ou *fitness* de la base b par rapport à la base a en position i, et l'évolution est quasi-neutre ($s \ll 1$). Cette dernière hypothèse permet d'écrire :

$$\frac{f^i_{a,b}}{f^i_{b,a}} \simeq e^{2Ns} \tag{3.18}$$

Par ailleurs, en supposant que les substitutions vérifient le bilan détaillé à l'équilibre, on peut écrire

$$\frac{f^i_{a,b}}{f^i_{b,a}} = \frac{w_{i,b} Q_{b,a}}{w_{i,a} Q_{a,b}} \tag{3.19}$$

Finalement, en combinant les équations 3.16, 3.18 et 3.19, on obtient la matrice de transition du modèle HB en position i sous la forme :

$$\boxed{R(i)_{a,b} \propto Q_{a,b} \frac{x \log x}{x - 1}} \tag{3.20}$$

où

$$x = \frac{w_{i,b} Q_{b,a}}{w_{i,a} Q_{a,b}} \simeq e^{2Ns} \tag{3.21}$$

traduit l'effet de la sélection. En développant autour de $x = 1$ et en restant au premier ordre, on a

$$\frac{x \log x}{x - 1} \simeq \frac{1}{2}(1 + x) \tag{3.22}$$

Ainsi, dans le cas neutre où $x = 1$, la matrice de transition se réduit à la matrice *background* : $R(i)_{a,b} = Q_{a,b}$. Cependant, lorsque la base b est plus conservée que la base a ($x > 1$), les substitutions de a vers b sont plus fréquentes que sous le modèle neutre : $R(i)_{a,b} > Q_{a,b}$.

3.1.4 Approches sans motifs ou *motif-blind*

Les algorithmes précédents utilisent en leur cœur un modèle de motif $\mathcal{M}$, généralement une PWM, permettant d'attribuer une probabilité $P(S|\mathcal{M})$ à une séquence donnée. Néanmoins, il existe certaines méthodes cherchant à décrire plus généralement la statistique des mots au sein des CRMs sans chercher à associer ces statistiques à des motifs ayant une caractérisation biochimique précise. De telles approches sont dites sans motifs (*motif-blind*). Nous en recensons ici quelques-unes (voir Kantorovitz et al. (2009) pour plus de détails).

- **Modèles basés sur des chaînes de Markov**

Plusieurs modèles basés sur des chaînes de Markov ont été proposés. Par exemple, l'algorithme PFRSampler de Grad et al. (2004) consiste en un apprentissage de modèles de Markov d'ordre 5 sur des séquences d'intérêt et sur des séquences *background*, ces séquences étant préalablement filtrées par la conservation phylogénétique. Il est ensuite possible de calculer la vraisemblance qu'une séquence donnée soit générée par l'un ou l'autre des modèles, de manière similaire à l'éq.3.3. Le score d'une séquence S de taille L est défini comme étant la différence des log-vraisemblances qu'elle soit générée par le modèle d'intérêt $\mathcal{M}_{\text{train}}$ et par le modèle *background* $\mathcal{M}_{\text{back}}$:

$$\text{Score}(S) = \log \frac{P(S|\mathcal{M}_{\text{train}})}{P(S|\mathcal{M}_{\text{back}})} = \sum_{i=1}^{L} \log \frac{T_{\text{train}}(S_i|S_{i-k,i-1})}{T_{\text{back}}(S_i|S_{i-k,i-1})} \tag{3.23}$$

où T_{train} et T_{back} sont les probabilités de transition associées aux deux modèles, $S_{i,j}$ est la séquence entre les positions i et j incluses, et k est l'ordre de la chaîne de Markov (ici $k = 5$). Cette méthode détecte donc la *signature* globale d'un CRM plutôt que la présence de sites de fixation pour des TFs particuliers. Cette méthode a aussi été implémentée par Ivan et al. (2008) sous le nom de *Markov Chain Discrimination* (MCD), avec la différence notable que les auteurs n'utilisent pas la phylogénie. Une généralisation de cette approche a été proposée par Kazemian et al. (2011) sous le nom d'*Interpolated Markov Model*. Au lieu d'utiliser une chaîne de Markov à un ordre donné, les auteurs réalisent une interpolation entre des chaînes de Markov d'ordres 0 à 5, en ne gardant pour chaque ordre que les transitions sur-représentées dans les séquences d'apprentissage. Ceci leur permet de capturer les signatures présentes à différentes résolutions.

- **Modèles basés sur des enrichissements en k-mers**

D'autres modèles sont basés sur la statistique des mots de k nucléotides (k-mers) dans les séquences d'apprentissage. Par exemple, Kantorovitz et al. (2007) ont introduit une mesure de similarité entre séquences basée sur le nombre de k-mers qu'elles ont en commun. Les auteurs définissent le score D_2 par

$$D_2(S_1, S_2) = \sum_{\{w\}} N_1(w) N_2(w) \tag{3.24}$$

où S_i est la séquence i, $\{w\}$ est l'ensemble des k-mers, et $N_i(w)$ est le nombre de k-mers w dans la séquence i. Ce score est grand si les séquences partagent de nombreux k-mers, c'est-à-dire si elles ont une régulation commune. Ce score est normalisé pour produire le z-score (c'est-à-dire le nombre d'écarts-type par rapport à la moyenne) de mesure de similarité $D2z$:

$$D2z(S_1, S_2) = \frac{D_2(S_1, S_2) - E(D_2)}{\sigma(D_2)} \tag{3.25}$$

où $E(D_2)$ et $\sigma(D_2)$ sont l'espérance et l'écart-type de la distribution de $D_2(S_1, S_2)$, calculés théoriquement sous l'hypothèse que les séquences S_1 et S_2 sont indépendantes et sont générées par un modèle *background* de type chaîne de Markov.

D'autres méthodes pour attribuer un score à une séquence par similarité de k-mers avec des séquences d'apprentissage ont été introduites par Kantorovitz et al. (2009). Étant données des séquences d'apprentissage, les 200 k-mers ($k = 6$) les plus représentés par rapport à un modèle *background* sont sélectionnés selon leur z-score, dans ce cas le nombre d'écarts-type séparant le nombre $n(w)$ de fois que le mots apparaît dans le training set du nombre de fois moyen $\lambda(W)$ qu'il devrait apparaître sous un modèle *background* (Sinha and Tompa, 2000). Étant donnés ces mots sur-représentés, il est possible de définir un score basé sur la statistique de Poisson (modèle PAC pour *Poisson Additive Conditional*) :

$$PAC(S) = \frac{1}{200} \sum_{w} F(\lambda(w), n(w) - 1) \tag{3.26}$$

où $F(\lambda, x)$ est la distribution de Poisson cumulative de paramètre λ, donnant une valeur faible (proche de 0) si $n(w) \simeq \lambda(w)$ et maximale (proche de 1) si $n(w) \gg \lambda(w)$. D'autres scores sont définis par une approche de classification linéaire pondérant les comptages de k-mers (WSC pour *Weighted Sum of Counts*)

$$WSC(S) = \sum_{w} \beta(w) n(w) \tag{3.27}$$

où $\beta(w)$ est un poids reflétant l'association avec l'ensemble d'apprentissage. Ce poids peut être le rapport de la fréquence du mot dans l'ensemble d'apprentissage et de sa fréquence dans le *background* (modèle HexDiff, Chan and Kibler (2005)), le logarithme de cette quantité (Rouault et al., 2010), ou encore le z score introduit précédemment mesurant la sur-représentation du k-mer dans l'ensemble d'apprentissage (méthode HexYMF, Kantorovitz et al. (2009)).

3.1.5 Autres méthodes utilisant des collections d'oligonucléotides

Alors que les méthodes basées sur l'enrichissement en k-mers présentées en 3.1.4 s'intéressent au contenu général d'un CRM en k-mers, d'autres méthodes tentent de regrouper les k-mers en groupes associés à un régulateur putatif. Par exemple, Cao et al. (2010) ont introduit un algorithme de recherche de motifs destiné à l'étude de données ChIP-seq. Ici, un motif est simplement défini comme une collection de k-mers. Le but est de trouver les motifs qui discriminent le mieux un ensemble de séquences positives (des pics de ChIP-seq) d'un ensemble de séquences *background*. L'algorithme énumère d'abord tous les k-mers, mesure leurs fréquences, et ajuste pour chacun un modèle de régression logistique mesurant sa capacité à classifier les séquences. Le k-mer le plus important est choisi comme graine. Puis toutes les variations à distance de Hamming de 1 et 2 (c'est-à-dire ayant un ou deux nucléotides différents) de cette graine sont énumérées, et sont ajoutées au motif si elles permettent d'améliorer la régression. Lorsqu'un motif final est obtenu, toutes ses occurrences sont masquées et un nouveau motif est appris. Un algorithme similaire, HOMER, a été développé par Heinz et al. (2010). La différence majeure est que HOMER utilise la collection de k-mers obtenue pour générer une PWM qui est ensuite raffinée sur les séquences.

3.2 Article

Dans l'article suivant, nous introduisons Imogene, un algorithme de génération de motifs *de novo* utilisant la phylogénie basé sur l'algorithme de Rouault et al. (2010) qu'il généralise au cas des mammifères. Plusieurs tests sont réalisés, montrant sa capacité à prédire des CRMs tissu-spécifiques ou encore à classer différents CRMs selon leur motif d'expression associé.

Published online

Nucleic Acids Research, , Vol. , No. ***1–27***
doi:10.1093/nar/gkn000

Imogene: identification of motifs and cis-regulatory modules underlying gene co-regulation

Hervé Rouault [1,2,+,*], Marc Santolini [3,+], François Schweisguth [1,2] and Vincent Hakim [3*]

[1] Institut Pasteur, Developmental Biology Department,75015 Paris, France, [2]CNRS, URA2578, F-75015 Paris, France, [3]Laboratoire de Physique Statistique, CNRS, École Normale Supérieure, Université P. et M. Curie, Université Paris-Diderot,
[+] Have contributed equally
[*] present address: Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA

ABSTRACT

Cis-regulatory modules (CRMs) and motifs play a central role in tissue and condition-specific gene expression. Their identification could be facilitated by the development of suitable bio-informatic tools. Here we present *Imogene* an ensemble of statistical tools that we have developed and implemented in a publicly available software. Starting from a small training set of mammalian or fly CRMs that drive similar gene expression profiles, *Imogene* determines *de novo* cis-regulatory motifs that underlie this co-expression. It can then predict on a genome-wide scale other CRMs with a regulatory potential similar to the training set. *Imogene* bypasses the need of large data sets for statistical analyses by making central use of the information provided by the sequenced genomes of multiple species, based on the developed statistical tools and explicit models for transcription factor binding site evolution. We test *Imogene* on characterized tissue-specific mouse developmental CRMs. Its ability to identify CRMs with the same specificity based on its *de novo* created motifs equals that of the previously evaluated best motif-blind methods. We further show, both in flies and in mammals, that *Imogene de novo* generated motifs are sufficient to discriminate CRMs related to different developmental programs. Notably, *Imogene* performs as well in this discrimination task purely based on sequence data, than a previously reported learning algorithm based on ChIP data for multiple transcription factors. We thus expect *Imogene* to be a useful tool to decipher transcriptional gene regulation in higher eukaryotes.

INTRODUCTION

The identification and functional characterization of the non-coding sequences that direct the spatio-temporal specificity of gene expression in eukaryotes is of fundamental importance in developmental biology (1) and can find crucial applications in medicine (2). These regulatory sequences are generally located distally from gene promoters and termed enhancers or more generically cis-regulatory modules (CRMs) since they can either enhance or repress gene expression (3). They usually are of the order of 500 nucleotides (nts) long and can be located as far as several mega base-pairs away from the transcription start sites (TSSs) of the genes that they regulate. CRMs are composed of transcription factor binding sites (TFBSs) which bring spatio-temporal specificity to the expression of their target promoters (4). Detailed studies in both flies and vertebrates (5) have shown that CRMs contain multiple binding sites for transcription factors (TFs) that can be either identical (homotypic clustering) or different (heterotypic clustering). Homotypic clustering can provide cooperative TF binding and sharp on-off gene expression whereas heterotypic clustering allows for combinatorial gene regulation. The extent to which the order and relative positioning of the different TFBSs in CRMs matter, remains however debated (6, 7).

With the advent of ChIP-seq techniques, genome-wide studies are providing large amount of data on the binding loci of tissue-specific transcription factors (8), as well as on other factors that regulate transcription e. g. by modifying chromatin structure (p300, CTCF, histone marks, etc) (9, 10). This protein binding data has helped the identification of numerous CRMs specific to well-defined developmental processes and it has brought important information on CRM structure. However, genome wide studies suffer from limitations. A full characterization of regulatory mechanisms would require ChIP-seq analysis to be performed for every potential regulatory factor, on every tissue, at multiple developmental stages. The results would also have to be obtained for the often heterogeneous cells that constitute the tissue of interest instead of being averaged over them as it usually needs to be the case. Finally, and very importantly, binding cannot be equated to functional regulation.

Therefore, *in silico* identification of CRMs forms a useful complement to genome-wide binding studies. Classic case-by-case studies or large scale binding data (11), as previously

*To whom correspondence should be addressed. Email: vincent.hakim@ens.fr

© The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

described, often provide a moderate number (about ten to a few tens) of CRMs, active in the co-regulation of a subset of genes, in specific biological systems or in the formation of different organs at various stages of development. Identifying the important binding sites on these known sequences would help to bypass some of the limitations of large scale studies by providing information on the factor involved, both known and new, as well as on the existence of a regulatory grammar (12). It should also help one to determine other CRMs providing specific expression patterns, a difficult task at present given the absence of close association (13) between CRMs and their target genes in higher eukaryotes. These labor-intensive experimental tasks could be eased by computational work. To this end, we have previously developed (14) statistical tools to determine cis-regulatory elements *de novo*, in a set of input DNA sequences encoding a common transcriptional regulation. They allow the determination of regulatory elements from input DNA sequences without any prior information on the transcription factors acting in cis or on their binding sites. They make central use of the phylogenetic information contained in the aligned DNA sequences of related species. The method was applied to the *D. melanogaster* gene expression program in sensory organ precursor cell (SOPs), a specific type of neural progenitor cells (14). Predicted motifs included already characterized TFBS as well as new motifs and were successfully tested by mutational analysis. These motifs were used to rank intergenic DNA fragments genome wide for their regulatory potential in SOPs. Of the top 29 predicted CRMs, 38% were found by transgenic assays to direct transcription in SOP. A larger fraction (65%) drove more generally transcription in neural precursors.

This successful application to a *Drosophila* transcriptional program led us to try and extend the method developed in ref. (14) to the case of mammalian CRMs. The task of determining cis-regulatory elements is even more difficult for mammalian genomes than for *Drosophila* ones since they are an order of magnitude richer in intergenic sequences (15, 16). To tackle this challenge, we have developed *Imogene*, a computer algorithm and software that we present here and characterize. *Imogene* predicts:

1. cis-regulatory sequences (of about 10 nt long) within a moderate set size of 10-30 CRMs, responsible for specific gene co-regulation, as well as a set of Probability Weight Matrices (PWM) or motifs (17, 18) characterizing the DNA-binding specificity of the associated putative factors.

2. novel CRMs at the genomic scale with the same expression pattern as the starting set of CRMs, based on the set of build PWMs.

Numerous algorithms have already been developed to try and map cis- underlying transcriptional regulation (see e.g. (3, 17, 19, 20, 21) for recent reviews). *Imogene* differs from previous methods in several respects. *Imogene* aim is most similar to the goal of the algorithms analyzed in (22). These algorithms have been specially designed to decode cis-regulatory regulation in a small set of CRMs, contrary to other algorithms which are aimed at the analysis of large datasets such as whole ChIP-seq peak regions (23). Both work *de novo* instead of using already characterized binding motifs (24, 25, 26, 27, 28, 29, 30, 31, 32). Faced to the weak statistical discriminative power offered by the starting set of characterized CRMs, the best algorithms of ref. (22) try and distinguish regulatory sequences by their entire content in short nucleotide sequences as also proposed in other works (33, 34, 35, 36, 37). On the contrary, *Imogene* insists on building cis-regulatory motifs since those are important for experimental work. It instead relies on conservation and the comparison of multiple sequenced genomes.

In the following, the general methodology of *Imogene* is first presented. Then, *Imogene* performance on mammalian CRMs is assessed. Imogene is trained on CRMs pertaining to neural tube and limb developmental programs during embryogenesis. It is shown to successfully classify other CRMs in the same class based on its *de novo* created list of best motifs which contained both new and already known motifs. We then consider the distinct but related task of discriminating CRMs with different specificities, rather than discriminating a set of specific CRMs from background intergenic sequences. *Imogene* is shown to accurately discriminate mammalian neural tube from limb CRMs on the basis of very few learned motifs. To further assess the performance of *Imogene*, it is applied to the discrimination of five sets of mesodermal fly CRMs, a task previously considered in ref. (38). The CRM classification solely based on *Imogene de novo* generated motifs is found to be of similar quality as the results obtained in ref. (38) based on ChIP binding data for multiple transcription factors at several developmental time points. Finally, the developed publicly available *Imogene* interface is presented.

MATERIALS AND METHODS

Genome alignments

The alignments were downloaded from `ftp://ftp.ensembl.org/pub/release-63/emf/ensembl-compara/epo_12_eutherian` for mammals and from `http://www.biostat.wisc.edu/~cdewey/fly_CAF1/data` for Drosophilae. For the latter case, we have used the alignments engineered by A. Caspi with the help of the Mercator and MAVID programs. In both cases, the alignments were processed through a customized script to produce alignments in fasta format, mask for coding sequences (CDS) and simple repeats (see below). These scripts are available in the *Imogene* distribution.

Annotations

The CDS coordinates were downloaded from `ftp://ftp.ensembl.org/pub/release-64/gtf/mus_musculus` for mammals (mm9 coordinates) and from `ftp://ftp.flybase.net/releases/FB2011_06/dmel_r5.38/gff` for Drosophilae (release 5 coordinates). In the case of mammals, the TSS coordinates were obtained separately from `http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database`. Mammalian alignments were already masked for repeat sequences. Drosophilae alignments were masked using the coordinates indicated in the *gff* file.

Phylogenetic trees

The phylogenetic trees used within *Imogene* are displayed in Figure 2. For drosophilae, the distances are taken from Heger and Pontig (39). For mammals, they are obtained from the Ensembl (16) website (`www.ensembl.org`).

Background sequences

Imogene computes the statistical over-representation of the predicted motifs by comparing them to 20 Mb of background intergenic DNA (10^4 regions of 2 Kb). The script that generates the random coordinates is included in the distribution of *Imogene* as well as the actual coordinates of the produced intergenic regions.

Training sets

The two used mammalian training sets (limb, neural tube) were obtained from `http://enhancer.lbl.gov`, based on the work of (11, 40). They were manually curated to produce a high-quality data set, with respectively 41 CRMs for the limb, and 33 for the neural tube. We further pruned out uninformative CRMs for which no motifs could be generated, either because of repeat masking or because of lack of conservation. More precisely, the reference species sequence was scanned using a window size corresponding to the motif size. If a sequence did not contain any masked nucleotide, we looked in the other species for any unmasked sequence in the surrounding neighborhood of ± 20nt, our flexibility criterion when defining a conserved instance. If putative orthologous sequences were found in enough species to satisfy our conservation requirements (see below), the site was declared as a putative conserved site for a regulatory motif. This filtering step resulted in final sets of 39 limb CRMs (minimal length 789 bp, maximal length 9052 bp, average length 3045 bp) and 29 neural tube CRMS (minimal length 585 bp, maximal length 3045 bp, average length 2419 bp).

The Drosophilae training sets were obtained from (38). Coordinate files are given as Supplementary Material.

Main program

The main program is written in C++ and adapted from the program used in a previous study (14). It is distributed under the GNU GPL license and available as a git repository at `http://github.com/hrouault/Imogene`. The user manual is available at `http://hrouault.github.io/Imogene/`. The program can be accessed through a web interface at `http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene`.

Binding site scores

A given motif is represented by a PWM with frequency $w_{i,b}$ for the base b at position i. The index i runs from 1 to l_m the size of the motif which is a parameter in the program which takes the same value for all considered motifs. The binding score of a sequence s_i for such a motif, is defined through the corresponding PWM as:

$$S = \sum_i \log_2 \left(\frac{w_{i,s_i}}{\pi_b} \right) \quad \textbf{(1)}$$

where π_b is the mean frequency of the base b within intergenic regions ($\pi_{\mathrm{A,T}} = 0.30$ and $\pi_{\mathrm{C,G}} = 0.20$) as measured on the "background sequences" (see methods *Background sequences* subsection for their detailed description). A sequence is considered as a binding site in the reference species (*D. melanogaster* or *Mus musculus*) when its score S is larger than the score threshold (S_s or S_g) defined by the user of *Imogene*.

Conservation requirements for binding sites

Imogene iteratively builds PWM from binding sites that have conserved instances in different species. The conservation requirement is that orthologous instances are found in at least 3 distant species, including the reference species. For mammals, the 5 following groups of related species are composed of: *Mus musculus* and *Rattus norvegicus*; *Callithrix jacchus*, *Macaca mulatta*, *Pongo abelii*, *Gorilla gorilla*, *Homo sapiens* and *Pan troglodytes*; *Bos taurus*; *Sus scrofa*; *Canis familiaris*; *Equus caballus*. Similarly for flies, there are 5 groups composed of: *Drosophilae melanogaster*, *sechellia*, *simulans*, *yakuba* and *erecta*; *Drosophila ananassae*; *Drosophilae pseudoobscura* and *persimilis*; *Drosophila willistoni*; *Drosophilae grimshawi*, *mojavensis* and *virilis*.

A site instance must be found in at least 3 of these 5 groups (with an allowed shift of up to 20 nt with the reference species) considered conserved by *Imogene*.

Evolutionary models

Imogene can use two different evolutionary models, which vary in complexity and computational time, to compare orthologous binding sites. In both models, the bases within a site evolve independently from each other.

Felsenstein model. The simplest models of TFBS nucleotides evolution are copied on models of neutral evolution for genomic nucleotides. This procedure has been proposed by Sinha *et al* (24, 41) with the Felsenstein model of neutral evolution (42). In this TFBS evolution model, the transition probability from nucleotide b to b' at position i in two sites at evolutionary distance d is defined as:

$$p^i_{b \to b'} = q\, \delta_{b,b'} + (1-q)\, w_{i,b'} \quad \textbf{(2)}$$

where $\delta_{b,b'}$ is the Kronecker symbol, $w_{i,b'}$ is the mean frequency of base b' at position i of the site (as given by the PWM model), and q is the probability of conservation for an evolutionary distance d under neutral selection (see below).

When two species are close to one another, $q \sim 1$ and the probability that the observed bases are identical is high. On the contrary, when the two considered species are distant ($q \sim 0$), the observed bases are uncorrelated and reflect the PWM probabilities $w_{i,b}$.

The probability of conservation q can then be computed within this model by setting the PWM probabilities $w_{i,b}$ to the mean genomic frequencies π_b:

$$q = \exp\left(-\frac{d}{1/2 + 4\pi_{\mathrm{A,T}}\pi_{\mathrm{C,G}}} \right) \quad \textbf{(3)}$$

with $\pi_{\mathrm{A,T}}$ (resp. $\pi_{\mathrm{C,G}}$) the common genomic frequency of A and T (resp. C and G).

Halpern-Bruno model. The Halpern-Bruno model (HB) (43) differs in two ways from the simplest *Felsenstein* model. It uses the more complex Hasegawa, Kishino and Yano model (HKY) (44) for the neutral evolution of nucleotides and adds a fixation probability based on fitness differences for the evolution of nucleotides within the TFBS.

The HKY model improves on the Felsenstein model by taking into account the observed dependence of the mutation rate on the chemical nature of the bases. Substitutions between bases of the same chemical nature (purine or pyrimidine), also called transitions, are generally more frequent than the other type of mutations, called transversions. This is encapsulated in the HKY model by the parameter κ which is the ratio of the transition rate over the transversion rate. It is measured to be $\kappa=2$ in flies and $\kappa=3.7$ in mammals (45).

Within a TFBS, the HB model extends the HKY model to take into account an additional purifying selection on the nucleotide identities (43). It is formulated by the following transition probabilities:

$$p_{b\to b'}=\exp(t\mathbf{H})_{b,b'} \tag{4}$$

where $\mathbf{H}$ is the rate matrix defined by:

$$H_{b,b'}=\begin{cases}\pi_b\, h_{b'\to b} & \text{if } b\neq b'\\ -\sum_{b'\neq b}H_{b,b'} & \text{if } b=b'\end{cases} \tag{5}$$

The evolutionary time t is expressed in term of the evolutionary distance by:

$$t=\frac{d}{1/2+4\kappa\pi_{\mathrm{A,T}}\pi_{\mathrm{C,G}}} \tag{6}$$

Finally, the transition rates are defined by:

$$h_{b\to b'}=\frac{w_{\mathrm{b'}}}{\pi_{\mathrm{b'}}}\,\frac{\log\left(\frac{\pi_{\mathrm{b}}w_{\mathrm{b'}}}{\pi_{\mathrm{b'}}w_{\mathrm{b}}}\right)}{w_{\mathrm{b'}}/\pi_{\mathrm{b'}}-w_{\mathrm{b}}/\pi_{\mathrm{b}}}\alpha_{b\to b'} \tag{7}$$

with $\alpha_{b\to b'}=\kappa$ for a transition and $\alpha_{b\to b'}=1$ for a transversion.

Inference

The algorithm infers in a Bayesian way the PWM $\mathbf{w}$ frequencies $w_{i,b}$ based on observations of binding sites, as previously described in (14). In a Bayesian framework, the posterior probability $\mathcal{P}(\mathbf{w}|\{\mathcal{A}\})$ that the matrix $\mathbf{w}$ represents the PWM binding to a set of aligned nucleotides $\{\mathcal{A}\}$ is proportional to the product of:

-the *a priori* probability $\mathcal{P}_{ap}(\mathbf{w})$, the '*prior*', that the matrix $\mathbf{w}$ represents a PWM

-the probability $\mathcal{P}(\{\mathcal{A}\}|\mathbf{w})$ of observing the set of aligned nucleotides given that they belong to binding sites for the PWM $\mathbf{w}$.

The prior is taken to be a Dirichlet distribution with parameters α_β at each PWM position

$$\mathcal{P}_{ap}(w_i)\propto\prod_{b\in\{A,T,C,G\}}w_{i,b}^{\alpha_b-1} \tag{8}$$

The nucleotides at different positions are assumed to be independent and the prior for the full site is taken to be the product of the $\mathcal{P}_{ap}(w_i)$ over the different positions. The parameters α_b are taken to be equal for Watson-Crick complementary nucleotides since a sequence and its reverse complement are not distinguished in the description of binding sites (i.e. we assume that binding is not biased toward a particular DNA strand). The two values of α_b are fully determined by assuming that i) TFBS *a priori* have the same nucleotide frequencies as the background and ii) that a PWM mean *a priori* information content is equal to the input threshold score S_g.

The probability $\mathcal{P}(\{\mathcal{A}\}|\mathbf{w})$ of observing the set of aligned nucleotides given the PWM $\mathbf{w}$ is computed in a standard way (42) by recursion for a given PWM $\mathbf{w}$ and a given evolutionary model.

The posterior distribution of the nucleotides frequencies at position i is thus obtained under the form,

$$\mathcal{P}(w_i|\{\mathcal{A}\})\propto\prod_{a\in\{\mathcal{A}\}}\mathcal{P}(a|w_{i,b})\prod_{b\in\{A,T,C,G\}}w_{i,b}^{\alpha_b-1} \tag{9}$$

where we omit the normalization factor.

In the idealistic case where the aligned nucleotides represent independent observations (infinitely distant species), the likelihood reduces to a multinomial distribution and the posterior is given by:

$$\mathcal{P}(w_i|\{\mathcal{A}\})\propto\prod_{b\in\{A,T,C,G\}}w_{i,b}^{N_b+\alpha_b-1} \tag{10}$$

where N_b is the number of times the base b is observed in $\{\mathcal{A}\}$. This formula allows simple analytic formulations for the estimator of mean and maximum posterior probability. The mean posterior estimate is expressed as:

$$\tilde{w}_{i,b}=\frac{N_b+\alpha_b}{\sum_b N_b+\alpha_b} \tag{11}$$

Eq. (**11**) coincides with the maximum likelihood estimate for a Dirichlet 'prior' with parameters α_b+1.

In the case of a non-trivial evolutionary tree (like those of Fig. 2), the orthologous sites are correlated by their evolution from common ancestors. The probability $\mathcal{P}(a|w_{i,b})$ is a polynomial function of the $w_{i,b}$'s. However, it generally lacks a simple analytical expression and the mean posterior estimate should be computed numerically.

Mean Posterior Estimation

The mean posterior estimate was initially computed using a Markov chain Monte Carlo (MCMC) procedure (46). This turned out to be a time-consuming step in the algorithm. To speed it up, we observed, as noted above, that the mean posterior estimate for a prior with Dirichlet parameters α_b coincided with the maximum likelihood estimate for a prior Dirichlet parameters α_b+1 in the case of uncorrelated observations as well as fully correlated ones (i.e. reducing to a single observation). We thus reasoned that maximization with this modified Dirichlet prior could give a quick satisfying approximation for the phylogenetic trees of Fig. 2, which was checked on different examples. This procedure is thus adopted in the present version of *Imogene* and for the results shown here. The posterior distribution obtained with the modified prior is maximized by using the Nelder-Mead simplex algorithm, as implemented in the GNU GSL. The initial value for the estimation is taken to be the mean estimator in the independent species regime given in Eq. (**10**). This allows one to start close to the quadratic region and ensures fast convergence.

A simple example of nucleotide inference using the two evolutionary models

To illustrate the inference of ancestral nucleotides and the main features of the two models, we consider in Figure S5 a dinucleotidic genome with bases X and Y and a simple phylogenetic tree with an ancestral species at equal evolutionary distance from the reference species and a daughter species. We suppose that the observed nucleotide at position i of an observed binding site is X both in the reference and the orthologous species.

Our goal is to infer the frequencies w_Y and $w_X=1-w_Y$. First, there are two simple cases. For $d=0$, the observations of the same nucleotide in the two evolutionary branches really constitute only one observation of X. On the contrary, for very long evolutionary branches $d\to\infty$, the two instances of nucleotide X form two independent observations. Using the previous result (Eq. (**11**)) with $\alpha_X=\alpha_Y=\alpha$, the estimator of the maximum transformed posterior distribution for N_X and N_y independent instances of X and Y is:

$$w_Y=\frac{N_Y+\alpha}{N_Y+N_X+2\alpha} \tag{12}$$

Thus, for $d=0$, the inferred frequency is:

$$w_Y=\frac{\alpha}{1+2\alpha} \tag{13}$$

while for $d\to\infty$, it tends toward:

$$w_Y=\frac{\alpha}{2+2\alpha} \tag{14}$$

Between these two extreme cases, an evolutionary model has to be used to estimate w_Y, for finite evolutionary branches of length d.

For the Felsenstein model, the likelihood function writes:

$$\begin{aligned}\mathcal{P}(\mathcal{A}|w)&=w_X\,[q+(1-q)w_X]^2+w_Y(1-q)^2w_X^2\\&=q^2w_X+(1-q^2)w_X^2\end{aligned} \tag{15}$$

where $\mathcal{A}$ stands for the simple alignment considered in Figure S5 and we used $w_X=1-w_Y$. From this expression it can clearly be seen that the evolutionary model simply interpolates between the independent species case ($d\to\infty$, $q=0$) where there are two observations of base X: $\mathcal{P}(w|\mathcal{A})=w_X^2$, and the fully correlated case ($d=0$, $q=1$) where the two species merge and we have only one observation: $\mathcal{P}(w|\mathcal{A})=w_X$. The corresponding mean, $w_{Y,me}$ and maximum posterior, $w_{Y,ma}$ analytic estimates for finite d read

$$\begin{aligned}w_{Y,me}&=\frac{\alpha}{2}\,\frac{1+q^2}{\alpha+1+\alpha q^2}\\w_{Y,ma}&=\frac{1}{4(\alpha+1)(1-q^2)}\Big[3\alpha+2-(\alpha+1)q^2\\&\quad-\sqrt{[\alpha+2-3(\alpha+1)q^2]^2+8q^2(1-q^2)(\alpha+1)^2}\Big]\end{aligned}$$

Note that for the maximum posterior estimate, $w_{Y,ma}$, the prior exponent $\alpha+1$ has been used instead of α as explained above. So, the two estimates coincides at $q=0$ and $q=1$. Both estimates are plotted as of function of the evolutionary distance d in Figure S5 ($\alpha=0.1$).

For the Halpern-Bruno model, the analogous results have been computed numerically and are also shown for comparison in Figure S5 . The Halpern-Bruno model results are seen to be closer to the large distance limit than the Felsenstein model ones. Moreover, the difference between the nature of the estimates is seen to be comparable to the difference between the evolutionary models.

Filtering of motifs coming from simple repeats

Imogene pre-processes the training set by masking repeated sequences with repeat masker (47) but this is not sufficient to eliminate the production of motifs corresponding to repeated sequences. These motifs have a non-poissonian distribution of binding sites on intergenic sequences: one binding site has a high probability to be followed by another one after a multiple of the repeat period. This anomalous distribution of binding sites biases motif ranking and diminishes the algorithm CRM predicting power (14). Motifs corresponding to repeated sequences are thus filtered out using the non-poissonian characteristics of their binding site distribution. The binding sites of each motif m are determined on the above-described set of $N_{bg}=10^4$ background sequences of length $L=2\times10^3$ nt. For a Poisson distribution, one would expect the number $N_m^{(p)}(j)$ of intergenic sequences containing j binding sites to be

$$N_m^{(p)}(j)=N_{bg}\frac{(\lambda_m^{(bg)}L)^j}{j!}\exp(-\lambda_m^{(bg)}L) \tag{16}$$

where $\lambda_m^{(bg)}$ is the computed density of binding sites of the motif m in the set of background sequences. The deviation from this theoretical Poisson distribution is quantitatively assessed by computing the χ^2-like value,

$$\chi^2(m) = \sum_j \frac{[N_m(j) - N_m^{(p)}(j)]^2}{N_m^{(p)}(j)} \Theta(N_m(j)) \quad \textbf{(17)}$$

where Θ is the Heaviside function ($\Theta(x) = 0$ for $x < 0, \Theta(x) = 1$ for $x > 0$) which restricts the sum to non-zero values of $N_m(j)$. Only the 75 % motifs with the lowest $\chi^2(m)$-value are retained for subsequent computations.

Distance between motifs

The similarity between two motifs is quantitatively assessed based on the overlap between the sets of their binding sites. The 'strict proximity' between motifs represented by two PWMs $\mathbf{w_1}$ and $\mathbf{w_2}$, is defined by

$$\mathrm{Prox}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) = 2\frac{\mathrm{Prob}\left\{\left[S(\mathbf{s}, \mathbf{w}^{(1)}) > S_{th}\right] \text{ and } [S\left(\mathbf{s}, \mathbf{w}^{(2)}\right) > S_{th}]\right\}}{\mathrm{Prob}\{S(\mathbf{s}, \mathbf{w}^{(1)}) > S_{th}\} + \mathrm{Prob}\{S(\mathbf{s}, \mathbf{w}^{(2)}) > S_{th}\}} \quad \textbf{(18)}$$

where $\mathrm{Prob}\{S(\mathbf{s}, \mathbf{w}) > S_{th}\}$ is the probability that a sequence $\mathbf{s}$ drawn at random with the background frequencies π_b has a binding score $S(\mathbf{s}, \mathbf{w})$ (Eq. (**1**)) above the threshold S_{th} for the frequency matrix $\mathbf{w}$. The strict proximity is computed analytically as explained in (14), where it was defined. To take into account potential shifs in the motifs or in their orientation, $\mathrm{Prox}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$is computed for all possible alignments of the two matrices (with a maximum shift of $l_m/2$ where l_m is motif size) in the two possible orientations. When shifted matrices are compared, they are completed by additional columns with the background frequencies (i. e. with no specifity). The proximity between two motifs is obtained simply by taking the maximum over the obtained strict proximities. It goes from 1 for two identical motifs to zero for motifs that do not share any binding site above the threshold. *Imogene* distance between two motifs is defined as minus the logarithm of their proximity.

Ranking motifs

The previous filtering step provides for each considered motif m, the density $\lambda_m^{(bg)}$ of its binding sites on the background sequences and ensures that these sites are approximately distributed in a poissonian way. The deviation from this baseline distribution on the CRM of the training set (t.s.) is used to score each motif. This is quantified by the poissonian log-likelihood of the training set

$$Pl(m) = \sum_{t \in \{\text{t.s.}\}} \log\left(\frac{\left(L_t \lambda_m^{(bg)}\right)^{k_t} \exp(-L_t \lambda_m^{(bg)})}{k_t!}\right) \quad \textbf{(19)}$$

where k_t is the number of instances of m on the training set sequence t of length L_t. Larger deviations from the baseline poissonian distribution are supposed to reflect motif specificity for the training set and correspond to more negative/better scores.

Scoring intergenic sequences

Given a list of motifs m_i, a CRM E is scored as follows:

$$S(E) = \sum_i n(E, m_i) \log(\lambda_i^t / \lambda_i^b) \quad \textbf{(20)}$$

where $n(E, m_i)$ is the number of binding sites for the motif m_i on E and λ_i^t, λ_i^b are the average number of binding sites per base on the training set and background respectively. It is important to note that the previously found motif binding sites are masked when scanning with successive motifs. Thus motifs with lower ranks that resemble high-ranking motifs, do not increase artificially the CRM weight by predicting the same binding sequences twice.

Selection of optimal intergenic sequences

When ranking genome-wide intergenic sequences, with a list of N motifs, the best intergenic sequence at a given position is determined as follows. The list of motifs is used to scan the genome for conserved binding sites above a given threshold. Binding sites are then grouped in successive CRMs of size L such as to maximize clustering. The position E_i of the center of the enhancer i is chosen to be the center of the motifs cluster:

$$E_i = \frac{X_1 + X_N + l_m - 1}{2} \quad \textbf{(21)}$$

where X_1 and X_N are the starting positions of the first and last TFBSs in the cluster and l_m is the width of the motif.

Mammalian predictions

Learning sets, test sets and background test sets. For each class, the CRMs were divided into a learning set composed of 15 CRMs chosen at random, the other CRMs (~ 20) defining the test set of 'True positives'. In addition, a set of background test regions was built using the 1Kb flanking sequences of the full list of CRMs.

Such an 'adapted' background test set was used to provide a more stringent and informative test of the algorithm. It prevents discrimination on the training set from the background test set, based on other features than the sought high-information-content motifs, such as a local composition bias. Furthermore, in order to avoid biasing the results towards the true positives, uninformative sequences for *Imogene* (i.e sequences where no binding site could possibly be found given *Imogene* conservation requirements) were also removed from this background test set. These regions were also filtered for uninformative elements. This yielded background test sets of 72 CRMs for the limb and 57 for the neural tube.

Cross-validation protocol. The learning set was used to learn the motifs content. The 10 best motifs were then used to score test set CRMs and background regions. Because the length of the training set CRMs could vary, we decided to

keep for each test sequence the best scoring 1kb fragment. This process was repeated 40 times, and both generation and scanning threshold were varied. The retrieval rate of test set CRMs (True Positives) among background elements (False Positives) as a function of the score was used to build a ROC curve. The Area Under ROC Curve or AUC, a quantity that varies between 0 for absolute misclassification, 0.5 for random classification, to 1 for perfect classification, was used to evaluate the quality of prediction. The parameter set yielding the highest AUC was chosen as the best set.

Leave-one-out cross-validation for the CRM discrimination task.

Let us note $\mathcal{C}_i$ the tissue class of interest. There are M_i corresponding CRMs. Let N_c denote the total number of classes. Our goal is to find the particular motif signature that distinguishes these M_i CRMs from the N_c-1 other classes of CRMs. This signature corresponds in our case to a number N of top ranked motifs with generation and scanning thresholds S_g and S_s. These are the three parameters we wish to constrain with a leave-one-out cross-validation (LOOCV) procedure.

Let us detail this procedure in the case where we distinguish class $\mathcal{C}_i$ from the other classes $\mathcal{C}_j$. The M_i CRMs of $\mathcal{C}_i$ are termed 'positive' CRMs and the M_j CRMs of each of the other classes are termed 'negative' CRMs. Let us note $M=\sum_i M_i$ the total number of CRMs. The LOOCV consists in withdrawing one 'test' CRM from these M CRMs, learn the motifs on the $M-1$ resulting CRMs, and use them to score the let alone test CRM. For the learning step, motifs are generated with threshold S_g on each class (one class being deprived of one CRM), yielding N_c sets of motifs: one set of positive motifs from class $\mathcal{C}_i$ and N_c-1 sets of negative motifs from the other classes. The N top ranked motifs from each set are then used to scan the M CRMs for conserved instances with scanning threshold S_s. Each CRM E is scored with respect to these N_c sets of motifs by:

$$S(E)=\sum_{j=1}^{N_c}(2\delta_{j,i}-1)S_N^{\mathcal{C}_j}(E) \qquad \textbf{(22)}$$

where $S_N^{\mathcal{C}_j}(E)$ is the CRM score for the N top motifs of class $\mathcal{C}_j$ as defined below in the 'Main program' description, and $\delta_{j,i}=1$ if $j=i$, and 0 otherwise. This score simply gives positive contributions if positive motifs are found on the CRM, and negative contributions if negative motifs are found. This scoring procedure allows to rank the test CRM among the other $M-1$ CRMs. Ties are resolved by attributing their mean rank to equally scored CRMs. The rank of the test CRM is used rather than its raw score to avoid potential bias stemming from score normalization. Indeed, the raw score is dependent on the generated motifs, which differ at each step of the LOOCV. This procedure is repeated over all M CRMs, yielding a corresponding list of M ranks. This list is finally used to build a ROC curve discriminating True Positives (CRMs from class $\mathcal{C}_i$) from False Positives (the other CRMs). The discrimination is quantified by the area under the ROC curve for a False Positive Rate FPR $\leq 20\%$, which we note AUC20 and that we want to maximize.

In our case, we used a $2D$ parameters grid with S_g varying between 7 and 13 bits by steps of 1, and S_s varying between S_g-5 and S_g by steps of 1. Both *Felsenstein* and *Halpern-Bruno* models were used for motif generation. For each parameter set, the number of motifs used for scanning was increased from 1 to a maximum number of 10 (actually never attained) until the addition of a new motif decreased the AUC20, yielding an optimal number of motifs N. Finally, for each class, the parameter set $\{S_g,S_s,N\}$ yielding the highest AUC20 was selected as the best parameter set.

Motifs identification

In order to identify the known TFs that might correspond to the *de novo* generated motifs, we used Transfac database (48). In order to avoid uninformative matches, we kept Transfac motifs that had an information content greater than 8 bits, a threshold approximately corresponding to 4 conserved nucleotides. This gets rid of 170 vertebrate motifs and 32 insect motifs, yielding a total of respectively 765 and 37 motifs.

Each *de novo* motif was compared to all Transfac motifs from the corresponding clade (vertebrates or insects) using the PWM distance introduced in (14). During the comparison, motifs are shifted to find the best match, with a minimal match of 5 nts. The shift is simply introduced by adding flanking nucleotides with background frequency on either side. The closest candidate was kept for identification.

Statistical analysis

All statistical analyses were performed using R (49).

RESULTS AND DISCUSSION

Description of Imogene

Imogene has two modes that can be used in succession, as sketched on Figure 1 and summarized here (see *Methods* for details of their implementation).

The first mode, *Genmot*, aims at extracting statistically meaningful PWMs from a "training set" of functionally related CRMs on a reference genome (the mouse *M. musculus* genome for mammals; the *D. melanogaster* genome for flies). The cumulated size of the training set could in principle be unlimited, but in practice computer execution time requires it to stay below 100 Kbp. It should also be above a few Kbp to provide a sufficient amount of information (a training set of about 20 Kbp appears as a good compromise). Starting from a chosen training set, *Genmot* performs its task in two steps (I and II in Figure 1): I. *Genmot* first enlarges the training set with aligned orthologous sequences in other related sequenced genomes (see *genome alignments* in *Methods*), as shown in Figure 2 (for the mouse, the 11 other aligned mammalian sequenced genomes with high coverage presently available on the Ensembl project (16), the 11 other *Drosophilae* sequenced genomes (15) for the fly). This comparative genomics step results in the creation of the "enlarged training set" (step I in Figure 1).

II. In this second central step, *Genmot* build PWMs of given length ℓ (10 nt is the default value) by scanning the training set, in an iterative manner (step II in Figure 1). Each sequence of ℓ nucleotides in the training set is used in turn to create an

initial PWM using a Bayesian prior. This PWM is then refined by scanning the training set to find all the PWM binding sites in the training set, i.e. all ℓ nucleotide long sequences in the training set that have a binding score above a generation threshold score S_g, chosen at the procedure onset ($S_g = 13$ bits is the default value). These binding sites are filtered using conservation, that is only sites that have orthologues in distant species are further considered (see *Conservation requirements for binding sites* in *Methods*). A shift in alignment between a binding site on the reference species and its orthologues in other species is allowed for the correction of eventual alignment errors (20nt is the shift default value). The ensemble of conserved binding sites and their orthologues serve, using an evolutionary model, to build a refined PWM. The procedure is then iterated by finding the binding sites of the refined PWM and using them to build a further refined PWM, until convergence to a stable set of binding sites.

The need of an evolutionary model to properly assemble binding sites (24, 25, 41) is simply explained. A binding site in the reference genome and its orthologues are all related through descent from their last common ancestor, and cannot therefore be considered as independent observations. In order to correctly quantify the amount of information provided by the observation of orthologous sites, one has to estimate their potential of change through mutation since their last common ancestor. To account for this, *Imogene* can, in its present implementation, make use of either one of two evolutionary models of TFBS evolution at the user choice. The first option, *"Felsenstein model"*, is a simple and computationally fast model proposed in (41). Mutations are generated at the same rate in a PWM binding site than in the background intergenic sequences. However, the mutated nucleotide in a binding site is drawn according to its frequency in the PWM at the mutated position. This is analogous to the simplest model of DNA evolution (42) but with nucleotides neutral relative abundances replaced by PWM nucleotide frequencies. This *Felsenstein* model is the simplest model that provides at evolutionary equilibrium, nucleotide frequencies that agree with those prescribed by the PWM at the different positions in the binding site. The second option, *"Halpern-Bruno model"* (43) uses an evolutionary model that is more complex than the Felsenstein model but that is also more clearly grounded on theoretical population genetics ideas. It has previously been used for TFBS evolution in (25). It allows for the inclusion of different mutational probabilities between different bases in the neutral background intergenic mutation model. Additionally, it includes a fitness-dependent fixation probability for a mutation in a TFBS, based on classical population genetics estimates for the fixation of a mutant allele appearing in an homogeneous population (50). The relative fitnesses of different nucleotides are determined by the requirement that binding site convergence to evolutionary equilibrium leads to the PWM nucleotide frequencies (see Methods for details).

The described procedure produces a PWM for each ℓ nucleotide long sequences in the training set. In a series of final steps (see *Methods* for a mathematically detailed description), this long list is pruned and ranked based on comparing the PWM bindings sites on the training set to a "background" set of intergenic sequences in the reference genome (20 Mb of *M. Musculus* or *D. melanogaster* genomic DNA). *Imogene* pre-processes the training set by masking repeated sequences with repeat masker (47) but, as noted in ref. (14), this is not sufficient to eliminate some PWMs corresponding to repeated sequences from the produced list of PWMs. These PWMs have statistically anomalous distributions of binding sites that bias their subsequent ranking. Therefore, in a filtering first step, PWMs corresponding to repeated sequences are discarded on the basis of their anomalous distribution of their binding sites in the background set (see *Filtering of motifs coming from simple repeats* in *Methods*). Then for each remaining PWM, the distribution of its conserved binding sequences on the training set is compared to the distribution of the PWM conserved binding sequences on the set of background intergenic sequences. The larger the statistical deviation between the two distributions, the larger its score and the more meaningful the PWM is deemed (see *Ranking motifs* in *Methods*). In a final step, PWMs in the ranked list are compared (see *Distance between motifs* in *Methods*)) and, among similar ones, only the highest scoring one is kept. Although the identity of the transcription factors corresponding to the different PWMs of interest is not directly assessable by the algorithm, the comparison between the produced PWMs and existing databases can provide relevant information on their identity, as will be shown in the following sections.

In its second mode, *Scangen*, *Imogene* determines intergenic sequences in the reference genome that are considered as putative CRMs with the same functional specificity as the training set. This second mode (step III in Figure 1) is based on the inferred PWMs in the *Genmot* mode. The algorithm scans the entire non-coding repeat-masked reference genome and find all the conserved binding sites above the scanning binding score S_s for the N first PWMs in the ranked list. The intergenic sequences of a given length (the default value is 1000 nt) are then scored according to their similarity to the training set in their content of PWM binding sites (see *Scoring intergenic sequences* in *Methods*) The closest the similarity in its motif content with the training set, the most likely an intergenic sequences is deemed to be functionally related to the training set.

Application to mammalian developmental programs

In order to assess *Imogene* performance on mammalian transcriptional regulation, we applied it to two sets of mammalian specific CRMs, that have previously been identified starting from p300 Chip-seq data and functionally tested in a transient transgenic assay for activity in stage 10 mouse embryo (11, 40). We chose CRMs active in neural tube and limb, as characterized in the VISTA website (`http://enhancer.lbl.gov`). For each developmental program, a subset of CRMs was visually selected for specificity and strength of expression in the tissue of interest, from the provided expression pattern. Among these selected sets, 2 limb CRMs and 4 neural tube CRMs contained no sequence that could possibly be used to learn motifs by *Imogene*, due to its conservation requirements, either because of repeat masking or because of low conservation (see *Methods*). Elimination of these uninformative sequences produced curated training sets of 29 neural and 39 limb CRMs (see *Training sets* in *Methods*).

A cross-validation scheme was then used to measure *Imogene* predictability power (see *Methods* for details). In brief, for each developmental program, the CRMs of the training set were divided into a learning set composed of 15 CRMs chosen at random, and a test set composed of the other CRMs used as True Positives.

The learning set was used for motifs generation using *Imogene Genmot* mode. This procedure was conducted for both evolutionary models using different values of the generation parameter S_g and scanning threshold S_s to obtain the optimal values of these parameters for each model and each learning set (see Figure 3 and Figure S1).

The test CRMs of the training set were then ranked, using motifs generated on the learning set, against a 'background test set', a set of ~ 60 regions of 1Kb taken from the flanking sequences of the initial set of CRMs (see *Methods*).

For different parameter sets, the test CRMs as well as the intergenic sequences of the background set were scored. The proportion of retrieved test set CRMs above a given score (True Positive Rate or TPR) was plotted against the proportion of appearing test background regions above the same score (False Positive Rate or FPR) as this score decreased, to produce a so-called ROC curve (51). The ROC curves corresponding to different parameters values were then compared using the Area Under ROC Curve (AUC), a quantity that is maximal at best prediction. Figure S1 shows the AUC as a function of the number of motifs N for different values of the scanning threshold S_s. One can see that the AUC increases quickly with the 5 first motifs generated, and has nearly converged to its maximum value when 10 motifs are kept. Therefore we restricted ourselves to $N=10$ motifs, and constrained the other parameters using AUC maximization. Figure 3 shows the ROC curves obtained for the optimal parameters which are seen to be similar for both models and both training sets. For the neural tube CRMs, 30% of the test set CRMs are retrieved at 1% FPR whereas an even larger proportion of 40% is obtained for the limb CRMs. The *Halpern-Bruno* and the *Felsenstein* models are seen in Figure 3 to yield very similar results in both cases. It should be noted that the test really provides only a lower estimate of *Imogene* success rate. Sequences of the background test set counted as 'False Positive' could, in reality, be *bona fide* positive CRMs.

The performance of *Imogene* is found to be comparable to the best motif-blind methods (22). Using a cross-validation protocol similar to the one used here, in which the CRMs to be tested were compared to flanking sequences, the 'HexMCD' was found to be top-scoring method for the set of limb CRMs. It recovered 60% of the training set for a $5-10\%$ FPR. For neural tube CRMs, the two best methods, 'PAC-rc' and 'D2z-cond-weights' recovered 80% and 74% of the test set for a $5-10\%$ FPR (see Figure S5 in ref. (22)).

One interesting feature of *Imogene* lies in its production of specific motifs. In our cross-validation procedure, different ranked lists of motifs were created for each randomly drawn test set. In order to provide a list of motifs generated by the algorithm, we ran *Imogene* on the full set of CRMs for each class. The corresponding 10 best motifs are shown in Figure S2. The closest TRANSFAC PWM assigned to each motif by *Imogene* PWM distance is also shown in Figure S2. Previously characterized motifs belonging to the considered developmental programs appear in each class (e.g. Oct1/Pou2f3 family and NeuroD motif in the neural CRMs). The motif content of each CRM is also provided in Figures S3, S4. It is seen that the 10 best motifs appear on most CRMs of the training set.

Discrimination of tissue-specific CRMs in the mouse

Given the ability of *Imogene* to distinguish specific CRMs from background sequences, we found it interesting to apply it to the related but distinct task of distinguishing different classes of CRMs. The question was previously considered for *D. melanogaster* CRMs based on ChIP-seq data at different developmental time points (38), as detailed in the next section. It consists in learning features that distinguishes the CRMs of a given class from the CRMs of other classes, in order to be able to predict the class of a newly observed CRM. The task differs from distinguishing CRMs from background intergenic sequences since learning motifs shared among different classes, for instance characterizing the binding of generic CRM factors, is of no use for discrimination purposes. As a test case, we considered the neural tube and limb sets of mammalian CRMs used in the previous section. Given the nature of the task, we selected in each set the CRMs with an expression that appeared mostly restricted to neural tube and limb. This yielded 12 neural and 15 limb CRMs.

As in ref. (38), we used a leave-one-out cross validation (LOOCV) scheme in which the learning set constituted all but one of the elements of a class, the remaining one being used as a test sequence. The process can be summarized as follow. We call the class of interest the positive class and the classes against which we wish to learn the negative classes. The LOOCV process begins with the exclusion of a (positive or negative) CRM which serves as an unobserved test CRM. Then, a set of N motifs is learnt on the remaining CRMs of each class, yielding positive and negative motifs. These motifs are used to build a simple linear classifier based on a weighted score giving positive (resp. negative) contributions to positive (resp. negative) motifs (see *Methods*). Finally, the test CRM is ranked among all CRMs by the build classifier and this rank is registered. A successful classification would rank positive CRMs on top of the list and attribute worse ranks to negative CRMs. Therefore, after processing all CRMs, the list of ranks for the positive and negative CRMs is represented as a ROC curve indicating the True Positives Rate and False Positive Rate for increasing rank. This serves to optimize the different parameters (the threshold for motifs generation S_g, the threshold for sequences scanning S_s, and the number of motifs N used to score sequences) by maximizing the Area Under the ROC Curve for a FPR ≤ 0.2.

The results are shown in Figure 4. We focus on the results obtained with the *Halpern-Bruno* evolutionary model. Results (motifs, thresholds) are very comparable in the two cases. Motifs are shown on the right of the ROC plots and were generated on the positive classes with optimal parameters. The two classes were optimally discriminated using only 2 motifs in each class, with specificities $S_g=11$, $S_s=8$, comparable to that found in the learning task of the previous section. The best ranking motif of the neural CRMs was found to be unequivocally associated to the Transfac Oct1/Pou2f3

Transcription Factor, known to be involved in the neural tube formation (52).

Discrimination of Drosophila tissue-specific CRMs

In order to further test the discriminating power of *Imogene de novo* generated motifs, we applied it to the CRM classification task reported in ref. (38). In this work, previously characterized *D. melanogaster* CRM were divided in 5 classes corresponding to the different tissue types in which they were active: mesoderm (Meso), somatic muscle (SM), visceral muscle (VM), mesoderm and somatic muscle (Meso & SM) and visceral and somatic muscle (VM & SM). Ref. (38) made use of a collection of Chip-seq binding data for different factors and at different developmental time points to attribute to each CRM a total of 15 peak height values. It was then tested whether classical machine learning techniques could be used to discriminate the different CRM classes, on the basis of these extensive data. This was indeed found possible with a high success rate in a standard cross-validation scheme: CRMs predicted with probability higher than 95% to belong to a given class were indeed found to belong to that class with a high success rate of 80%.

This led us to wonder whether *Imogene* would succeed in classifying these different CRMs, without using any binding data, but rather on the basis of combinations of *de novo* motifs that it would itself generate. We used the set of well-characterized CRMs belonging to 5 different classes assembled in ref. (38). We then proceeded as in the previous case of mammalian CRMs.

Imogene results are shown together with the machine learning results of ref (38) in Figure 5. For clarity, we here show results obtained with the *Felsenstein* model. Results obtained with the *Halpern-Bruno* model are comparable. Strikingly, without any binding data *Imogene* prediction rates are comparable to the machine learning ones, in the specificity range (FPR $\leq 5\%$) used for CRM prediction in (38). Its performance is even better for the Meso and SM classes at high score. The latter case is of particular interest. The machine learning algorithm essentially used Mef2 ChIPseq peak heights to predict SM CRMs, resulting in an incorrect classification at high scores since this TF is required for the differentiation of all muscle types. However, the use of the specific Mef2 motif obtained *de novo* from the SM training set allows one to restore a correct classification at high score (Figure 5C).

On the side of each ROC plot, the *de novo* motifs generated on the whole training set are displayed. The number of motifs shown is the optimal number used for CRM scoring in the leave-one-out cross-validation. Among the generated motifs, one can recognize $4/5$ TFs for which ChIPseq data was used in (38), namely Twist (motif 2, Meso & SM), Mef2 (motif 1, SM), Bin and Tin (motifs 1 and 2, VM). The Bap motif was not found by the algorithm, and correspondingly it was not shown to be of importance in ref. (38).

In summary, our analysis indicates that *Imogene* not only determines *de novo* functionally relevant binding sites within a set of CRMs but can also be used to identify the more subtle differences in binding sites that underlie functional differences between related sets of CRMs.

Web interface

The ensemble of developed statistical tools and the allied computer codes are freely available at `http://github.com/hrouault/Imogene`. In addition, they can be used through a user-friendly web interface (`http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene`) that provides motif and CRM predictions for the community. This interface is powered by the Pasteur Institute Internet server through the mobyle framework (53). The input web page and an example output web page are shown in Figure 6 and 7.

The input form (see Figure 6) is divided into several sections. One of the two available algorithm modes should be chosen at start:

- Genmot: given a list of coordinates of typically 15 enhancers of 1 kb (training set), generates *de novo* motifs ranked by their score ($Pl(m)$ in *Methods*).

- Scangen: given the previously generated motifs, produces a list of genome-wide predicted CRMs with conserved binding sites. The rank of a CRM is based on a poissonian score that takes into account the motif content (as described in *Methods*)

The group of species considered should also be specified. The algorithm can be used on Drosophilae (with reference species *D. melanogaster*) or mammals (with reference species *Mus musculus*). The different algorithm parameters such as the sought motif width, threshold specificity for binding sites or allowed position shifts between different species (see *Methods* for a detailed description) are set by default to values that have been found to provide reasonable results. They can be modified by the user to optimize the results for other training sets.

In mode *Genmot*, the user should enter the training set CRM coordinates. The chosen evolutionary model for the TFBS should also be specified. The *Felsenstein* mode is computationally faster than the *Halpern-Bruno* one. The results of the two modes have been found to be comparable (see Figure 3 and 4).

In mode *Scangen*, the algorithm scores and ranks intergenic sequences in the reference species, using a list of motifs, as described in the first *Results* section and in *Methods*. The list of *de novo Genmot* motifs can be used as input. The user can set the length of the ranked sequences (1 Kb is the default value) and the number of scoring motifs (5 is the default value). The default values have been chosen for computational efficiency but changes can improve results (see Figure S1).

An example of *Imogene* output is displayed in Figure 7. The *Genmot* mode creates from the provided training set a list of ranked motifs together with their significance and over-representations (see *Methods*). The positions of these motifs on the CRM of the training set and on their homologous sequences in other species are also provided, as illustrated in Figure 7A for 2 motifs. Figure 7B shows the output of the *Scangen* mode for these two motifs. The ordered list of best-ranking intergenic sequences is given together with information on the closest TSSs.

DISCUSSION AND CONCLUSION

We have presented *Imogene*, a set of statistical tools and a computer software able to predict *de novo* relevant motifs in a moderate size set of functionally related CRMs and able to infer novel CRMs with a low false positive rate in both Drosophilae and mammalian genomes. *Imogene* mode of inference internally makes use of quantitative models for binding site evolution. This allows it to systematically exploits the information available in multiple sequenced-genomes, and to work efficiently from a CRM set of modest size. It leads it to achieve a performance comparable to the best motif-blind algorithms (22).

Phylogenetic conservation between multiple sequenced genomes has previously been shown to provide useful information on cis-regulatory motifs (54, 55, 56) but cannot *per se* address the question of specific spatio-temporal expression. The necessary information is provided to *Imogene* by the training set of CRMs with well-characterized expression. *Imogene* aim is to extract it optimally by making full use of several sequenced genomes, instead of focusing on a single genome (26) analysis, simply comparing the reference genome with another one (57, 58, 59) or simply adding orthologous sequences (60). Similarly to the *Monkey* algorithm of ref. (25), *Imogene* uses a model for the evolution of motif binding sites, to properly weigh this additional information. The two algorithms are however complementary since *Imogene* creates *de novo* motifs from the training set while *Monkey* tests already well-characterized binding motifs.

The algorithm which lies at *Imogene* core was previously applied to gene co-regulation in *Drosophila* (14). Motifs predicted to be important for Sensory-Organ-Precursors development were confirmed by site-directed mutagenesis. A significant fraction of top predicted new CRMs were also shown to direct expression in SOP or more generally in the peripheral nervous system. The ability of the algorithm to provide meaningful information on cis-regulatory elements in *Drosophila* was further confirmed in a subsequent application to epidermal morphogenesis and trichome development (61). The algorithm provided an informative PWM for the master regulator Ovo/Shavenbaby and predicted as well a functionally important novel motif.

In spite of its successful application to gene co-regulation in *Drosophila*, it was not clear that the method could be successfully extended to decipher cis-regulatory information in the notoriously more difficult case of mammalian gene expression. We have here provided bioinformatics evidence that our developed algorithm indeed provides meaningful results in this case also. *Imogene* was shown to successfully recognize CRMs belonging to neural and limb development programs solely based on motifs that it has constructed *de novo* from the analysis of other CRMs. Furthermore, the created PWMs appear to comprise both known and new motifs, in strong analogy with the previous studied cases in the fly.

There is currently numerous cases for which a small number of CRMs belonging to the same program of gene expression has been characterized. At the same time a large number of PWMs remain to be found. This is even more the case for CRMs. Therefore, the use of *Imogene* with its *de novo* motif building ability and allied CRM identification, should provide helpful service to the community.

We have further shown that *Imogene* can discriminate between classes of CRMs. In this task, it should usefully complement ChIP-seq data that are currently obtained for many developmental programs. Whereas ChIP-seq provides information on the binding of already known factors, *Imogene* is able to propose new motifs and help to identify new involved DNA-binding cofactors and their binding sites. We thus believe that *Imogene* is a useful addition to existing algorithms and softwares (26). We hope that it will serve as a helpful and timely tool in the difficult deciphering of gene regulation in higher eukaryotes.

ACKNOWLEDGMENTS

We wish to thank I Leroux, S Meilhac and B Robert who helped us to characterize the patterns of expression of the mammalian CRMs used in the present work and S. Eddy for his critical reading of the manuscript. We acknowledge the Centre d'Informatique pour la Biologie at the Pasteur institute for its help in the design of a mobyle front-end to *Imogene*. This work was supported by core funding from Centre National de la Recherche Scientifique, Ecole Normale Supérieure and Institut Pasteur and by a specific grant from the Agence Nationale pour la Recherche (ANR-08-BLAN-0235).

Conflict of interest statement. None declared.

REFERENCES

1. Davidson, E. H. (2006) The regulatory genome: gene regulatory networks in development and evolution, Academic, Burlington, MA.
2. Dorer, D. E. and Nettelbeck, D. M. (Jul, 2009) Targeting cancer by transcriptional control in cancer gene therapy and viral oncolysis. *Adv Drug Deliv Rev,* **61**(7-8), 554–71.
3. Hardison, R. C. and Taylor, J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.,* **13**(7), 469–483.
4. Lelli, K. M., Slattery, M., and Mann, R. S. (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.,* **46**, 43–68.
5. Levine, M. (Sep, 2010) Transcriptional enhancers in animal development and evolution. *Curr. Biol.,* **20**(17), R754–763.
6. Arnosti, D. N. and Kulkarni, M. M. (Apr, 2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?. *J Cell Biochem,* **94**(5), 890–8.
7. Swanson, C. I., Evans, N. C., and Barolo, S. (Mar, 2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell,* **18**(3), 359–70.
8. Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (Jun, 2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science,* **316**(5830), 1497–502.
9. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (May, 2007) High-resolution profiling of histone methylations in the human genome. *Cell,* **129**(4), 823–37.
10. Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (Aug, 2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature,* **448**(7153), 553–60.
11. Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (Feb, 2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature,* **457**(7231), 854–8.
12. Arnosti, D. N. and Kulkarni, M. M. (Apr, 2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?. *J. Cell. Biochem.,* **94**(5), 890–898.
13. Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (Jan, 2009) Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell,* **16**(1), 47–57.
14. Rouault, H., Mazouni, K., Couturier, L., Hakim, V., and Schweisguth, F. (Aug, 2010) Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proc Natl Acad Sci U S A,* **107**(33), 14615–20.
15. Clark, A., Eisen, M., Smith, D., Bergman, C., Oliver, B., Markow, T., Kaufman, T., Kellis, M., Gelbart, W., Iyer, V., et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature,* **450**(7167), 203–218.
16. Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., and Searle, S. M. J. (Jan, 2012) Ensembl 2012. *Nucleic Acids Res,* **40**, D84–90.
17. Wasserman, W. W. and Sandelin, A. (Apr, 2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.,* **5**(4), 276–287.
18. Stormo, G. and Fields, D. (1998) Specificity, free energy and information content in protein–DNA interactions. *Trends in biochemical sciences,* **23**(3), 109–113.
19. Su, J., Teichmann, S. A., and Down, T. A. (2010) Assessing computational methods of cis-regulatory module prediction. *PLoS Comput. Biol.,* **6**(12), e1001020.
20. Elnitski, L., Jin, V. X., Farnham, P. J., and Jones, S. J. (Dec, 2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.,* **16**(12), 1455–1464.
21. Aerts, S. (2012) Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Curr. Top. Dev. Biol.,* **98**, 121–145.
22. Kantorovitz, M., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G., Göttgens, B., Halfon, M., and Sinha, S. (2009) Motif-Blind, Genome-Wide Discovery of cis-Regulatory Modules in Drosophila and Mouse. *Developmental Cell,* **17**(4), 568–579.
23. Machanick, P. and Bailey, T. L. (Jun, 2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics,* **27**(12), 1696–1697.
24. Siddharthan, R., Siggia, E., and van Nimwegen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol,* **1**(7), e67.
25. Moses, A. M., Chiang, D. Y., Pollard, D. A., Iyer, V. N., and Eisen, M. B. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol,* **5**(12), R98.
26. Herrmann, C., Van de Sande, B., Potier, D., and Aerts, S. (Aug, 2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res,* **40**(15), e114.
27. Berman, B., Nibu, Y., Pfeiffer, B., Tomancak, P., Celniker, S., Levine, M., Rubin, G., and Eisen, M. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences,* **99**(2), 757.
28. Halfon, M., Grad, Y., Church, G., and Michelson, A. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome research,* **12**(7), 1019.
29. Rebeiz, M., Reeves, N., and Posakony, J. (2002) SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. *Proceedings of the National Academy of Sciences,* **99**(15), 9888.
30. Schroeder, M., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E., and Gaul, U. (2004) Transcriptional control in the segmentation gene network of Drosophila. *PLoS biology,* **2**, 1396–1410.
31. Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell,* **124**(1), 47–59.
32. Pierstorff, N., Bergman, C., and Wiehe, T. (2006) Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics,* **22**(23), 2858.
33. Nazina, A. and Papatsenko, D. (2003) Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency. *BMC bioinformatics,* **4**(1), 65.
34. Abnizova, I., te Boekhorst, R., Walter, K., and Gilks, W. (2005) Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test. *BMC bioinformatics,* **6**(1), 109.
35. Chan, B. and Kibler, D. (2005) Using hexamers to predict cis-regulatory motifs in Drosophila. *BMC bioinformatics,* **6**(1), 262.
36. Leung, G., Eisen, M., and Provart, N. (2009) Identifying Cis-Regulatory Sequences by Word Profile Similarity. *PLoS ONE,* **4**(9), e6901.
37. Brody, T., Yavatkar, A. S., Kuzin, A., Kundu, M., Tyson, L. J., Ross, J., Lin, T.-Y., Lee, C.-H., Awasaki, T., Lee, T., and Odenwald, W. F. (Jan, 2012) Use of a Drosophila genome-wide conserved sequence database to identify functionally related cis-regulatory enhancers. *Dev Dyn,* **241**(1), 169–89.
38. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. (Nov, 2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature,* **462**, 65–70.
39. Heger, A. and Ponting, C. P. (Nov, 2007) Variable strength of translational selection among 12 Drosophila species. *Genetics,* **177**, 1337–1348.
40. May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Afzal, V., Simpson, P. C., Rubin, E. M., Black, B. L., Bristow, J., Pennacchio, L. A., and Visel, A. (2011) Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.,* **44**, 89–93.
41. Sinha, S., van Nimwegen, E., and Siggia, E. D. (2003) A probabilistic

method to detect regulatory modules. *Bioinformatics,* **19 Suppl 1**, i292–301.
42. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol,* **17**(6), 368–76.
43. Halpern, A. L. and Bruno, W. J. (Jul, 1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol,* **15**(7), 910–7.
44. Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol,* **22**(2), 160–74.
45. Seplyarskiy, V. B., Kharchenko, P., Kondrashov, A. S., and Bazykin, G. A. (Aug, 2012) Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. *Mol. Biol. Evol.,* **29**(8), 1943–1955.
46. Bishop, C. et al. (2006) Pattern recognition and machine learning, Springer New York, .
47. Bao, Z. and Eddy, S. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research,* **12**(8), 1269–1276.
48. Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (Jan, 2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res,* **34**(Database issue), D108–10.
49. R Development Core Team R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna, Austria (2011) ISBN 3-900051-07-0.
50. Kimura, M. (Jun, 1962) On the probability of fixation of mutant genes in a population. *Genetics,* **47**, 713–9.
51. Hastie, T., Tibshirani, R., Friedman, J., et al. (2001) The elements of statistical learning: data mining, inference, and prediction, Springer New York, .
52. Kiyota, T., Kato, A., Altmann, C. R., and Kato, Y. (Mar, 2008) The POU homeobox protein Oct-1 regulates radial glia formation downstream of Notch signaling. *Dev Biol,* **315**(2), 579–92.
53. Neron, B., Menager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P., and Letondal, C. (Nov, 2009) Mobyle: a new full web bioinformatics framework. *Bioinformatics,* **25**(22), 3005–3011.
54. Xie, X., Lu, J., Kulbokas, E., Golub, T., Mootha, V., Lindblad-Toh, K., Lander, E., and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature,* **434**(7031), 338–345.
55. Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J., and Birney, E. (2005) The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biology,* **6**(12), R104.
56. Stark, A., Lin, M., Kheradpour, P., Pedersen, J., Parts, L., Carlson, J., Crosby, M., Rasmussen, M., Roy, S., Deoras, A., et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.. *Nature,* **450**(7167), 219.
57. Wang, T. and Stormo, G. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics,* **19**(18), 2369.
58. Grad, Y., Roth, F., Halfon, M., and Church, G. (2004) Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D. pseudoobscura. *Bioinformatics,* **20**(16), 2738.
59. Zhao, G., Schriefer, L., and Stormo, G. (2007) Identification of muscle-specific regulatory modules in Caenorhabditis elegans. *Genome research,* **17**(3), 348.
60. Busser, B. W., Taher, L., Kim, Y., Tansey, T., Bloom, M. J., Ovcharenko, I., and Michelson, A. M. (Mar, 2012) A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLoS Genet,* **8**(3), e1002531.
61. Menoret, D., Santolini, M., Fernandes, I., Spokony, R., Zanet, J., Gonzalez, I., Latapie, Y., Ferrer, P., Rouault., H., White, K., Besse, P., Hakim, V., Aerts, S., Payre, F., and Plaza, S. (2012) Genome-wide analyses of Shavenbaby target genes reveals distinct features of enhancer organization. *(Submitted),*.

FIGURES

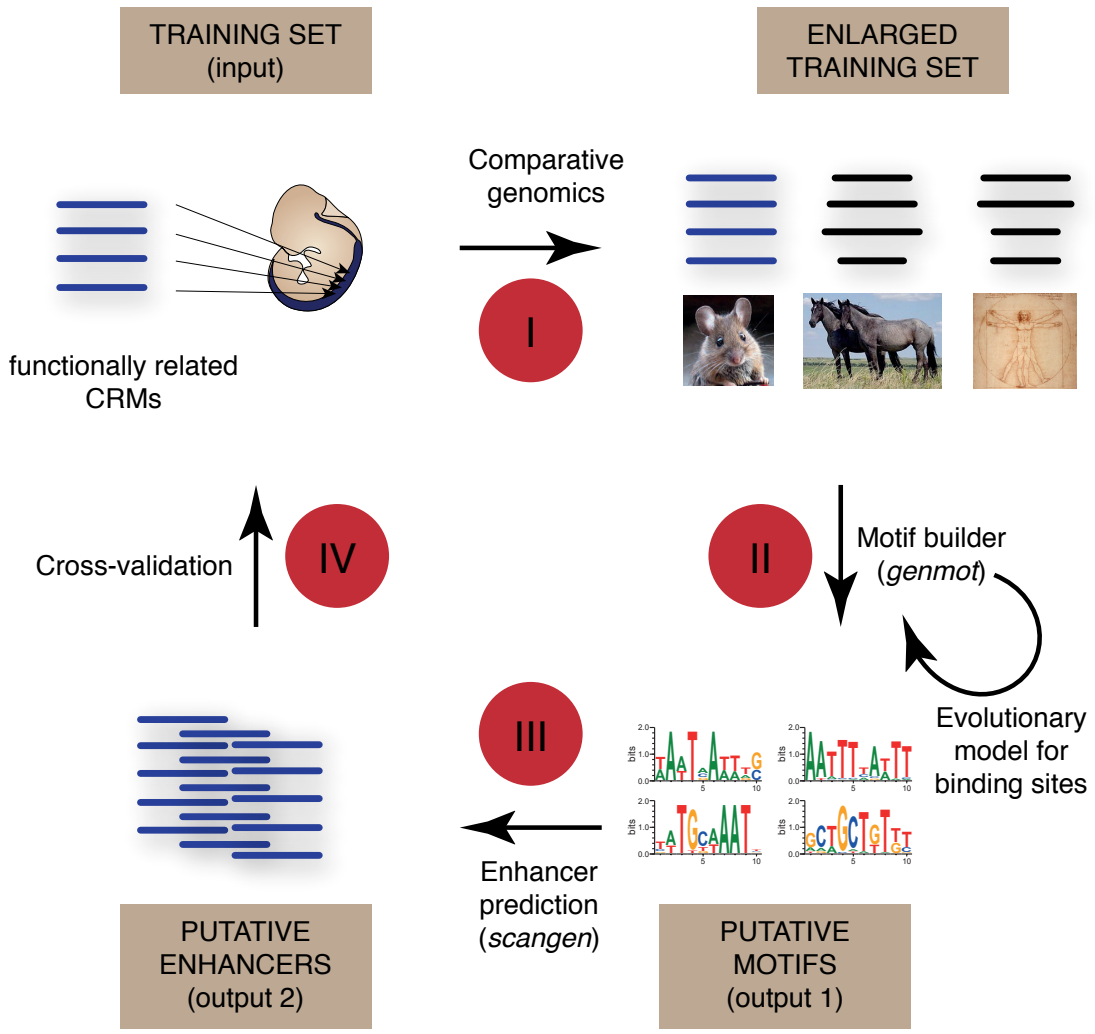
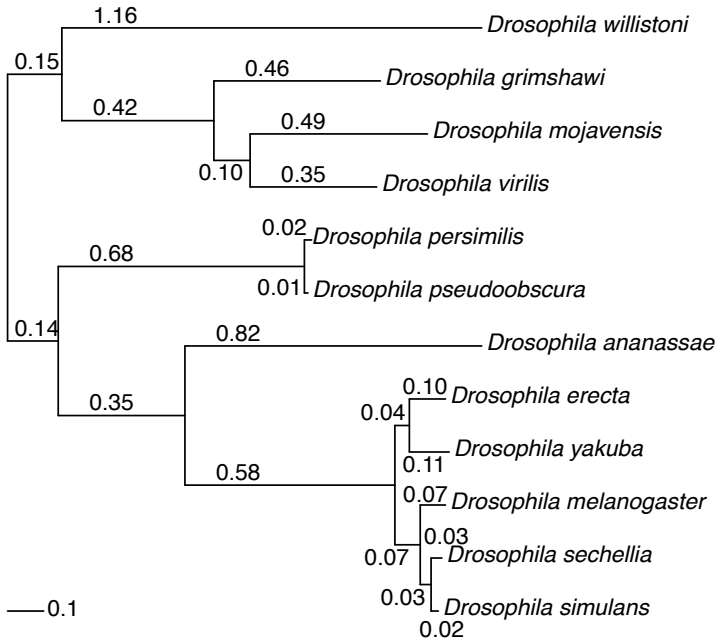


Figure 1. ***Imogene* workflow.** The algorithm takes as input a list of functionally related CRMs. Homologous sequences from closely related species are automatically retrieved (I) and scanned in order to generate a list of putative transcription factor motifs (II). These motifs fuel the last step consisting in the inference of related novel CRMs (III). These predicted CRMs can finally be compared to a set of test CRMs to evaluate the predictability power of the whole procedure (IV).

A **Drosophilae**



B **Eutherian**

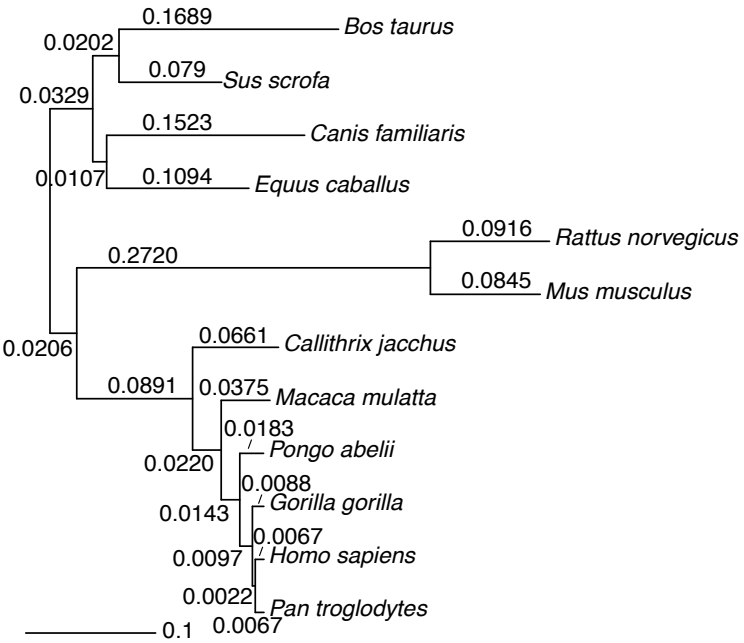


Figure 2. **Phylogenetic trees and phylogenetic distances used by *Imogene*.** The branch lengths represent the evolutionary distances d used by the evolutionary models at the motif construction stage.

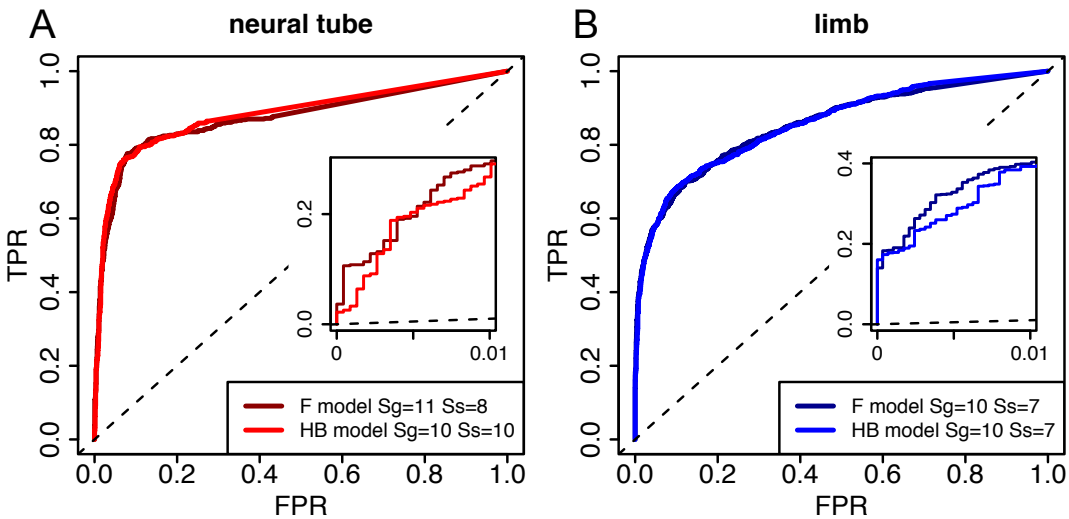


Figure 3. Analysis of well characterized developmental processes. We tested the algorithm on mammal CRMs driving expression at E11.5 in neural tube (A) and limb (B). For each class, CRMs were divided into a training set and a test set. Motifs were learned on the training set and used to score CRMs from the test set along with background regions consisting of the CRMs 1kb flanking sequences (see *Methods*). The displayed ROC curves show the proportion of test set CRMs recovered above a given score (True Positive Rate denoted by TPR) *vs.* the proportion of recovered background sequence at the same score for the Felsenstein (F) and Halpern-Bruno (HB) models. The shown ROC plots are the results of 40 trials. The FPR $\leq 1\%$ region of each curve is replotted in the insets for better visibility. For each test set and each evolutionary model, the thresholds S_g and S_s used for motifs generation and sequences scanning are given in the figures. Black dashed lines show random discrimination.

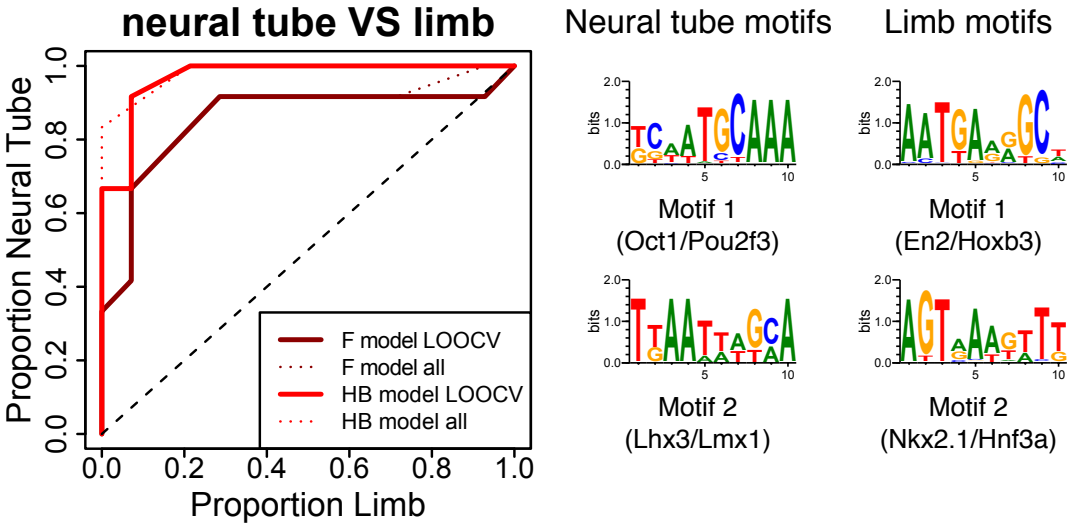


Figure 4. Pattern recognition (mammals). ROC plots showing the discrimination between limb and neural CRMs using a simple linear classifier. Neural and limb classes are compared to each other. Thick lines correspond to a leave-one-out cross-validation (LOOCV) scheme with a score function based on the *de novo* generated motifs from *Imogene*. The results obtained with the two evolutionary models are shown (Felsenstein model (F) solid dark red line, with threshold parameters $S_g = 11$, $S_s = 9$, and Halpern-Bruno (HB) model, solid light red line, with threshold parameters $S_g = 11$, $S_s = 8$). The analogous discrimination curves based on learning motifs on the whole training set (with the same threshold parameters) are shown for comparison (colored dashed lines). With this latter procedure, the discrimination is improved but still comparable to that computed by the LOOV, indicative of no strong overfitting of the training set. The corresponding discriminative motifs are shown for the whole training set learning with HB model (similar motifs are obtained with the F model). Black dashed line show random discrimination.

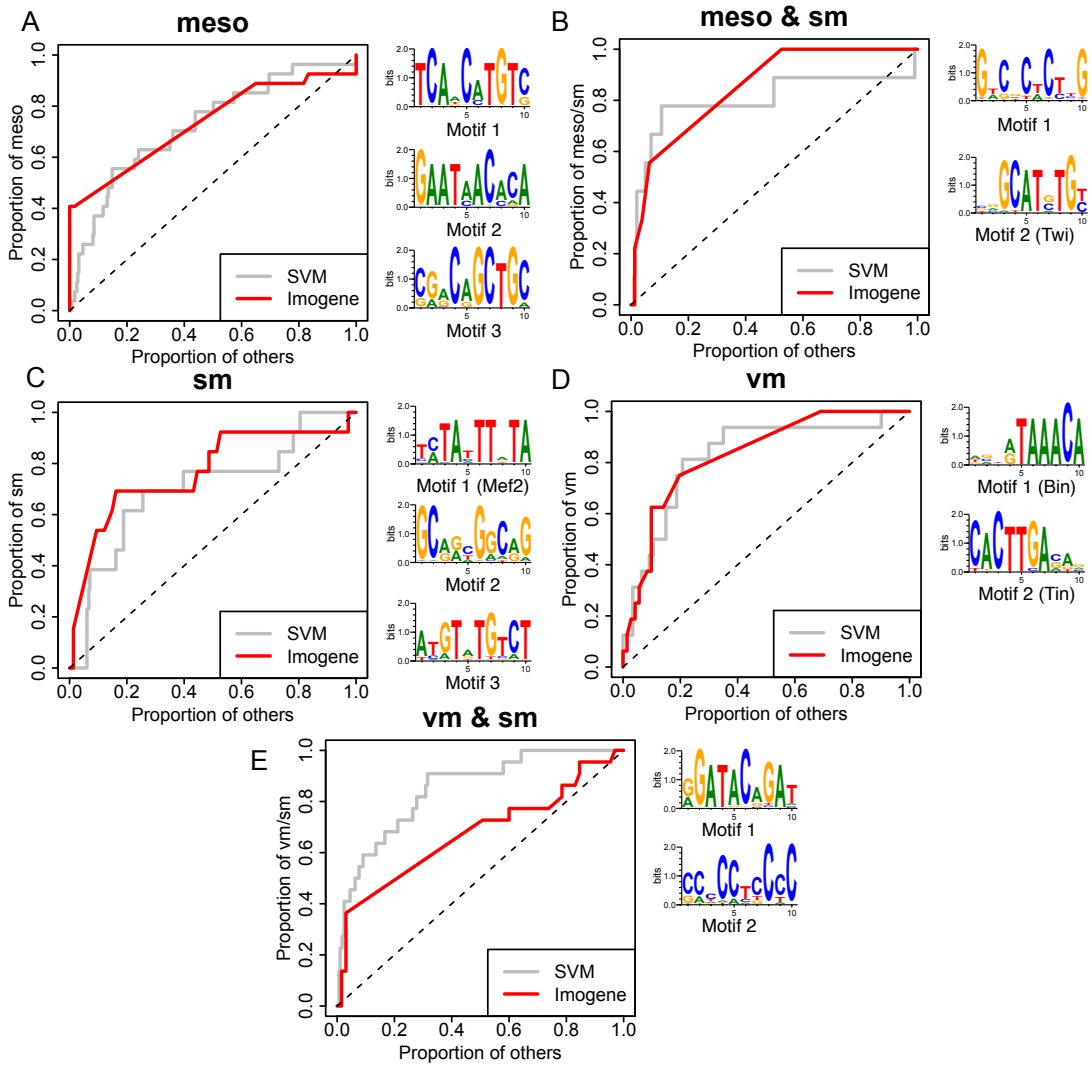


Figure 5. Pattern recognition (Drosophilae). Recognition of classes of CRMs expressed in 5 tissue types: mesoderm (meso), somatic muscle (sm), visceral muscle (vm) , mesoderm and somatic muscle (meso & sm) and visceral and somatic muscle (vm & sm). ROC plots are obtained using a leave-one-out cross-validation scheme. Two classifiers are compared: a Support Vector Machine using 15 ChIPseq peak heights (grey, replotted using the data and the program provided in ref. (38)), and *Imogene* using the *de novo* generated motifs with Felsenstein evolutionary model (red) and a simple linear classifier (see *Methods*). The following thresholds were used: meso ($S_g = 12$, $S_s = 12$), meso & sm ($S_g = 10$, $S_s = 10$), sm ($S_g = 9$, $S_s = 4$), vm ($S_g = 10$, $S_s = 10$), vm & sm ($S_g = 11$, $S_s = 8$).

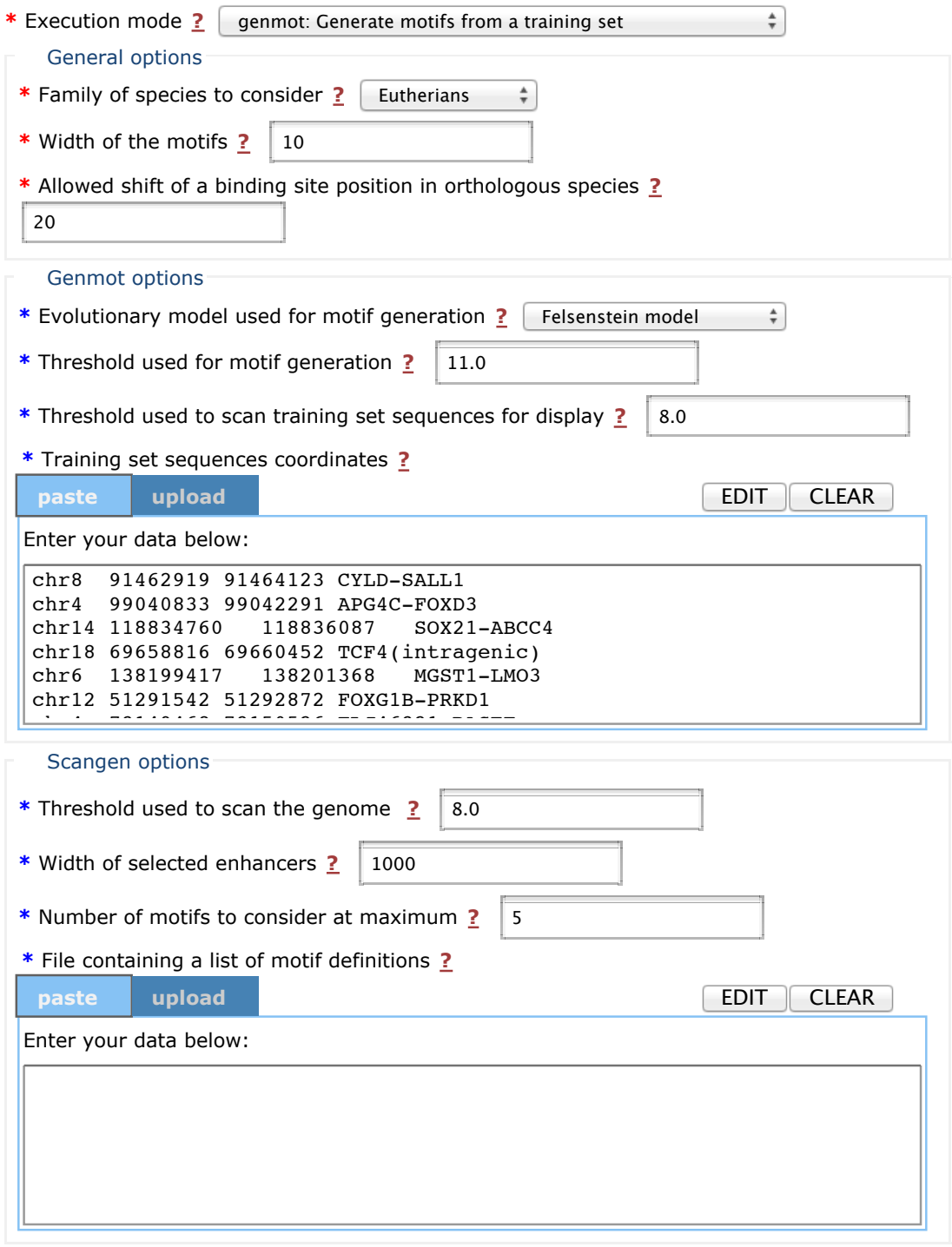
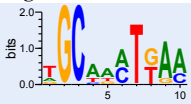
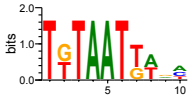


Figure 6. Web based interface : input web page. A copy input web page for *Imogene* powered by the mobyle bioinformatics framework is shown.

A

Motifs

Color	Rank	Logo	P-value (log10)	Over-representation
	1		-79.4772	55.122
	2		-76.2578	38.1757

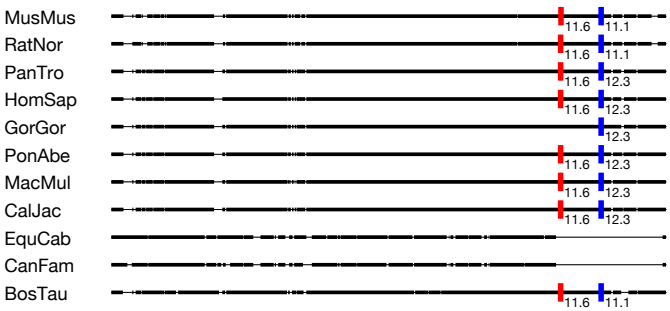
Motifs instances in the training set

>MusMus MRPS9(intragenic)_1_42945168_42946091 1 42945168 42946091

```
CAACTTGTTA  CACGGATGGG  TTGCACGCAG  CGAAGCTGTG  GAAAATCTGT  GCCTTTTAAC
TTTTCTACTT  AATCACGGTT  GTAGCATTGC  CTTTAGACTG  TATGCTACAT  TAATTCTCTT
CCTGCCTTCT  GCCTTCATCC  CAAGTTTCAC  GGGAAAAAGT  AAAGTGTGCA  GGTCTTACAG
AGGAGCCTTA  TCAAACAGCT  GTCATCTGAC  AAGCCATTTG  CATTTGTTTT  GGCTGAAATG
GAGCAACCCA  AGGGCAAGAT  CTTTTGTTGC  ATTCCATCAT  AATGAAGAAA  TTACACATTG
TGTAAGAGGC  CTGGCTTTAT  TTTTAGTTTG  CTTGTGTGCT  TTAAAAGGTA  TTGCTCCAGA
AACTGATGGG  ATAGAATTTT  ACCG
```

Motifs presence in alignments

MRPS9(intragenic)_1_42945168_42946091



B

Score	Coordinate	Closest TSS	Relative distance to closest TSS (bp)	5 surrounding TSSs
48.1146	chr15:81014639-81015638	Mkl1	7048	Sgsm3;Mkl1;Mkl1;4930483J18Rik;Mchr1;
34.2492	chr3:143836754-143837753	Lmo4	29042	A830019L24Rik;Gm6260;Lmo4;Lmo4;Lmo4;
34.2492	chr12:51291776-51292775	Prkd1	458934	Foxg1;3110039M20Rik;Prkd1;G2e3;Scfd1;
33.8818	chr14:23564465-23565464	Gm10248	349828	Zfp503;1700112E06Rik;Gm10248;Kcnma1;Dlg5;
30.9743	chr2:63807707-63808706	Fign	128862	Gca;Kcnh7;Fign;Grb14;Cobll1;

Figure 7. **Web based interface : output web page.** Example of an output web page for *Imogene* powered by the mobyle bioinformatics framework. A. Result page for the *Genmot* mode. Two motifs were generated from the neural tube full training set (default is 5), using the same parameters as in Figure 3. Results are shown for the training set sequence MRPS9(intragenic). For display purposes, the beginning of the sequence, which contains no instances for the motifs, was cut in the middle panel. In the alignments, thick lines correspond to sequences and thin lines to gaps. B. Result page for the *Scangen* mode. The two generated motifs were used to score putative regulatory sequences of 1kb in the mouse genome at optimal threshold $S_s = 10$. The 5 best ranking sequences are shown (default is 200).

Imogene: identification of motifs and cis-regulatory modules underlying gene co-regulation

Hervé Rouault, Marc Santolini , François Schweisguth, Vincent Hakim

July 15, 2013

Supplementary Figures

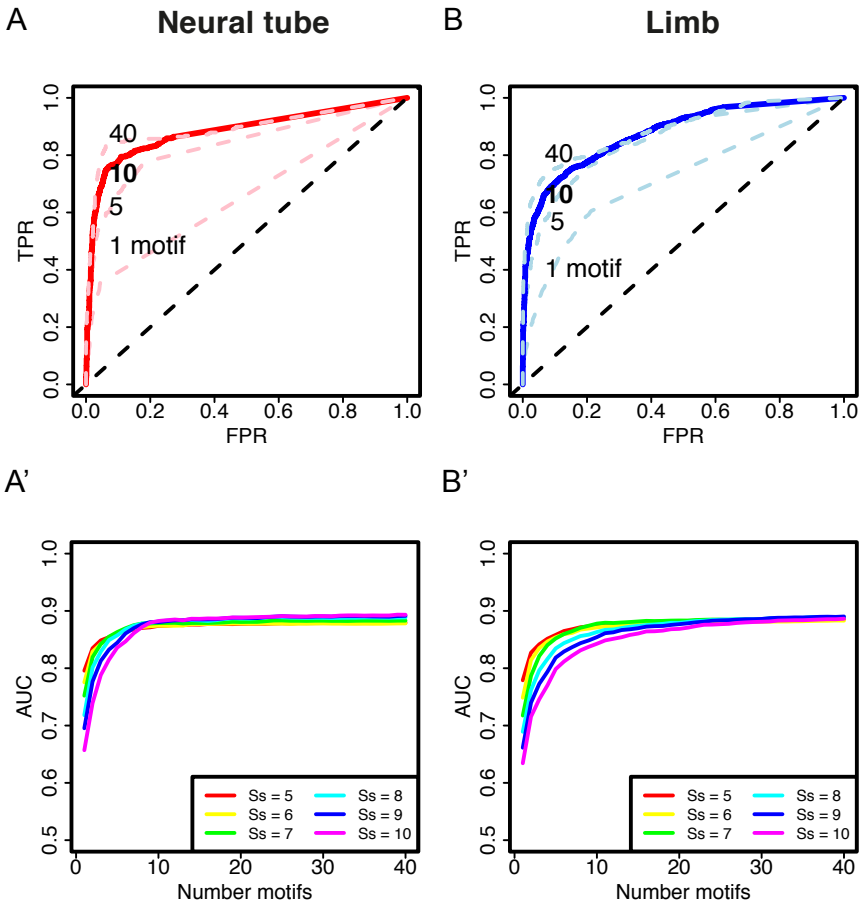


Figure S1. Dependence of the predictions on the number of scoring motifs ROC plots obtained at optimal scanning threshold using the Halpern-Bruno evolutionary model are shown for the neural tube (A) and limb (B) cases. Different curves are shown corresponding to sequences scored with different number of motifs: 1, 5 and 40 (light-color dashed lines), 10 (thick line). The ROC curves obtained for 10 motifs correspond to the ones shown in Fig. 3. To assess the degree of convergence, we computed the Area Under ROC Curve as a function of the number of motifs used (A',B',C'). We show the curves corresponding to the choice of different scanning thresholds S_s. In all cases, 10 motifs were sufficient for the AUC to reach convergence. The optimal S_s was chosen as the one maximizing the AUC for 10 motifs.

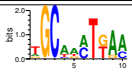
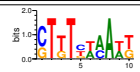
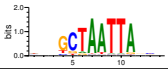
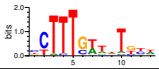
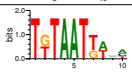
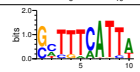
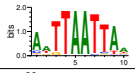
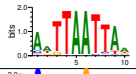
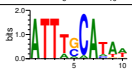
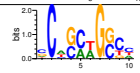
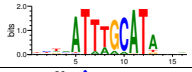
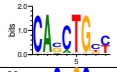
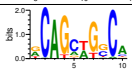
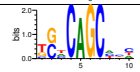
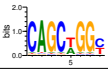
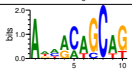
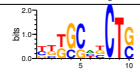
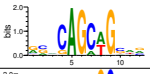
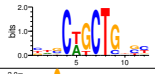
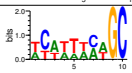
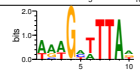
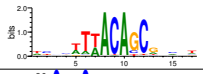
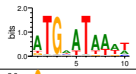
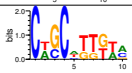
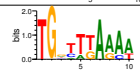
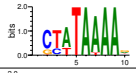
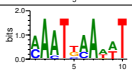
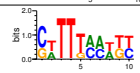
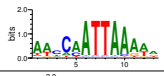
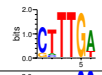
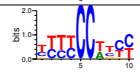
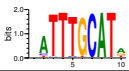
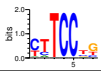
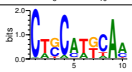
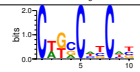
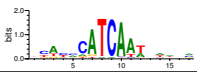
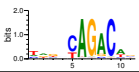
NEURAL		LIMB	
	Motif 1		Motif 1
	V$CHX10_01		V$TCF3_01
	Motif 2		Motif 2
	V$LHX3_01		V$LHX3_01
	Motif 3		Motif 3
	V$OCT2_01		V$MYOGENIN_Q6
	Motif 4		Motif 4
	V$HEB_Q6		V$CBF1_QX
	Motif 5		Motif 5
	V$NEUROD_02		V$NEUROD_02
	Motif 6		Motif 6
	V$RHOX11_01		V$POU1F1_Q6
	Motif 7		Motif 7
	V$CBF1_QX		V$TATA_C
	Motif 8		Motif 8
	V$NKX61_01		V$LEF1_Q2
	Motif 9		Motif 9
	V$OCT1_B		V$ETS2_Q6
	Motif 10		Motif 10
	V$PBX1_04		V$SMAD_Q6_01

Figure S2. Motifs learnt on the full training sets. The 10 best ranking motifs generated on the CRMs training sets are shown together with the closest Transfac motifs (see *Distance between motifs* in *Methods* for details of motif distance computation).

	Mot1	Mot2	Mot3	Mot4	Mot5	Mot6	Mot7	Mot8	Mot9	Mot10
ZIC4-ZIC1_9_91261697_91263041	2	0	1	0	0	0	2	1	1	0
TCF4(intragenic)_18_69658816_69660452	0	0	0	1	1	2	0	0	0	0
CEI-IRX1_13_72435297_72436784	3	4	2	2	0	3	0	1	3	2
NBEA(intragenic)_3_55768657_55770664	0	1	1	2	1	0	0	0	3	1
AKT3(intragenic)_1_179080168_179081586	2	1	1	3	2	1	2	0	2	0
FOXG1B-PRKD1_12_51291542_51292872	4	2	1	0	2	0	0	3	1	0
DACH1(intragenic)_14_98553917_98556433	5	0	2	4	0	2	3	1	3	2
FAM44A-CPEB2_5_42914188_42915270	1	0	1	3	1	1	2	0	1	0
IRX4-IRX2_13_73170587_73173631	0	0	0	2	0	0	0	0	4	0
EBF1(intragenic)_11_44469978_44471372	2	3	1	1	0	3	4	1	0	0
ATG4C-FOXD3_4_99240573_99241457	0	0	0	0	0	0	0	0	0	0
CYLD-SALL1_8_91462919_91464123	0	0	1	1	1	0	0	0	1	0
POU2F1(intragenic)_1_167864366_167866439	5	0	4	1	0	0	0	3	0	3
APG4C-FOXD3_4_99040833_99042291	0	0	0	0	0	0	0	0	0	0
MGC14798-HH114_2_115363420_115365044	2	2	2	0	1	0	0	2	0	0
MGST1-LMO3_6_138199417_138201368	5	1	1	1	1	3	0	3	1	1
APG4C-FOXD3_4_98961102_98962673	2	2	2	2	3	1	4	0	0	0
FLJ46321-RASEF_4_73149468_73150526	0	0	2	4	0	1	1	1	0	1
TCF12(intragenic)_9_71823775_71824538	1	0	0	1	2	0	1	0	0	1
BMPER(intragenic)_9_23182371_23184296	2	1	1	1	0	2	1	0	1	0
SOX21-ABCC4_14_118834760_118836087	1	6	2	1	3	0	1	3	3	2
FANCL-BCL11A_11_25256346_25257683	0	2	1	0	3	0	2	0	0	0
DERA(intragenic)_6_137772070_137773298	1	5	0	1	1	1	2	0	0	0
MRPS9(intragenic)_1_42945168_42946091	1	1	0	2	1	1	1	0	0	0
YTHDF3-BHLHB5_3_16776170_16778776	2	2	0	1	0	0	0	0	1	0
STXBP6-NOVA1_12_47121350_47122759	1	4	2	3	0	2	2	0	0	0
IDH3B-CPXM1_2_130177541_130178125	0	0	0	0	0	0	0	0	0	0
LOC347487-SOX3_X_57972482_57973750	3	0	1	2	1	1	2	2	1	3

Figure S3. Neural CRMs and motifs. List of the neural CRMs used in this study. The number of motifs of different types on each CRM is given for the 10 best-ranking neural motifs shown in Figure S2.

	Mot1	Mot2	Mot3	Mot4	Mot5	Mot6	Mot7	Mot8	Mot9	Mot10
hs1435_7_106105018_106107143	1	1	2	2	0	3	3	0	3	0
hs126_14_97485454_97486724	5	1	2	0	0	2	1	3	1	0
hs1477_2_59400401_59401189	2	0	1	1	2	1	1	1	1	0
hs521_1_91610325_91611486	0	1	2	4	0	1	0	0	8	0
mm422_2_4477190_4478921	0	0	1	1	0	0	0	0	0	0
hs1432_13_91326599_91329775	0	0	0	1	0	0	0	0	0	0
hs1433_3_30003454_30008202	8	4	8	5	5	5	4	5	6	1
hs208_9_100171947_100173392	2	2	3	3	5	1	1	1	4	2
hs1507_1_75765578_75770167	1	0	5	4	3	0	1	0	7	1
hs774_3_5329674_5330756	4	2	1	0	0	2	0	2	0	0
hs919_15_50496379_50498196	3	1	1	0	2	1	1	1	2	3
hs326_19_45568075_45569359	1	0	4	1	2	3	3	0	0	1
hs72_8_91978407_91979282	1	1	2	3	2	2	0	0	2	1
hs1484_4_97888231_97891318	0	1	0	0	2	0	0	1	1	0
mm423_2_4508631_4509808	0	0	1	0	0	0	0	0	0	0
mm428_5_38308981_38309833	0	2	1	0	0	0	1	0	4	0
hs741_3_66874217_66875516	4	2	1	0	1	2	0	2	1	0
hs1148_12_119941220_119942766	0	0	1	0	0	0	0	0	0	0
hs1109_13_79503055_79504129	2	1	1	1	0	1	1	0	1	0
hs2041_9_96280544_96283360	2	0	0	0	0	0	1	0	0	0
hs1473_13_56260379_56262548	1	1	7	1	8	0	0	0	1	0
hs1434_14_23833434_23842485	1	1	7	5	3	0	4	2	4	3
hs1465_6_51144711_51148222	0	2	6	3	1	1	1	0	3	0
mm94_6_122342623_122346341	0	0	2	1	1	0	0	0	3	0
hs1452_10_45612931_45614502	0	0	0	0	0	0	2	0	1	2
hs1468_10_125358093_125366026	0	0	1	0	0	1	0	0	0	0
hs1586_13_15640807_15642666	0	1	1	1	1	2	0	0	3	0
hs1273_12_9344323_9346407	2	2	2	1	3	4	1	4	3	1
hs1278_2_137073444_137074711	1	1	5	3	0	0	0	1	0	1
hs1500_14_22281464_22282917	0	0	4	1	2	0	0	0	2	0
mm458_15_63025492_63026343	2	0	1	0	0	1	0	0	4	0
hs388_12_26576441_26577229	4	4	2	2	1	0	0	0	0	1
hs1491_14_25804749_25806653	1	0	6	0	6	0	0	0	3	3
hs1428_3_99469238_99471067	0	2	4	2	2	0	1	0	3	0
hs1430_6_52917020_52919645	5	1	4	1	2	1	1	0	2	0
hs1475_16_72685882_72688547	0	0	1	0	1	0	1	4	0	1
hs1448_2_171555881_171562133	1	3	5	1	0	0	1	0	1	0
hs644_12_34884495_34885741	0	5	4	1	0	1	1	0	2	0

Figure S4. Limb CRMs and motifs. List of the limb CRMs used in this study. The number of motifs of different types on each CRM is given for the 10 best-ranking limb motifs shown in Figure S2.

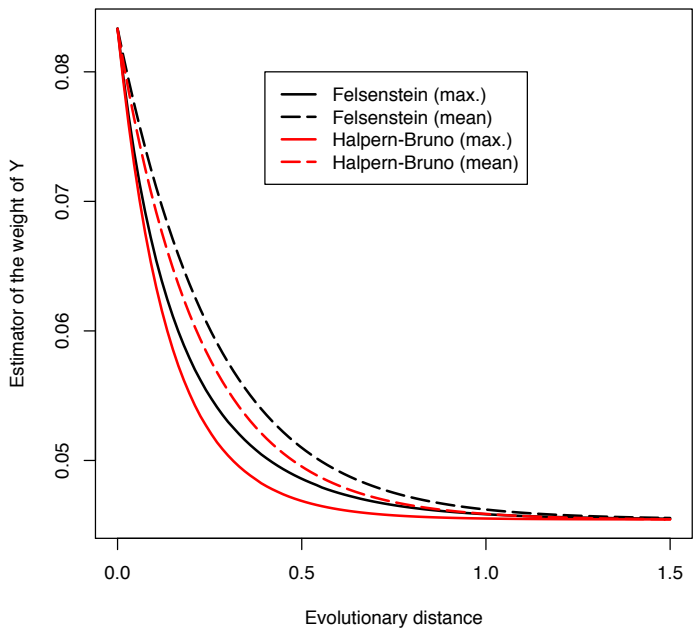


Figure S5. Simple example of motif inference with Felsenstein and Halpern-Bruno evolutionary models The inference of an ancestral base is compared in the simple case of two species at a phylogenic distance d from their common ancestor, for a two nucleotide alphabet,X and Y. The mean and maximum likelihood estimate of observing Y in the common ancestor given that the two species share an X is shown as a function of evolutionary distance d, for the Felsenstein or Halpern-Bruno evolutionary models. The likelihood is always smaller with the Halpern-Bruno model, reflecting the model greater evolutionary rate.

3.3 Calcul de la moyenne de la postérieure par une méthode MCMC

Nous avons voulu savoir si l'approximation réalisée par Imogene lors du calcul du maximum de l'estimateur de la postérieure modifiée (obtenue avec un prior de Dirichlet de paramètres $\alpha_b + 1$) donnait un résultat effectivement proche du calcul de la moyenne de l'estimateur de la postérieure non modifiée (avec un prior de Dirichlet de paramètres α_b) :

$$\underset{w_i}{\operatorname{argmax}}\, \mathcal{L}(\{\mathcal{A}\}|w_i)\mathrm{Dir}(w_i, \alpha_b + 1) \simeq \int w_i \mathcal{L}(\{\mathcal{A}\}|w_i)\mathrm{Dir}(w_i, \alpha_b)\mathrm{d}w_i \tag{3.28}$$

où w_i est le vecteur de poids de la PWM à la position i, $\{\mathcal{A}\}$ est l'ensemble des alignements de nucléotides observés à cette position dans les sites de fixation, $\mathcal{L}(\{\mathcal{A}\}|w_i)$ dénote la vraisemblance de l'alignement connaissant le vecteur de poids w_i et $\mathrm{Dir}(w_i, \alpha_b)$ est le prior de Dirichlet de paramètres α_b. Pour des distances phylogénétiques nulles ou infinies, cette approximation est valide, et pour des distances intermédiaires, elle semble fonctionner dans le cas simplifié de l'arbre en étoile de la figure S5 de l'article. Qu'en est-t-il dans le cas réèl avec un arbre à la topologie plus complexe ? L'estimation directe de la moyenne de cette distribution est difficile, puisque nous n'avons pas de moyen simple de l'échantillonner. Afin de contourner ce problème, nous avons eu recours à une méthode de Monte-Carlo par chaînes de Markov ou MCMC basée sur l'algorithme de Metropolis-Hastings (Krauth, 2006). Nous présentons la comparaison de la moyenne de la postérieure obtenue par l'approche MCMC et du maximum de la postérieure modifiée par descente de gradient sur un cas réel simple en utilisant l'arbre des vertébrés (figure 2 de l'article) et le modèle d'évolution *Felsenstein*.

3.3.1 Principe de l'algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) permet d'échantillonner une distribution donnée en utilisant le parcours d'une chaîne de Markov ayant cette distribution pour loi stationnaire. Un tel processus de Markov est défini par des probabilités de transition $P(w \to w')$ entre deux états w et w'. Il converge vers une distribution stationnaire $\pi(w)$ unique sous deux conditions : (1) les transitions sont réversibles et le processus satisfait le bilan détaillé $\pi(w)P(w \to w') = \pi(w')P(w' \to w)$, (2) le processus est ergodique, c'est-à-dire que tout état est et reste accessible. L'algorithme de Metropolis-Hastings repose sur la construction d'une chaîne de Markov ayant ces propriétés et dont la distribution d'équilibre $\pi(w)$ est la probabilité que l'on cherche à échantillonner $P(w)$. Pour cela, on part de l'équation du bilan détaillé, que l'on peut écrire

$$\frac{P(w \to w')}{P(w' \to w)} = \frac{P(w')}{P(w)} \tag{3.29}$$

La transition $P(w \to w')$ est ensuite décomposée en deux sous-étapes, la proposition (*proposal*) et l'acceptation (*acceptance*) :

$$P(w \to w') = \underbrace{g(w \to w')}_{\text{proposition}} \cdot \underbrace{A(w \to w')}_{\text{acceptation}} \tag{3.30}$$

En insérant dans l'éq. 3.29 on obtient

$$\frac{A(w \to w')}{A(w' \to w)} = \frac{P(w')}{g(w \to w')} \frac{g(w' \to w)}{P(w)} \tag{3.31}$$

Plusieurs choix de la fonction d'acceptation sont possibles pour satisfaire cette équation (Hastings, 1970). Un choix courant, dit choix de Metropolis, est :

$$A(w \to w') = \min\left(1, \frac{P(w')}{g(w \to w')} \frac{g(w' \to w)}{P(w)}\right) \tag{3.32}$$

On remarque que cette quantité est invariante sous multiplication de la distribution $P(w)$ par un facteur non nul. Autrement dit, la distribution n'a pas besoin d'être normalisée. Dans un cadre bayésien, cela veut dire que l'on peut remplacer la postérieure par le produit de la vraisemblance et du *prior*.

La méthode de Metropolis-Hastings se résume donc ainsi :

1. Initialiser w à une certaine valeur.
2. Choisir un nouvel état w' tiré selon $g(w \to w')$
3. Accepter l'état avec une probabilité donnée par $A(w \to w')$. Si le nouvel état n'est pas accepté, alors $w' = w$.
4. Itérer jusqu'à convergence

Au final, w étant tiré selon la distribution $P(w)$, sa moyenne est estimée en sommant les poids $w(t)$ obtenus au cours des N itérations réalisées :

$$\langle w \rangle \simeq \hat{w}_N = \frac{1}{N} \sum_{t=1}^{N} w(t) \tag{3.33}$$

Quant au critère de convergence, une possibilité est d'utiliser le Théorème Central Limite (TCL). Celui-ci stipule que la moyenne de n variables aléatoires indépendantes et identiquement distribuées selon une loi de moyenne μ et d'écart-type σ tous deux de valeurs finies suit, pour n grand, une loi normale de moyenne μ et d'écart-type $\sigma/\sqrt{n}$. Dans l'approche MCMC, les échantillons successifs w_i ne sont pas indépendants à cause du fait qu'on les tire selon la loi $g(w \to w')$. Il faut donc calculer le temps de décorrélation T pour lequel $\langle w(t) w_i(t+T) \rangle \simeq 0$, puis utiliser les échantillons $w(t)$ obtenus toutes les T itérations comme variables indépendantes. L'application du TCL permet alors d'arrêter les itérations lorsqu'une certaine précision désirée est atteinte, par exemple lorsque $\sigma/\sqrt{n} < 0.05 \cdot \mu$, c'est-à-dire lorsque les variations de l'estimateur de la moyenne sont de l'ordre de 5% de la valeur de la moyenne. L'écart-type σ étant lui même estimé à partir des échantillons de l'algorithme, il faut aussi s'assurer qu'il a convergé vers une valeur stable pour appliquer le TCL.

3.3.2 Application au calcul de la postérieure

Dans notre cas, nous souhaitons utiliser l'algorithme de Metropolis-Hastings pour calculer la valeur moyenne du vecteur de poids w_i en position i de la PWM selon la distribution postérieure $\mathcal{P}(w_i|\{\mathcal{A}\})$:

$$\langle w_i \rangle = \int w_i \mathcal{P}(w_i|\{\mathcal{A}\}) \mathrm{d}w \tag{3.34}$$

Il nous faut pour cela définir une loi de proposition $g(w_i \to w_i')$ pertinente. Dans notre cas, les poids w_i doivent rester dans le simplexe de dimension 3 défini par $w_A, w_C, w_G > 0$ et $w_A + w_C + w_G < 1$, le poids w_T étant entièrement déterminé par $w_T = 1 - w_A - w_C - w_G$. La

distribution naturelle possédant cette propriété est la loi de Dirichlet $\mathrm{Dir}(\alpha)$, de paramètres $\alpha = \{\alpha_A, \alpha_C, \alpha_G, \alpha_T\}$ et de densité de probabilité

$$f(w) = \frac{1}{B(\alpha)} \prod_{b \in \{A,C,G,T\}} w_{i,b}^{\alpha_b - 1} \tag{3.35}$$

où $B(\alpha)$ est la fonction bêta multinomiale permettant la normalisation. Cette distribution est la même que celle obtenue dans le cas d'observations indépendantes (cf article). Afin d'accélérer l'échantillonnage MCMC, nous avons cherché à régler les paramètres α de manière à être au plus proche de la distribution $\mathcal{P}(w_i|\{\mathcal{A}\})$. Dans le cas de N sites indépendants, celle-ci suit une loi de Dirichlet de paramètres $\alpha_p + N_i$, où α_p est le vecteur de pseudo-counts et N_i le vecteur donnant les nombres d'observations des nucléotides en position i au sein des différentes séquences. Dans le cas d'un arbre phylogénétique corrélant les séquences, le nombre *effectif* d'observation est moins grand que le nombre total de séquences. Nous avons donc défini les paramètres de notre proposition comme étant

$$\alpha_b = \alpha_p + N_{\text{eff}} \cdot w_i \tag{3.36}$$

où $N_{\text{eff}} = N_{\text{sites}} \cdot N_{\text{spe}}/2$ avec N_{sites} le nombre d'alignements observés et N_{spe} le nombre d'espèces dans l'alignement (12 dans les deux cas, Drosophile et mammifères). Grossièrement, cela revient à dire que le modèle d'évolution réduit d'un facteur 2 le nombre de séquences indépendantes. Nous avons calculé que le taux d'acceptation (proportion de mouvements proposés qui sont acceptés) pour ce paramètre était de l'ordre de 50%, une valeur généralement considérée comme raisonnable (Krauth, 2006). Nous obtenons finalement l'expression pour la proposition :

$$g(w \to w') = \frac{1}{B(\alpha)} \prod_{b \in \{A,C,G,T\}} (w'_{i,b})^{\alpha_{p,b} + N_{\text{eff}} \cdot w_{i,b} - 1} \tag{3.37}$$

Le vecteur w_i est initialisé à la valeur qu'il prendrait si toutes les séquences orthologues étaient des observations indépendantes (cf article) :

$$w_{i,b}(0) = w_{i,b}^{\text{inde}} = \frac{N_{i,b} + \alpha_b}{N_{\text{tot}} + \sum_b \alpha_b} \tag{3.38}$$

Le poids $w_i(1)$ suivant est tiré selon la probabilité de transition $g(w_i(0) \to w_i(1))$. Les différentes quantités de l'équation 3.32 sont ensuite calculées et la transition est acceptée avec probabilité $A\,(w_i(0) \to w_i(1))$.

3.3.3 Illustration sur un exemple

Étudions maintenant un exemple concret. Nous présentons la méthode sur le cas présenté en fig. 3.2A et nous utilisons le modèle d'évolution *Felsenstein*. Le vecteur de poids w_i est initialisé au cas indépendant w_i^{inde}. La proposition $g(w_i^{\text{inde}} \to w)$ (éq. 3.37) est montrée en fig. 3.2B. On voit notamment comment la distribution de Dirichlet permet de rester dans le simplexe dans les cas w_A et w_T proches de 0. La valeur finale de l'estimation de la moyenne de la postérieure obtenue après convergence de la chaîne (voir ci-dessous) est aussi montrée : elle est relativement proche de la moyenne de la distribution, indiquant que le choix de la valeur initiale est effectivement judicieux.

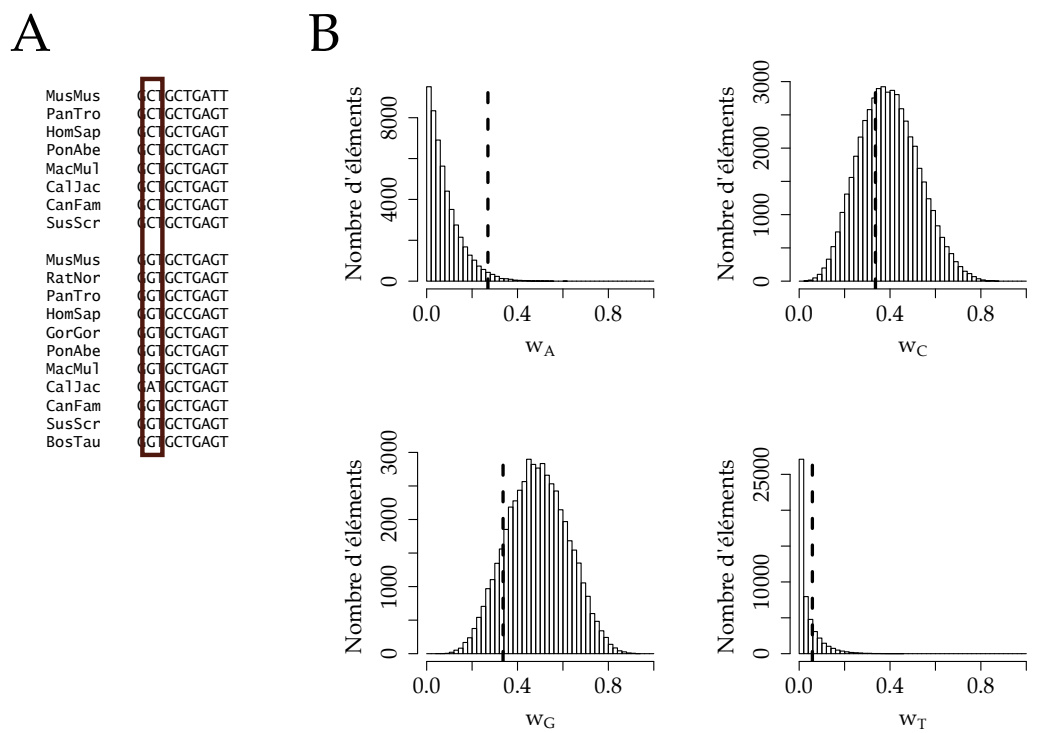


FIGURE 3.2 – **Conditions initiales.**
(A) Sites utilisés pour le MCMC. On s'intéresse ici à la 2ème position, contenant au total 10 G, 8 C et 1 A.
(B) Distribution de la distribution de Dirichlet correspondant à la proposition $g(w_i^{\text{inde}} \to w)$. L'échantillonnage consiste en 50,000 tirages aléatoires, et le bin est de 0.02. Les lignes verticales en pointillés indiquent l'estimation de la moyenne obtenue après convergence (voir fig. 3.4).

La chaîne MCMC est ensuite lancée. Le taux d'acceptation est calculé comme valant 62%. Les 500 premiers échantillons de w_i sont montrés en figure 3.3A. On note que certains points sont corrélés : diminutions ou augmentations successives de la valeur courante $w_i(t)$ sur plusieurs itérations. Pour quantifier cet effet, nous avons mesuré la corrélation temporelle des échantillons. Celle-ci est donnée par

$$C_b(\tau) = \frac{1}{N}\sum_{t=1}^{N-\tau} w_{i,b}(t)w_{i,b}(t+\tau) - \left(\frac{1}{N}\sum_{t=1}^{N} w_{i,b}(t)\right)^2 \tag{3.39}$$

avec dans ce cas $N = 50,000$. Le logarithme de cette quantité est montrée en figure 3.3B. L'intérêt du logarithme est de mettre en exergue le caractère exponentiel de la décorrélation :

$$C_b(\tau) \propto e^{-\tau/T_b} \tag{3.40}$$

Les temps de décorrelation $T_b \sim 10$ sont estimés en ajustant une droite sur les 20 premiers points de la courbe. Maintenant que l'on connait le temps de corrélation entre deux échantillons, il est possible d'obtenir des échantillons indépendants en les choisissant à des

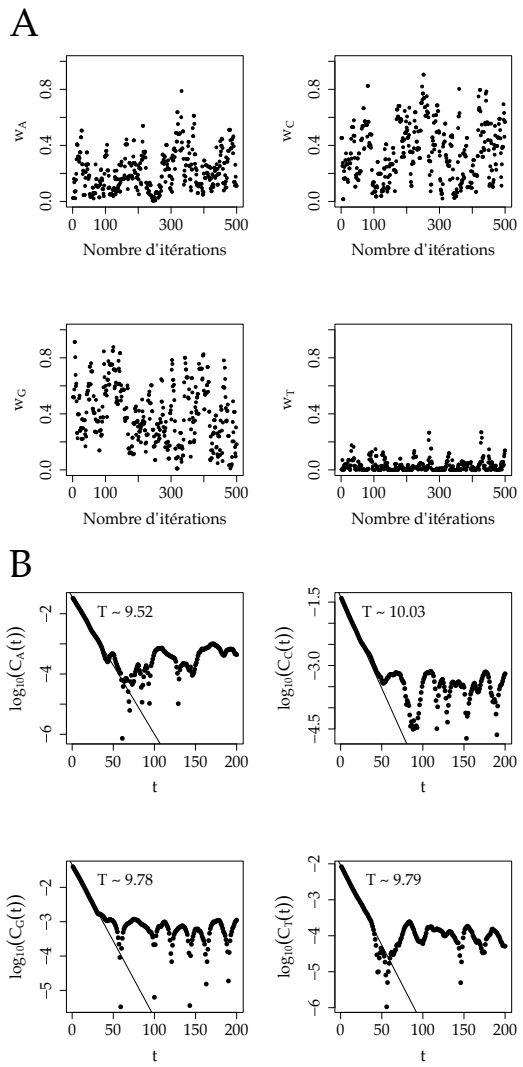


FIGURE 3.3 – **Corrélations entre les échantillons.**
(A) Extrait de l'échantillonnage de w_i. Les 500 premiers échantillons w_i de la chaîne MCMC sont montrés.
(B) Fonction d'auto-corrélation de la chaîne MCMC. La corrélation $C_b(t)$ est réalisée entre des échantillons séparés de $t-1$ itérations. Une droite est ajustée sur les 20 premièrs points, permettant d'obtenir la pente $-T_b$ donnant le temps de décorrélation.

intervalles plus grands que T_b. Dans notre cas, nous avons choisis un intervalle de 30 itérations.

Nous souhaitons maintenant étudier la convergence de la chaîne MCMC. Pour cela, nous utilisons le Théorème Central Limite (TCL, cf 3.3.1). Pour un nombre n suffisamment grand d'échantillons indépendants, l'écart-type de la distribution de l'estimateur empirique de la moyenne $\hat{w}_n$ se comporte comme $\sigma_b/\sqrt{n}$, où σ_b est l'écart-type de la distribution $\mathcal{P}(w_i|\mathcal{A})$. Ce dernier doit lui-même être estimé à partir de la chaîne MCMC. Nous présentons en figure 3.4A la valeur de l'estimation $\sigma_b(n)$ obtenue pour n itérations indépendantes. On voit que cette valeur atteint rapidement en $\sim$ 100 itérations une valeur stable, et que dans tous les cas

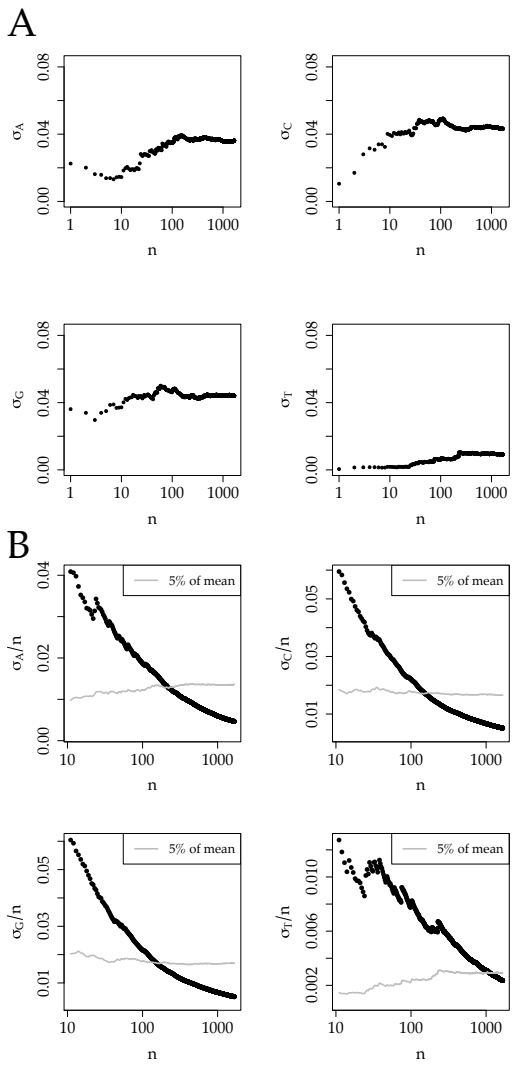


FIGURE 3.4 – **Estimation de la convergence.**
(A) Écart-type σ_b de la distribution $w_{i,b}$ estimé à partir de n échantillons indépendants. Ces échantillons sont pris toutes les 30 itérations au sein de la chaîne MCMC, soit environ 3 fois le temps de décorrélation. On observe qu'au bout de 100 itérations l'écart-type est stabilisé. Par ailleurs les fluctuations sont relativement faibles, au plus de l'ordre de $\sim$ 10% de la moyenne.
(B) La convergence est estimée grâce au Théorème Centrale Limite. L'écart-type de l'estimateur de la moyenne se comporte comme $\sigma_b/\sqrt{n}$. La précision demandée correspond à un écart-type $\leq$ 5% de la moyenne pour les 4 bases, ce qui correspond à l'intersection la plus tardive entre les courbes noire et grise (dans notre cas le cadran du bas à droite).

les fluctuations sont faibles, de l'ordre de 10% de $\hat{w}_n$. Il paraît donc raisonnable d'utiliser cette valeur de σ_b pour le TCL. Nous traçons en figure 3.4B la quantité $\sigma_b(n)/\sqrt{n}$. La chaîne est considérée comme convergée lorsque cette valeur est inférieure ou égale à 5% de la valeur moyenne estimée $\hat{w}_n$ (ligne grise) dans les 4 cas. Ce seuil est bien entendu arbitraire et dépend de la précision voulue par l'utilisateur sur l'estimation. Néanmoins, une plus grande précision

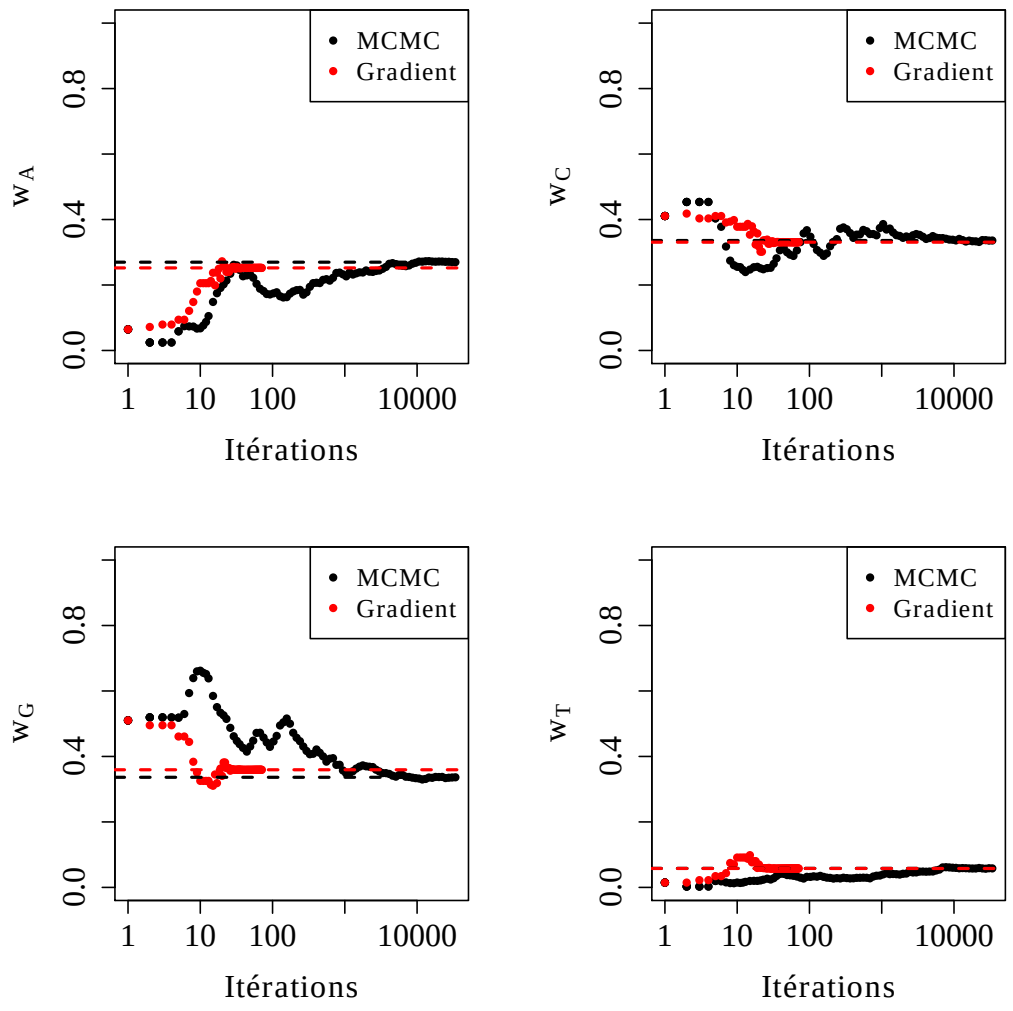


FIGURE 3.5 – **Comparaison des approches par MCMC et par descente de gradient.** L'approche de maximisation de (l'opposé de) la postérieure modifiée par descente de gradient (rouge, cf article) est comparée à l'approche de calcul de la moyenne de la postérieure par MCMC (noir). Les deux méthodes sont initialisées au même w_i^{inde}. Alors que la méthode par descente de gradient converge rapidement ($\sim$ 100 itérations), l'approche MCMC converge plus lentement, dans ce cas plus de 10,000 itérations. Au final les deux approches convergent vers des quantités proches (lignes pointillées rouges et noires).

implique un plus long temps de calcul.

Nous comparons en figure 3.5 la valeur de l'estimation du Maximum A Posteriori (MAP) de la postérieure modifiée (voir article) obtenue avec la méthode de descente de gradient et celle de la moyenne de la postérieure obtenue avec l'approche MCMC, en fonction du nombre total d'itérations. L'approche MCMC converge vers un état proche de celui donné par la descente de gradient, On note que la convergence de l'approche MCMC est beaucoup plus lente : plus de 10,000 itérations, alors que la descente de gradient n'en requiert que 100. Au

vu de la faible différence entre les deux résultats, nous utilisons dans Imogene l'approche de maximisation de la postérieure modifiée, ce qui permet un gain de temps considérable pour l'algorithme (au minimum un facteur 10).

3.4 Conclusion et perspectives du chapitre 3

Nous avons présenté Imogene, un algorithme bayésien utilisant la phylogénie de recherche de motifs et modules conduisant une régulation commune. Imogene est basé sur l'algorithme introduit par Rouault et al. (2010) dans le cas des Drosophiles, et l'étend au cas des mammifères. Nous avons présenté des tests d'Imogene sur des CRMs possédant une expression déterminée chez l'embryon de souris (tube neural et bourgeon de membre), et avons montré la capacité d'Imogene de prédire des CRMs conduisant à une expression similaire au sein de séquences intergéniques. Parmi les motifs générés par Imogene, certains sont associés à des régulateurs connus des étapes du développement considérées. Par ailleurs, nous avons montré que les motifs générés par Imogene pouvaient être utilisés pour générer un classifieur linéaire permettant d'associer un CRM donné à une classe liée à une expression spécifique. Ce classifieur montre des performances similaires à un classificateur basé sur des données biologiques spatio-temporelles extensives (Zinzen et al., 2009), mais ne nécessite que la connaissance de quelques séquences de chaque classe et fournit en plus la connaissance des motifs régulateurs.

Imogene peut être utilisé à partir de l'interface Mobyle de l'Institut Pasteur `http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene`. Nous espérons ainsi qu'il pourra servir à des biologistes souhaitant mettre à jour des régulateurs putatifs dans des CRMs fonctionnels et détecter dans le génome des CRMs possédant une activité similaire.

Chapitre 4

Étude de la différenciation épidermale chez la drosophile

4.1 Concept d'optimum de Pareto 143
4.2 Article 144
4.3 Conclusion et perspectives du chapitre 4 215

Introduction du chapitre 4

Nous présentons maintenant une application de Imogene au cas de la différenciation des trichomes (poils) chez la Drosophile, réalisée en collaboration avec l'équipe de Serge Plaza à l'Université Paul Sabatier. Plusieurs motifs ont été générés à partir de 14 CRMs connus pour réguler le processus de différenciation des trichomes. Parmi les motifs générés, deux d'entre eux ont montré une meilleure capacité à distinguer les CRMs positifs de l'ensemble d'apprentissage de CRMs négatifs. Le critère de distinction est basé sur l'optimisation de Pareto, caractérisant la satisfaction de plusieurs contraintes à la fois. Dans notre cas, les contraintes sont de maximiser le nombre de CRMs positifs trouvés par les motifs (maximisation de la sensibilité) tout en minimisant le nombre de faux positifs (maximisation de la spécificité). Ces critères définissent une frontière de Pareto de motifs optimum. En variant les différents paramètres de Imogene, nous avons trouvé deux motifs sur la frontière de Pareto. Parmi les deux motifs, l'un d'eux (« svbf7 ») correspond au régulateur maître du processus de différenciation des trichomes, et l'autre (« blue motif ») est un motif nouveau. L'importance des deux motifs pour la régulation est montrée par mutagenèse. Par ailleurs, ces motifs permettent de distinguer des ChIP-seq pour *svb* liés à une régulation (dont le gène le plus proche subit une perte d'expression lors du KO de *svb*) des ChIP-seq ne l'étant pas. Ce travail montre donc un exemple de la possibilité d'utiliser Imogene sur sun un petit ensemble (ici 14 CRMs) de données biologiques fonctionnelles pour détecter des motifs nouveaux et fonctionnels.

4.1 Concept d'optimum de Pareto

Dans l'article qui suit, nous utilisons le principe d'optimum de Pareto, lié à la satisfaction simultanée de plusieurs contraintes. Le problème est le suivant. Étant donnés 14 CRMs positifs liés à une même régulation de la différenciation des trichomes chez l'embryon de Drosophile, et 25 CRMs négatifs ne conduisant aucune expression au stade de développement considéré, quels motifs permettent le mieux de distinguer les deux classes ? Le problème est similaire à celui de *pattern recognition* introduit dans l'article précédent en section 3.2. Néanmoins, nous avons ici adopté une démarche légèrement différente.

Des motifs sont appris sur les CRMs positifs pour différents seuils S_g (variant entre 7 et 13 bits). Ces motifs sont ensuite utilisés pour prédire les CRMs positifs parmi les CRMs initiaux. Un CRM est déclaré comme positif s'il contient au moins n sites conservés ($n = 1$, 2, ou 3) au-dessus d'un seuil S_s (variant entre 7 et 13 bits). Pour des paramètres S_g, n et S_s donnés, un motif détectera un nombre FP de Faux Positifs (les CRMs négatifs qui sont prédits positifs) et omettra un nombre FN de Faux Négatifs (les CRMs positifs qui ne sont pas prédits comme tels). L'optimisation de Pareto consiste à trouver les paramètres S_g, n et S_s qui minimisent à la fois FP et FN. Il est possible de minimiser différentes fonctions de coût pour FP et FN, attribuant des poids plus ou moins importants à l'une ou l'autre des contraintes. Dans notre cas, nous avons défini l'optimum de Pareto comme minimisant la fonction $|\mathrm{FN} + \mathrm{FP}|$. La droite $|\mathrm{FN} + \mathrm{FP}| = cte$ contenant les meilleurs optima de Pareto pour les différents motifs est la frontière de Pareto.

Dans notre cas, deux des motifs générés sont sur la frontière de Pareto : le motif svbF7 et le motif bleu (fig. 4.1), pour $S_g = 10$ bits, $n = 1$, $S_s = 8.5$ bits. Ce sont par ailleurs les 2 premiers motifs générés à ce seuil. Les 3 motifs suivants sont montrés en gris. Le motif svbF7 correspond au TF *svb* (*Shavenbaby*), régulateur maître de la différenciation des trichomes. Le motif OvoQ6 correspond à la PWM de Transfac pour *svb*, et n'apporte pas une aussi bonne classification que svbF7. Le motif bleu, quant à lui, n'est pas connu. Nous montrons aussi le motif jaune, introduit dans l'article, qui n'est pas prédit par Imogene mais correspond à un motif de régulation ultra-conservé dans les de différentes espèces de drosophiles (Elemento and Tavazoie, 2005), et possédant un rôle fonctionnel investigué par mutagenèse.

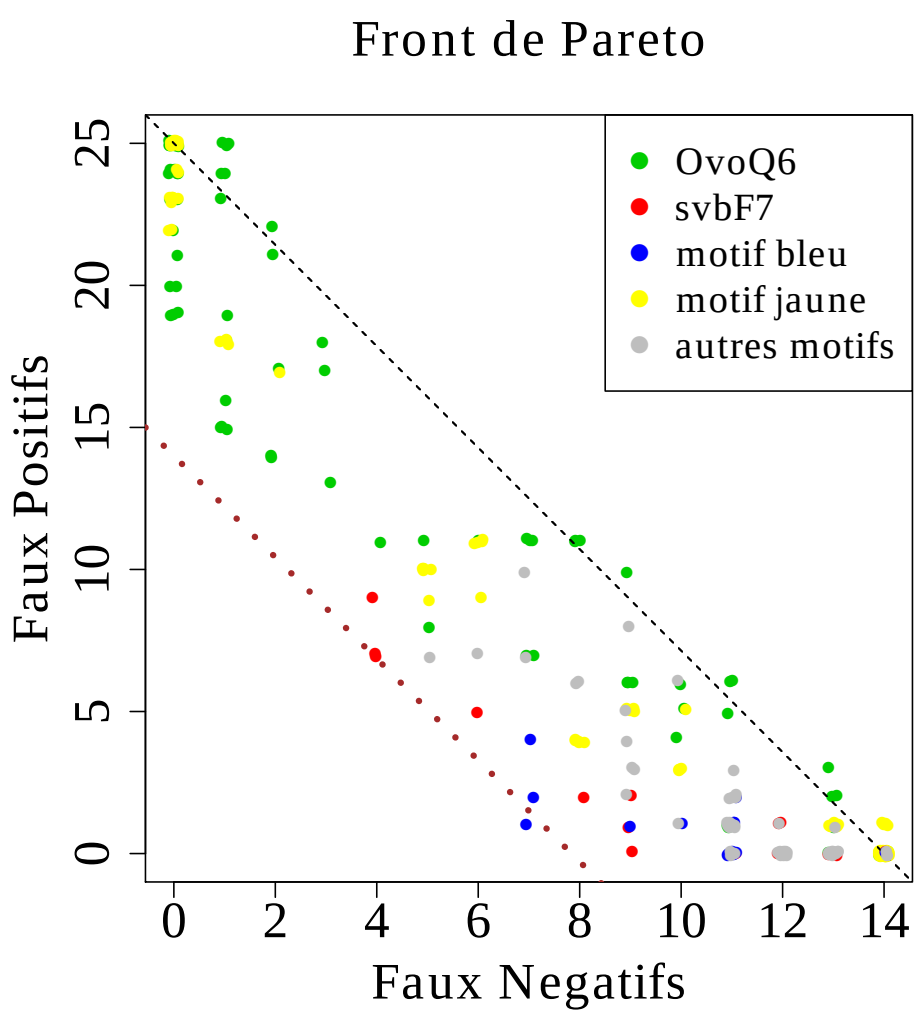


FIGURE 4.1 – **Frontière d'efficacité de Pareto.**
Illustration de l'optimisation de Pareto permettant la sélection des paramètres lors de la génération de motifs. Ici, nous voulons à la fois minimiser le nombre de Faux Positifs (FP) et le nombre de Faux négatifs (FN). Différents motifs sont successivement utilisés pour classer les séquences comme positives ou négatives, une séquence étant déclarée comme positive pour un motif si elle contient au moins n sites conservés au-dessus d'un seuil S_s. Les différents points correspondent aux différentes performances. Un léger bruit a été ajouté aux points pour rendre visibles ceux qui sont superposés (fonction `jitter` de R, paramètre `factor`= 0.5). Le motif svbF7 et le motif bleu sont tous deux optimaux pour des paramètres bien définis $S_g = 10$ bits, $n = 1$, $S_s = 8.5$ bits : ils dessinent la frontière de Pareto $|FP + FN| = cte$ en pointillés bruns.

4.2 Article

Genome-wide analyses of Shavenbaby target genes reveals distinct features of enhancer organization

Delphine MENORET[1,2*], Marc SANTOLINI[3*], Isabelle FERNANDES[1,2,4], Rebecca SPOKONY[5], Jennifer ZANET[1,2], Ignacio GONZALEZ[6,7], Yvan LATAPIE[1,2], Pierre FERRER[1,2], Hervé ROUAULT[3,9], Kevin P. WHITE[5], Philippe BESSE[6,7], Vincent HAKIM[3], Stein AERTS[8], Francois PAYRE[1,2#] and Serge PLAZA[1,2#]

[1] Centre de Biologie du Développement, Université de Toulouse, UPS, Toulouse, F-31062, France

[2] CNRS UMR5547, Toulouse, F-31062, France

[3] Laboratoire de Physique Statistique, CNRS, Université Pierre & Marie Curie, Université Denis Diderot, ENS, Paris, France

[4] present address: Department of Biology, McGill University, Montreal, QC, Canada

[5] Institute for Genomics and Systems Biology, Department of Human Genetics, The University of Chicago, Chicago, IL, USA

[6] Université de Toulouse, INSA, Toulouse, France

[7] Institut de Mathématiques, CNRS UMR5219, Toulouse, France

[8] Laboratory of Computational Biology, Center for Human Genetics KU Leuven, Belgium

[9] present address: Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, VA, USA

* The two first authors contributed equally

Corresponding authors: serge.plaza@univ-tlse3.fr, francois.payre@univ-tlse3.fr

Running title: Functional architecture of trichome enhancers

keywords: Drosophila, enhancer, epidermis, shavenbaby, morphogenesis

Abstract

Background: Developmental programs are implemented by regulatory interactions between Transcription Factors (TFs) and their target genes, which remain yet poorly understood. While recent studies have focused on regulatory cascades of TFs that govern early development, little is known on how are selected and controlled the ultimate effectors of cell differentiation. We addressed this question during late *Drosophila* embryogenesis when the finely tuned expression of a TF, Ovo/ Shavenbaby (Svb), triggers the morphological differentiation of epidermal trichomes.

Results: We defined a sizeable set of 39 Svb downstream genes and used *in vivo* assays to delineate 14 enhancers driving their specific expression in trichome cells, with highly similar pattern and dynamics. Coupling computational modeling to functional dissection, we investigated the regulatory logic of these enhancers. Genome-wide approaches, further extending the repertoire of epidermal effectors, support that the regulatory models learned from this first sample are representative of the whole set of trichome enhancers. We find that these "terminal" enhancers harbor remarkable features with respect to their functional architectures. They display weak if any clustering of Svb binding sites. The *in vivo* function of each site relies on its intimate context, with a critical importance of the flanking nucleotides. We identify two additional cis-regulatory motifs, retrieved in a broad diversity of composition and positioning among trichome enhancers, and that critically contribute to their activity.

Conclusion: Our results show Svb directly regulates a large set of terminal effectors of the remodeling of epidermal cells. Furthermore, these data reveal that trichome formation is underpinned by unexpectedly diverse modes of regulation providing fresh insights into the functional architecture of enhancers governing a terminal differentiation program.

Introduction

Many studies have established that transcriptional networks control development, through determining specific programs of genome expression [1]. These Gene Regulatory Networks (GRNs) are implemented by Transcription Factors (TFs) that bind to regulatory DNA sequences, known as enhancers or Cis-Regulatory-Modules (CRMs), to control the transcription of nearby genes. Although recruited to target genes via their DNA binding properties [2], TFs recognize only short and often degenerate motifs (reviewed in [3, 4]). Consequently, thousands of putative Binding Sites (BS) are scattered throughout the genome hampering efficient prediction of cis-regulatory regions [3, 5, 6]. The fine structure of enhancers as well as putative general rule(s) underlying their organization remain(s), however, poorly understood.

Although animals encode hundreds of TFs, only a few of them have been studied in detail to elucidate the regulatory logic of their target enhancers [7, 8]. In *Drosophila*, current knowledge of enhancer structure mainly comes from works on early development, *e.g.* TFs controlling segmentation and mesoderm specification [9-12]. Within these early acting networks, a number of studies have shown that the local enrichment for BS (homotypic or heterotypic clustering) in evolutionarily conserved regions is a general signature of active enhancers [13-15]. Functionally related enhancers (driving similar expression pattern) often share a combination or code of cis-regulatory motifs, together defining a specific program of expression [11, 16-18]. Whether enhancers rely on a constrained organization of cis-regulatory motifs or can accommodate flexibility in their number, composition and positioning remains debated (reviewed in [4, 19, 20]). While several studies have shown that regulatory codes are efficient to predict expression pattern [9, 11, 16], recent large-scale work suggests that developmental enhancers may have a more flexible architecture [10,

20]. However, in depth analyses of individual enhancers [21-24] have revealed an unexpected level of functional constraint in their intimate architecture. It has been proposed that constrained enhancers could be critical when TFs display limiting concentrations [25], *e.g.* to accurately integrate gradients [26]. On the other hand, enhancers that do not hold integrative properties might be of simpler architecture [27, 28]. Distinguishing between these possibilities thus requires detailed analyses of the structure and regulatory logic of CRM/TF interactions that occurs at late developmental stages.

Here, we focus on a GRN that controls cell morphogenesis during terminal differentiation of the *Drosophila* embryonic epidermis. The subset of epidermal cells that express the TF Ovo/Shavenbaby (Svb) [29] undergo localized changes in cell shape leading to the formation of dorsal hairs and ventral denticles, collectively referred to as trichomes [30]. Svb triggers the expression of various classes of cellular effectors in trichome cells. Developmental and genetic analyses have established that trichome formation relies on their collective action, acting together as a developmental module to promote cell shape reorganization [31-33]. The mechanisms underlying the co-expression of Svb-regulated genes in trichome cells remained yet poorly understood. A first level of regulation resides in the activity of Svb itself that is controlled in a post-translational manner, in response to small peptides encoded by the gene *polished-rice (pri)* [34]. Pri peptides trigger N-terminal truncation of the Svb protein, switching its activity from a full-length repressor to a cleaved activator [34], therefore providing a temporal control to the program of trichome formation [32]. However, little is known concerning how this TF recognizes and selects its target genes. Besides definition of DNA-binding specificity *in vitro* [35] and the identification of few targets regulated by Ovo germline-specific isoforms [35,

36], only a single epidermal enhancer dependent on Svb has been identified so far [31]. Thus, whether or not Svb targets genes that are co-expressed in trichome cells co-opted similar cis-regulatory elements remained an open question.

To address this question, we designed a set of computational modeling coupled to experimental approaches to identify and investigate the cis regulatory logic of Svb-dependent enhancers. By systematic *in vivo* assays, we first identified a robust set of Svb target effectors, specifically expressed in trichome cells at the time of their morphological differentiation. We then searched for and identified 14 Svb-dependent epidermal enhancers driving their expression in trichome cells and investigated their functional organization. Computational analyses and experimental dissection led to a refinement of the Svb BS bound *in vivo* and the identification of two additional motifs required for enhancer activity. Our studies further reveal that the distribution of these *cis*-regulatory motifs does not follow a stereotypical organization. Coupled to chromatin immunoprecipitation (ChIP-seq) and microarray profiling, the models built from these fine scale experiments allow efficient genome-wide identification of new enhancers, which drive the specific expression of trichome effectors. In summary, our results show that enhancers driving co-expression in cells of a late GRN present various composition and respective organization of cis regulatory motifs, extending the idea that co-expressed developmental enhancers can have diverse cis-regulatory architectures [11, 37], including for those mediating terminal stages of cell differentiation.

Results

Enrichment of conserved binding sites in Svb downstream genes

Previous work has identified a dozen of genes activated by Svb, each contributing to epidermal cell remodeling [31, 33, 38, 39]. To investigate the cis-regulatory logic of Svb-dependent targets, we first sought to define a larger set of Svb downstream genes appropriate for *in silico* analyses. We therefore analyzed additional candidates selected because of their expression in subsets of epidermal cells (BDGP) using *in situ* hybridization. Among 57 candidates, we identified 21 Svb-dependent genes, *i.e.*, downregulated in *svb* mutants and upregulated following *svb* ectopic expression (Fig. 1A, S1A), while the other 36 epidermal genes were found independent of Svb (Fig. S1B). Together with genes identified previously [31, 33, 38, 39], this constituted a robust set of 39 genes activated by Svb to be expressed in trichome cells. We used these 39 Svb targets to examine whether they displayed an evolutionarily conserved signature in their non-coding regions, when compared with all *Drosophila* genes, or the 36 epidermal genes independent of *svb* as a negative control. The recent method cisTargetX aims at detecting motifs enriched among a group of co-expressed genes, *e.g.* to predict direct targets of a TF [40]. It exploits a library of >3000 motifs, including TF binding sites and ultra-conserved DNA words [41], each motif being ranked with a score representative both of clustering and evolutionary conservation [40]. When applied to Svb targets, 4 of the top 5 motifs match the consensus CnGTT (Fig. 1B, S1C), characteristic of the Ovo/Svb BS CnGTTa as defined *in vitro* [35]. From the 39 input genes, CisTargetX determined an optimal subset of 16 Svb direct targets, having the highest scores for the OvoQ6 motif (Fig. 1B and S1C) [35, 36]. OvoQ6 was specific to Svb targets since it was not

detected in control epidermal genes (Fig. S1C). In contrast, motifs matching the BS of TFs involved in general epidermis differentiation such as Grainy head [42] or Vrille/c-EBP [43] were highly ranked in Svb independent genes (Fig. S1C), whilst detected only at low score in Svb downstream genes. Hence, OvoQ6 motifs appear as a signature of a subset of genes activated by Svb, a result consistent with their direct regulation.

Distribution of Svb BS clusters poorly correlates with enhancer activity.

We then examined the genomic distribution of OvoQ6 motifs within Svb target loci showing significant enrichment when compared to random *Drosophila* genes. We found that each target gene contained evolutionarily conserved OvoQ6 scattered throughout intergenic and intronic regions (Fig. 2A,B), instead of high OvoQ6 clusters enriched locally (even using relaxed conditions of at least 2 sites *per* kb). To delineate which regions mediate epidermal expression, we generated a series of transgenic reporters scanning systematically two Svb downstream genes. We focused on *singed* since it encodes Fascin, a conserved regulator of actin organization [44], and *shavenoid* that encodes a pioneer protein but displays an extreme trichome phenotype upon its inactivation [31]. Although most regions with OvoQ6 sites did not show embryonic expression, we identified three sequences one in *singed* (*snE1)* and two in *shavenoid* (*sha1* and *sha3)* that drove expression in the epidermis, specifically in trichome cells (Fig. 2A,B). Somehow unexpectedly, one of the three sequences, *sha1,* displays a single recognizable OvoQ6 motif (see below) in *D. melanogaster,* as well as in sibling species. The activity of all three regions was lost when introduced in *svb* null mutant background, showing that they are functional Svb target enhancers (Fig. 2A,B). cisTargetX predicts the location of putative

enhancers within each gene [40] and two out of three enhancers defined *in vivo* matched these predictions, in one case (*sha3*) to the highest rank for this gene (Fig. 2C). We therefore investigated whether evolutionarily conserved OvoQ6 sites were sufficient to predict trichome enhancers and assayed 18 additional regions (Fig. 2C) taken from the top 100 predictions. Transgenic reporter assays identified 4 novel sequences from *CG15589*, *cypher*, *dusky-like* and *neyo* driving expression in the epidermis. We verified in each case that they were specifically expressed in all (*dyl2, nyo1*) or subsets (*15589, cyrA*) of trichome cells where Svb is active. Consistently, these four enhancers depended on Svb since they displayed a strong reduction of their expression in the absence of *svb* (Fig. 2C). Hence, analysis of Svb downstream targets shows that they are enriched in OvoQ6 BS, a feature well conserved across *Drosophila* species. However, putative trichome enhancers predicted from evolutionary conservation and clustering of OvoQ6 sites were validated only at a rate of 28% (6/21, Fig. 2C), most tested regions being devoid of activity in embryos, suggesting that other criteria distinguishes enhancers from negative regions.

We noticed that OvoQ6 clusters failed to predict a number of active enhancers. This was the case of *sha1* (Fig. 2) or *Emin*, an epidermal enhancer previously identified for *miniature* [31]. Examination with Cluster-Buster [45] or Swan [46] did not detect supplementary OvoQ6 in *sha1* or *Emin* sequences (even in *D. melanogaster* only), explaining that these enhancers, containing a single Svb BS, escape from *in silico* predictions. 6 additional enhancers identified during initial stages of this study using alternative prediction criteria (see Fig. S1C) were not highly ranked by cisTargetX, because they lack BS clustering and/or evolutionary conservation. These data therefore show that BS clustering is not an absolute requisite for Svb regulation (Fig.

2C), suggesting that additional sites are required to discriminate between enhancers and inactive regions.

***De novo* motif discovery identifies a specific signature of Svb BS active *in vivo*.** To search for these putative sites, we compared the two sets of experimentally tested regions, *i.e.* the 14 enhancers (positive) and 25 inactive regions (negative), using *Imogene*, an algorithm designed for *de novo* motif discovery [47]. Briefly, we systematically searched, *ab initio*, for 10bp motifs that are evolutionarily conserved across *Drosophilidae* and display a distribution within each region statistically different from background sequences. We then evaluated how well each motif discriminated between enhancers and inactive regions and ranked these *de novo* motifs accordingly (Fig. 3A). Strikingly, the most discriminative motif overlaps OvoQ6 (CnGTTa), with a similar core consensus but extended to adjacent nucleotides (**A**CHGTT**AK**). A second discriminative motif (WAGAAAGCSR) called blue motif was also found, and will be studied below.

The **A**CHGTT**AK** motif, hereafter called svbF7, was sufficient to detect 10 out of 14 enhancers (Fig. 3B). The proportion of svbF7-positive enhancers reached 13/14, when relaxing the penalty imposed for poor conservation [47]. In contrast, svbF7 was found in only 6/25 negative regions (Fig. 3B), even when lowering the threshold (data not shown). Once added to the cisTargetX library, svbF7 is the most significant motif found in the set of 39 Svb downstream genes (Fig. S1C,D). It also increased the accuracy of enhancer predictions, with three additional positives *32159, Emin* and *EminB* while 9 negatives were removed from the top100 cisTargetX regions (Fig. S1C). Hence, svbF7 performs better than OvoQ6 or any other related motifs [48] (Fig. 3B,S1D). To evaluate whether this slight extension of the Svb BS

was relevant for activity, we substituted nucleotides flanking the core CnGTTa in the single svbF7 of *Emin*, *i.e.* altering the svbF7 motif without disrupting OvoQ6 consensus (Fig. 3C). When assayed *in vivo*, different patterns of flanking substitutions including a single point mutation of the 5' A residue were sufficient to strongly reduce *Emin* expression (Fig. 3C). This demonstrated the functional importance of flanking nucleotides within the svbF7 motif for CRM activity. Hence, the computational analysis of Svb-dependent enhancers has discovered a refined nucleotide sequence required for *in vivo* regulation.

Trichome enhancers use different combinations of cis-regulatory motifs.

Having shown the role of svbF7 in *Emin*, we investigated its functional significance in other enhancers. We focused on enhancers containing from 1 to 3 predicted SvbF7 sites, to address the importance of single versus clustered BS for trichome cell expression. As observed for *Emin*, disruption of the single svbF7 site abolished the activity both of *sha1* and of *nyo1* (Fig. 4A,B). The mutation of svbF7 also decreased the activity of *tyn2*, albeit weakly and only in ventral cells (arrowhead, Fig. 4C). In this enhancer, we detected however a second putative site that appears less conserved across species. Its inactivation strongly reduced expression (Fig. 4), showing that this site mainly contributes to *tyn2* activity. For *sha3* and *dyl2* that contain 2 or 3 svbF7 respectively, simultaneous inactivation of these sites abrogated expression (Fig. 4D-E). The individual disruption of svbF7 led nonetheless to varying defects. The two svbF7 sites of *sha3* are partly redundant, with a similar and limited impact of individual KO when compared to the simultaneous KO (Fig. 4D-G). In contrast, a single svbF7 plays a major role in *dyl2* activity, whereas the two others contribute marginally to expression pattern or levels (Fig. 4E,H). Hence, the disruption of svbF7

leads to a reduced expression for all enhancers that have been tested, confirming the functional importance of this motif. Nevertheless, the introduction of two copies of the svbF7 motif within negative regions (*sha2* and *12063*) was not sufficient to promote expression in trichome cells. In addition, the individual inactivation of multiple svbF7 sites has different consequences on enhancer activity, suggesting that additional elements are likely to modulate, locally, the *in vivo* function of svbF7.

We thus searched for additional cis-regulatory motifs and evaluated their contribution to the activity of trichome enhancers. In a first approach, we performed a systematic mutagenesis of the *Emin* enhancer by linker scanning (Fig. 5A). In addition to svbF7 whose inactivation abolished *Emin* activity (F7mt), the mutation of three regions (8mt, 9mt and 10mt) strongly decreased epidermal expression, two others (3mt, 4mt) affecting only the *Emin* pattern ventrally (Fig. 5A). These results show that while Svb acts as a main switch for *Emin* activity, other motifs are required for complete expression. Interestingly our *de novo* motif discovery identified a second discriminative motif (WAGAAAGCSR), hereafter called the blue motif, enriched in positive regions and evolutionarily conserved in 7 out of 14 enhancers (Fig. 3A,B, 5B). Mutations that disrupted the blue motif (9mt & 8mt) of *Emin* displayed the strongest effect, besides svbF7 KO (Fig. 5A). These unbiased data show that the blue motif represents an element that, in addition to svbF7, is critical for *Emin* activity. To further test its contribution to the activity of trichome enhancers, we mutated the blue motif in two other enhancers that contain a single occurrence of it (Fig. 5B). As observed for *Emin*, disruption of the blue motif reduced *snE1* expression (Fig. 5C). Furthermore, the blue motif plays a key role in *sha3* activity, its inactivation abolishing expression (Fig. 5C) similarly to the simultaneous mutation of both svbF7 sites (Fig. 4D). In addition, we noticed that one important region for *Emin* expression (10mt, Fig.

5A) matches a 8mer (TTATGCAA), previously predicted as a regulatory element from discovery of ultra-conserved DNA words in the genome of distant *Drosophila* species [41]. Although not sufficient by itself to discriminate between active enhancers and negative regions (data not shown), this motif, which we called yellow motif, was nevertheless retrieved in 6 additional trichome enhancers (Fig. 5B). To further assay *in vivo* the role of yellow motifs, we generated mutant versions of the *17058* & *nyo1* enhancers that disrupt their yellow motif. As observed for *Emin*, mutation of the yellow motif led to a strong decrease in the expression driven by both *nyo1* and *17058* (Fig. 5D), showing that the yellow motif represents a functional cis-regulatory element in a subset of enhancers.

Taken together, these data therefore support that svbF7 is a main feature of Svb targets, this motif being shared by the vast majority (13/14) of active enhancers. Our analyses have discovered two additional cis-regulatory elements, the blue and yellow motifs, present in overlapping subsets of trichome enhancers (9/14 & 7/14, respectively). While the 3 motifs show various patterns and combinations, functional assays demonstrated that each of them contributes to the *in vivo* activity of this sample of trichome enhancers.

Genome-wide prediction of Shavenbayby target enhancers.

To address whether these cis-regulatory motifs were a relevant signature of the genome-wide set of enhancers regulated by Svb, we undertook ChIP-seq to obtain an extensive cartography of Svb binding sites in epidermal cells. To improve specificity, we used a Svb::GFP transgene driven in ventral and dorsal trichome cells by two complementary *svb* cis-regulatory regions [34], likely at levels comparable to endogenous since it rescues *svb* mutant phenotypes [49]. ChIP-Seq data indicated

that Svb was bound to almost 6000 genomic sites, a large number of binding events being a feature shared by several *Drosophila* TFs [6, 8, 15]. Analysis of ChIP peaks with *i*-cisTarget [50] showed that svbF7 and OvoQ6 are the most enriched motifs. A strong cross correlation between conserved svbF7 and the center of ChIP peaks confirmed the importance of this motif (Fig. 6A). As observed in our pilot analysis of enhancers, we did not detect high svbF7 clustering, multiple svbF7 motifs being rarely found within genome-wide ChIP peaks. Blue motifs (and to a lesser extent yellow motifs) also displayed a significant but weaker correlation with Svb peaks, consistent with wider genomic distribution (Fig. 6A).

With the large number of Svb bound regions detected by ChIP-Seq, it was unlikely that all of them were functional, as being involved in the regulation of target genes [5, 15]. Therefore, in order to identify the entire set of genes regulated by Svb, we performed microarray profiling, comparing wild type to mutant embryos. In mRNA samples prepared from *svb* whole embryos, we often detected only a modest reduction in the levels of validated targets (Fig. 6B, S3), challenging unambiguous identification of Svb downstream genes. In the absence of *pri,* Svb behaves as a dominant repressor [34] and consistently we observed a stronger decrease in the levels of known Svb targets in *pri* mutants (Fig. 6B,C,S3), therefore providing an additional criteria to identify the genes regulated by Svb. Henceforth, we selected the genes down-regulated in *svb* mutants and that also displayed a further (>2 fold) reduction of their expression in *pri* mutants, as benchmarked for known Svb targets. This defined a set of 150 genes encompassing 16/39 Svb targets validated *in vivo* (Fig. S1A), as well as 42 additional epidermal candidates (Fig. S3). Among these, we examined 23 genes by *in situ* hybridization and confirmed that 21 of them required Svb to be expressed in trichome cells (Fig. 6B,C and S4). These results therefore

show that microarray profiling has defined a representative set of genes activated by Svb in trichome cells.

Focusing on this genomic set of Svb-regulated genes, we found 172 peaks associated with 85 genes (Fig. S3), including 11 out of 14 active enhancers (Fig. S7). Within the whole set of relevant Svb-bound regions, we retrieved the characteristic features of cis motifs as defined previously. Although retrieved in many Svb-bound regions (Fig. 6A, S5), the enrichment in yellow motifs within ChIP peaks associated with Svb regulated does not reliably reach a significant threshold, consistent with a broad genomic distribution [41]. In contrast, we found clear association of svbF7 motifs and to a lesser extent of blue motifs (Fig. S5). Importantly, these motifs were not detected in peaks associated with a control set of genes independent of Svb (Fig. S5), strongly supporting that they are hallmarks of Svb-target enhancers. As an independent way to evaluate this conclusion, we used *ab initio* analysis of ChIP peaks using PeakMotif [51]. This identified the motif ACAGTTA characteristic of peaks associated with Svb downstream genes that extensively matches svbF7 (Fig. S6). A second sequence (TGAAAAG) partly matching the blue motif was also detected in about 50% of peaks, again only in Svb-regulated genes and not among control genes (Fig. S6).

Hence, we interpret these results to imply that svbF7, and to a lesser extent the blue and/or yellow motifs, would allow predicting the location of additional trichome enhancers (Fig. 7A). To evaluate this, we tested ChIPed regions containing svbF7 alone *(12017, 14395)*, svbF7 in association with either the blue motif *(mey2, EminC, actn, 12017-2)* or the yellow motif *(31022, 4914),* or all three motifs together *(9095, 11175)* (Fig. 7B and S7). We found that 8/10 (80%) of these regions act as Svb-dependent enhancer when assayed *in vivo* (Fig. 7B). Indeed, they drove robust

expression, specifically in trichome cells, and their activity was reduced in *svb* mutant embryos (Fig. 7B). Moreover, these data confirm that trichome enhancers are generally built from different combinations of the three cis-regulatory motifs. For example, only a subset of newly predicted trichome enhancers relies on the blue motif, since *mey2*, *EminC, 9095* and *11175* contain conserved blue motifs whereas *12017, 31022* and *4914* do not (Fig. 7B , S7). In the case of the *actn* enhancer, there are four partly degenerate blue motifs in the sequence from *D. melanogaster* and sibling species, while not retrieved in more distant species suggesting a turnover of cis-regulatory motifs (Fig. S8). However, aside a couple of fast evolving enhancers, we found in many cases a remarkable conservation of the pattern of svbF7, blue and yellow motifs within individual enhancers across distantly related *Drosophila* species (Fig. 8, S8).

Therefore, the regulatory signatures learned from modeling and experimental dissection of a subset of enhancers helps understanding how the Svb TF selects the genomic set of its direct targets. Furthermore, they collectively allow efficient identification of cis-regulatory regions that specify the program of trichome-specific expression in response to Svb.

Discussion

It is well established that the Shavenbaby TF determined the trichome fate [29, 32, 52], however little was known on the repertoire of its direct target genes and mechanistic insights into the functional organization of trichome enhancers were lacking. Combining functional dissection, computational modeling and genome-wide profiling, we provide here a molecular map of the ultimate repertoire of genes and cis-regulatory elements implementing the network of trichome differentiation.

Physical elements of the GRN governing trichome formation

Our results identify a high-confidence set of more than 150 genes activated by Svb in trichome cells. We confirmed 60 of those, showing complete or partial down-regulation in the absence of active Svb protein. While most genes are expressed in all trichome cells, some are restricted to trichome subsets suggesting that they can contribute to the diversity in trichome shape and organization observed along the body [52]. Functional annotation (Gene Ontology and manual curation) indicates that Svb controls terminal players of trichome differentiation. In addition to novel factors of F-Actin organization [31, 39], ECM remodeling [31, 33], cuticle formation [31, 38] and pigmentation [31], we identify enzymes involved in oxidation-reduction, proteolysis and cell trafficking, further extending the repertoire of cellular functions involved in the terminal differentiation of trichome cells. Hence, a major role of Svb in trichome formation is to directly activate the expression of a battery of cell morphogenesis effectors. In support of this, ChIP-Seq peaks are present in >70% of these Svb-dependent effector genes. Experimental assays further validated 22 functional enhancers driving the expression of genes encoding factors involved in cytoskeletal or ECM reorganization, sugar binding, proteolysis and additional enzymes.

Recent work has established that apparently redundant, or shadow, enhancers ensure robust expression of transcription factors [53, 54]. For example, the transcription of *svb* itself involves separate enhancers that buffer the trichome pattern against variations in the genetic background and external conditions [53]. It has been proposed that shadow enhancers are required to drive an acute expression of some key developmental regulators [55]. We define within both *shavenoid* and *miniature* separable enhancers (*sha1*, *sha3* & *Emin, EminB, EminC*) that mediate Svb regulation. These data indicate that apparently redundant enhancers may not be limited to regulatory factors operating at high hierarchic positions in gene networks. Instead, we provide evidence that several "blue collar" effector genes display a similar regulatory architecture, suggesting that multiple enhancers represent an overlooked feature of the successive tiers of gene networks.

Binding site clustering as a general signature of active enhancers?

Early acting enhancers often comprise multiple BS for a given transcription factor [56, 57]. For example, conserved BS clusters have identified target enhancers of Dorsal [13] or Bicoid [58] and feature functional Twist-bound regions [15]. Of note, most algorithms developed for enhancer detection extensively use motif clustering as an important predictor [59]. We found a clear enrichment in putative Svb BS (OvoQ6 motif) in its downstream genes. However, only a small proportion of these motifs mediate *in vivo* regulation. There is very limited, if any, clustering of Svb BS in ChIP peaks associated with Svb target genes, and even genome-wide. Within the trichome enhancers we validated experimentally, 13 out of 22 display a single Svb site. Furthermore, for the enhancers *tyn2*, *sha3* and *dyl2* that contain 2-3 Svb BS, the inactivation of individual sites has often limited consequences, as also reported for

other TFs [60]. Even if some sites have been missed by computational approaches, the presence of multiple BS within a short region is not a deterministic feature of active Svb-dependent enhancers.

These findings highlight a paradoxical discrepancy between the enrichment of putative BS accumulated in Svb downstream genes and the limited number of those acting as cis-regulatory elements. Is there a role for this evolutionary accumulation of Svb-like motifs in Svb targets? For example, these sites presumably of weaker affinity (at least *in vivo*) can increase the local concentration of the TF facilitating regulation through a few BS stably bound *in vivo*, as it has been suggested on thermodynamics grounds [61] or to explain the existence of thousands of binding events that are transcriptionally inactive [5, 15].

Trichome enhancers rely on diverse combinations of cis-regulatory motifs

We found that the motif bound by Svb *in vivo* is more constrained than the consensus defined from *in vitro* [35] or one-hybrid approaches [48]. This shows that slight sequence differences, not detected *in vitro*, can play a key role within genomic context [62], for instance revealing the influence of co-factors [63].

In addition, other motifs influence which Svb BS are functional as regulatory elements, a notion well in line with recent results on the *in vivo* specificity of Hox factors [64]. Our statistical approaches identified a wider spread “blue” motif. Importantly, only half of the enhancers comprise blue motif(s), indicating that there are several ways to build Svb-responsive enhancers. Indeed, the systematic dissection of Emin disclosed an additional motif (TTATGCAA) ultra-conserved across *Drosophilidae* [41] and contributing to its activity. This “yellow” motif is retrieved in half of the trichome enhancers, with or without blue motifs. It is however barely

enriched in Svb-bound regions and therefore was not predicted by computational analyses, showing the importance of unbiased functional dissection to disclose the full spectrum of cis-regulatory elements. Indeed, the disruption of either blue or yellow motifs strongly affects enhancer function in all tested cases, providing experimental evidence of their cis-regulatory activity.

Trichome enhancers thus display various combinations of motifs, from those containing only Svb BS (5/22), Svb plus yellow (4/22), Svb plus blue (6/22) or all three together (7/22). These different motif compositions do not appear to correlate with distinct subclasses of gene function (DM unpublished). Furthermore, multiple enhancers from the same gene can harbor distinct combinations, as exemplified by *shavenoid* and to a lesser extent by *miniature* (Fig. 8, S6). Several studies have shown that motif composition may correlate with a given spatio-temporal pattern, *e.g.* for neurogenic or muscular GRNs [11, 16]. Since most trichome enhancers are often active in the very same population of cells, with highly similar dynamics, it is surprising to observe such diversity in their motif compositions. There are yet four enhancers restricted to dorsal trichome cells, but again they accommodate different motif compositions, with *EminB* & *4702B* that contain blue motifs *vs cyrA* & *31559* without. These data thus indicate that trichome enhancers display diverse distribution of functional motifs, supporting that distinct cis-regulatory architectures drive highly similar spatio-temporal expression.

Flexibility in cis-regulatory motifs among enhancers *vs* across species

Although highly constrained sequences, such as the interferon-B enhanceosome, do not seem widely spread [20], developmental enhancers may yet require some “grammar” for motif positioning [23], *e.g.* with an optimal pair-wise spacing of motifs

[64] that could reflect the cooperative binding of TFs. For trichome enhancers we did not detect any obvious bias in the number or respective arrangement of the cis-regulatory motifs they rely on (Fig. S2A). Likewise, recent results from the analysis of *Drosophila* cardiac enhancers support that similar expression patterns can be generated from divergent compositions and positioning of motifs [10, 65].

That several different inputs lead to similar enhancer outputs does not, however, formally rule out the existence of constraints, even though they are not detected by “horizontal” comparison of different enhancers within a same species. An independent way to evaluate this possibility is to look at the evolution of individual regulatory regions throughout species [15, 21]. Across *Drosophilidae*, trichome enhancers often display similar number and organization of cis-regulatory motifs (Fig. 8, S6). Furthermore, besides turnover of some motifs, svbF7, blue and yellow motifs are often imbedded within short-sized islands of high evolutionary conservation, when compared to neighboring sequences (Fig. 8). Similar strong evolutionary conservation was also noticed for the binding site of Twist [62] and its partner TFs [15], although these studies did not examine evolution of the detailed pattern of motif positioning. These data therefore suggest that despite diverse arrangements of motifs, patterns of evolutionary conservation likely represent the signature of functional constraints that locally shape the architecture of individual enhancers.

Materials and Methods

Fly strains and transgenic constructs

We used *btd, svb*^{1} or *svb*^{R9} [30, 31] and *pri*^{1} [34] stocks kept over GFP balancers. To delineate the epidermal enhancer of *sn* and *sha*, transgenic lines were initially generated using P-element mediated transformation (Fly Facility) and at least three independent insertions were analyzed for each construct. We then switched to the PhiC31 system (Bestgene) to quantify effects of mutations, with all constructs integrated at the same location (*zh-86F*), except for *sha1*, *sha3* and *snE1* for which mutant versions were assayed in P-elements for homogeneity (see supplemental methods). Genomic regions were amplified and cloned into pCasper or pAttB lacZ derivatives. QuickChangeXL site-directed mutagenesis (Stratagene) was used to introduce point mutations in enhancers, or CCGCCGGCGG stretches for linker scanning of *Emin*. All constructs were verified by sequencing.

Embryo staining

Dig- or biotin-labeled antisense RNA probes were used for *in situ* hybridization following standard protocols and embryos imaged using a Nikon Eclipse90i microscope. For immunodetection of lacZ reporter expression, 10-14h embryos were stained using anti-β–galactosidase (Cappel, 1/1000) and Alexafluor488 (Molecular Probes). Pictures were taken with a Leica SP2 confocal microscope, using the same settings to allow quantitative comparisons.

Microarrays

13-15h *svb*^{R9} or *pri*^{1} embryos were hand-selected using GFP balancers. 200 embryos were subjected to trizol (Invitrogen) extraction and RNA quality was monitored using Agilent Chip. Five independent samples of each genotype were used for microarrays (Affymetrix; IGBMC, Strasbourg). Data extraction and normalization were performed using Affymetrix software

and statistical analyses with R. A >2 fold difference in expression levels between mutant genotypes was the most efficient criteria to retrieve Svb downstream genes (with a false discovery rate (FDR) of 0.01 for *pri*). The top 150 genes downregulated both in *pri* and in *svb* mutants defined the set of Svb-regulated genes. 100 genes showing irrelevant variation of their expression (p-value>0.8, FDR>0.99) were used as a negative control set.

ChIP-seq

A *svb* rescue construct (RSQ8) {Kondo, 2010 #165} was used for ChIP-seq experiments. It expresses a Svb-GFP protein under the control of two *svb* enhancers (medial and proximal) driving specific in epidermal trichome cells. Stocks were expanded to fill three population cages. Adults were allowed to lay eggs for 2 hours on apple juice plates covered with yeast. Embryos deposited on the plates were aged for 12 hours at 25°C. Chromatin was collected from approximately 100 mg of whole embryos for each replicate chromatin collection. Chromatin immunoprecipitation was done with an anti-GFP antibody as described [8]. Data presented are from two independent replicates. Peaks were called for single replicates using MACS p<0.00001 for downstream computational analyses. MACS was used to call loose criteria peaks for two replicates of RSQ8 12-14 hr embryo. Those peaks were then used for an IDR analysis, IDR=0.02. DNA sequencing libraries were generated with Nextera DNA Sequencing Libarary Kits. Additional details are given in supplemental methods.

Motif detection and genome analysis

Detection of motifs enriched in Svb-dependent and Svb-independent epidermal genes was performed using *cisTargetX* [40]. For *de novo* motif discovery, genomic sequences of enhancer and negative regions were processed through a C++ program and statistical operations performed within the R software, as described [47]. To compute the cross-correlation between conserved motif instances and Svb ChIP-Seq data, we defined a 10 Kb window centered around each ChIP peak, collected distances of each motif to the peak center and plotted these values using a 500bp bin. In the cases of Svb-regulated and control

genes, each ChIP peak was associated with the nearest transcription start site. Additional details are given in supplemental information.

Acknowledgements

We are grateful to the Bloomington Drosophila Stock Center and Drosophila Genomic Resource Center for providing us with flies and molecular clones. We are indebted to B. Ronsin (Toulouse RIO Imaging), P. Valenti and O. Bohner for excellent technical assistance. We also thank C. Hermann, N. Negre, E. Preger-Ben Noon, A. Vincent and members of the SP/FP lab for critical reading of the manuscript. This work was supported by grants from ANR Blanc “Netoshape”, Association pour la Recherche contre le Cancer (n°3832, n°1111, SFI20101201669 and fellowships to IF & DM) and University Paul Sabatier (cisdecode). SA was supported by grants from FWO (G.0704.11N and G.0640.13) and HFSP (RGY0070/2011). SA, FP and SP designed the experiments, MS, HR and VH statistical analyses; DM, MS, IF, RS, JZ, IG, YL, PF, SA & SP performed the experiments and are listed according to their contributions. SP and FP wrote the paper and all authors commented on the manuscript. The authors declare no conflict of interest.

FIGURE LEGENDS

Figure 1: Enrichment in binding sites defines an evolutionarily conserved signature of *svb* downstream genes

A: Expression of *svb* mRNA determines the epidermal cells that form trichomes, visible on the larval cuticle. *In situ* hybridization shows mRNA expression of two *svb* downstream genes, *shavenoid (sha/koj)* and *CG15589*, in wild type (top) and *svb* mutant embryos (bottom). **B**: ROC curve showing significant enrichment in putative Svb binding sites (OvoQ6 position weight matrix) among the 39 Svb downstream genes (y axis) compared to a randomized set of 1000 *Drosophila* genes (x axis) using cisTargetX. The blue curve shows the detection of Svb downstream genes, the red one a random distribution, and the green curve shows a 2 sigma interval from random.

Figure 2: A subset of Svb binding sites correspond to functional enhancers

Svb-dependent trichome enhancers were identified by transgenic reporter gene assays, from a systematic scanning of the *singed (sn)* (A) and *shavenoid (sha)* (B) genes and regions predicted by cisTargetX (C). **A,B**. Vertical black lines represent evolutionarily conserved OvoQ6 clusters (at least two motifs in a 1kb window), as predicted by cistargetX. Horizontal boxes summarize regions tested by transgenic assays, using immunostaining of the lacZ reporter (green). Negative regions (pink) do not drive specific expression in the embryonic epidermis. Regions drawn in cyan display enhancer activity reproducing endogenous expression in trichome cells (as assayed by mRNA *in situ* hybridization, purple). The *snE1* and *sha3* enhancers are under the control of Svb, as demonstrated by a reduced expression in *svb* mutants. **C**. Putative enhancers (CRMs) predicted by cis-TargetX, from clustering and/or evolutionary conservation of OvoQ6 sites. Pictures show expression of positive enhancers (cyan) in wild type (top) and *svb* mutant (bottom) embryos. Additional regions (pink) showed no detectable activity during embryogenesis.

Figure 3: Computational analyses allow refining functional Svb binding sites

A. Statistical analysis of positive enhancers versus negative regions (intergenic genomic sequences used as background) was performed for *de novo* discovery of motifs, showing evolutionary conservation across *Drosophila* species and characteristic of active enhancers. **B.** The svbF7 and blue motifs perform best in discriminating between positive enhancers and negative regions as illustrated by the Pareto plot. **C**. While disruption of the core CnGTT OvoQ6 motif abolish *Emin* activity, point mutations that affect the 5' and 3' flanking nucleotides strongly reduce epidermal expression, as shown by anti-lacZ immuno-staining and quantification of fluorescence signals.

Figure 4: *In vivo* role of svbF7 motifs in Svb-dependent enhancers

Anti-LacZ staining (green) shows modifications of reporter gene expression resulting from individual and simultaneous inactivation of svbF7 motifs in *sha1* (**A**), *nyo1* (**B**), *tyn2* (**C**), *sha3* (**D**) and *dyl2* (**E**) enhancers. **F-H** Quantification of residual activity following individual disruption of svbF7 sites. Red boxes schematize evolutionarily conserved svbF7 motifs; the open black box, a site that does not appear conserved across *drosophilidae*; *** indicates a p-value <0,01, *<0,05.

Figure 5: Svb-dependent enhancers use various combinations of cis-regulatory elements

A. Linker-scanning mutagenesis of *Emin* identifies other 10bp regions required for full transcriptional activity, as deduced from altered pattern of lac-Z immuno-staining (green). Positions of SvbF7, blue and yellow motifs are indicated on top. **B**. Black, red, blue and yellow boxes schematize the distribution, number and orientation of OvoQ6, svbF7, blue and yellow motifs, respectively. Filled boxes represent motifs conserved across *Drosophila* species; open boxes those detected only in *D. melanogaster*. **C**. Point mutations that disrupt the blue motif in *Emin*, *snE1* and *sha3* reduce the activity of all three enhancers, to 40 +/-14

% (*), 44 +/- 6% (***), 8 +/- 4% (***) of wild-type levels, respectively. **D**. Point mutations that disrupt the yellow motif in *Emin*, *17058* and *nyo1* reduce the activity of all three enhancers, to 20 +/- 7 % (*), 16 +/- 8% (*), 6 +/- 3% (*) of wild-type levels, respectively. *** indicates a p-value <0,001, *<0,03.

Figure 6: Genome-wide profiling of embryonic genes regulated by Svb

A. Cross-correlation between conserved svbF7, blue or yellow motif instances and Svb ChIP-Seq peaks throughout the whole genome. Plots show the number of motifs found in a 10kb window on each side of the center of peaks. The p-value for correlation (Chi2 test) is <1E-46, <1E-9 and <1E-2, respectively. **B**. Modifications in mRNA levels as measured by microarrays between wild-type and *svb* (left) or *pri* (right) embryos in Svb-regulated (green) and control (blue) set of genes. Dark green dots represent known Svb targets (Fig. S1A), light green novel target genes as validated by *in situ* hybridization (Fig. S4) and open dots additional candidates. **C**. Whole mount *in situ* hybridization of *CG1273*, a Svb downstream target identified from microarray profiling, down regulated in trichome cells of *svb* mutants and showing a further reduced expression in *pri* mutant embryos.

Figure 7: Identification of Svb direct targets and their trichome enhancers using computational and *in vivo* experimental approaches

A. Chart flow diagram summarizing the pipeline used for enhancer prediction and validation. **B.** Motif distribution coupled to ChIP-seq allows predicting location of enhancers in Svb downstream targets. Graphs show ChIP intensity at the time of trichome formation (12-14h of embryogenesis). Active enhancers are drawn as cyan rectangles. Pictures show reporter gene expression driven by corresponding regions in wild type (*wt*) and *svb* mutant embryos, as revealed by anti-lacZ immunostaining (green). The composition, orientation and respective positioning of svbF7 (red), blue and yellow motifs is schematized by filled (evolutionarily conserved) and open (not traceable across species) boxes.

Figure 8. Pattern of evolutionary conservation of the three enhancers driving *miniature* expression in trichome cells

The position of epidermal enhancers is shown by cyan boxes and their respective architecture with respect to svbF7 (red), blue and yellow motifs are schematized across *Drosophila* species. Orthologous sequences were identified by BLAST and manually adjusted for optimized alignment. Motif search was performed in individual sequences taken independently, using a same threshold for each motif in all cases. Bottom histograms represent the pattern of evolutionary conservation across *Drosophila* species, focusing on individual regions harboring identified cis-regulatory motifs (color coded).

REFERENCES

1. Stathopoulos A, Levine M: **Genomic regulatory networks and animal development.** *Dev Cell* 2005, **9:**449-462.

2. Ptashne M: **Regulation of transcription: from lambda to eukaryotes.** *Trends Biochem Sci* 2005, **30:**275-279.

3. Rister J, Desplan C: **Deciphering the genome's regulatory code: the many languages of DNA.** *Bioessays* 2010, **32:**381-384.

4. Yanez-Cuna JO, Kvon EZ, Stark A: **Deciphering the transcriptional cis-regulatory code.** *Trends Genet* 2013, **29:**11-22.

5. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al: **Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm.** *PLoS Biol* 2008, **6:**e27.

6. Slattery M, Negre N, White KP: **Interpreting the regulatory genome: the genomics of transcription factor function in Drosophila melanogaster.** *Brief Funct Genomics* 2012, **11:**336-346.

7. MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keranen SV, et al: **Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions.** *Genome Biol* 2009, **10:**R80.

8. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al: **A cis-regulatory map of the Drosophila genome.** *Nature* 2011, **471:**527-531.

9. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U: **Predicting expression patterns from regulatory sequence in Drosophila segmentation.** *Nature* 2008, **451:**535-540.

10. Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, Furlong EE: **A transcription factor collective defines cardiac cell fate and reflects lineage history.** *Cell* 2012, **148:**473-486.

11. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE: **Combinatorial binding predicts spatio-temporal cis-regulatory activity.** *Nature* 2009, **462:**65-70.

12. Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M: **Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo.** *Genes Dev* 2007, **21:**385-390.

13. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.** *Proc Natl Acad Sci U S A* 2002, **99:**763-768.

14. Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED: **Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila.** *BMC Bioinformatics* 2004, **5:**129.

15. He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J: **High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species.** *Nat Genet* 2011, **43:**414-420.

16. Markstein M, Zinzen R, Markstein P, Yee KP, Erives A, Stathopoulos A, Levine M: **A regulatory code for neurogenic gene expression in the Drosophila embryo.** *Development* 2004, **131:**2387-2394.

17. Khoueiry P, Rothbacher U, Ohtsuka Y, Daian F, Frangulian E, Roure A, Dubchak I, Lemaire P: **A cis-regulatory signature in ascidians and flies, independent of transcription factor binding sites.** *Curr Biol* 2010, **20:**792-802.

18. Erives A, Levine M: **Coordinate enhancers share common organizational features in the Drosophila genome.** *Proc Natl Acad Sci U S A* 2004, **101:**3851-3856.

19. Levine M: **Transcriptional enhancers in animal development and evolution.** *Curr Biol* 2010, **20:**R754-763.

20. Spitz F, Furlong EE: **Transcription factors: from enhancer binding to developmental control.** *Nat Rev Genet* 2012, **13:**613-626.

21. Crocker J, Erives A: **A closer look at the eve stripe 2 enhancers of Drosophila and Themira.** *PLoS Genet* 2008, **4:**e1000276.

22. Papatsenko D, Goltsev Y, Levine M: **Organization of developmental enhancers in the Drosophila embryo.** *Nucleic Acids Res* 2009, **37:**5665-5677.

23. Swanson CI, Evans NC, Barolo S: **Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer.** *Dev Cell* 2010, **18:**359-370.

24. Rowan S, Siggers T, Lachke SA, Yue Y, Bulyk ML, Maas RL: **Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity.** *Genes Dev* 2010, **24:**980-985.

25. Papatsenko D, Levine M: **A rationale for the enhanceosome and other evolutionarily constrained enhancers.** *Curr Biol* 2007, **17:**R955-957.

26. Crocker J, Tamori Y, Erives A: **Evolution acts on enhancer organization to fine-tune gradient threshold readouts.** *PLoS Biol* 2008, **6:**e263.

27. Etchberger JF, Lorch A, Sleumer MC, Zapf R, Jones SJ, Marra MA, Holt RA, Moerman DG, Hobert O: **The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron.** *Genes Dev* 2007, **21:**1653-1674.

28. Laurencon A, Dubruille R, Efimenko E, Grenier G, Bissett R, Cortier E, Rolland V, Swoboda P, Durand B: **Identification of novel regulatory factor X (RFX) target genes by comparative genomics in Drosophila species.** *Genome Biol* 2007, **8:**R195.

29. Payre F, Vincent A, Carreno S: **ovo/svb integrates Wingless and DER pathways to control epidermis differentiation.** *Nature* 1999, **400:**271-275.

30. Sucena E, Delon I, Jones I, Payre F, Stern DL: **Regulatory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism.** *Nature* 2003, **424:**935-938.

31. Chanut-Delalande H, Fernandes I, Roch F, Payre F, Plaza S: **Shavenbaby couples patterning to epidermal cell shape control.** *PLoS Biol* 2006, **4:**e290.

32. Chanut-Delalande H, Ferrer P, Payre F, Plaza S: **Effectors of tridimensional cell morphogenesis and their evolution.** *Semin Cell Dev Biol* 2012, **23:**341-349.

33. Fernandes I, Chanut-Delalande H, Ferrer P, Latapie Y, Waltzer L, Affolter M, Payre F, Plaza S: **Zona pellucida domain proteins remodel the apical compartment for localized cell shape changes.** *Dev Cell* 2010, **18:**64-76.

34. Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y: **Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis.** *Science* 2010, **329:**336-339.

35. Lee S, Garfinkel MD: **Characterization of Drosophila OVO protein DNA binding specificity using random DNA oligomer selection suggests zinc finger degeneration.** *Nucleic Acids Res* 2000, **28:**826-834.

36. Lu J, Oliver B: **Drosophila OVO regulates ovarian tumor transcription by binding unusually near the transcription start site.** *Development* 2001, **128:**1671-1686.

37. Brown CD, Johnson DS, Sidow A: **Functional architecture and evolution of transcriptional elements that drive gene coexpression.** *Science* 2007, **317:**1557-1560.

38. Andrew DJ, Baker BS: **Expression of the Drosophila secreted cuticle protein 73 (dsc73) requires Shavenbaby.** *Dev Dyn* 2008, **237:**1198-1206.

39. Bejsovec A, Chao AT: **crinkled reveals a new role for Wingless signaling in Drosophila denticle formation.** *Development* 2012, **139:**690-698.

40. Aerts S, Quan XJ, Claeys A, Naval Sanchez M, Tate P, Yan J, Hassan BA: **Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in Drosophila uncovers a regulatory basis for sensory specification.** *PLoS Biol* 2010, **8:**e1000435.

41. Elemento O, Tavazoie S: **Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach.** *Genome Biol* 2005, **6:**R18.

42. Mace KA, Pearson JC, McGinnis W: **An epidermal barrier wound repair pathway in Drosophila is mediated by grainy head.** *Science* 2005, **308:**381-385.

43. Szuplewski S, Kottler B, Terracol R: **The Drosophila bZIP transcription factor Vrille is involved in hair and cell growth.** *Development* 2003, **130:**3651-3662.

44. Jayo A, Parsons M: **Fascin: a key regulator of cytoskeletal dynamics.** *Int J Biochem Cell Biol* 2012, **42:**1614-1617.

45. Frith MC, Li MC, Weng Z: **Cluster-Buster: Finding dense clusters of motifs in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3666-3668.

46. Kim J, Cunningham R, James B, Wyder S, Gibson JD, Niehuis O, Zdobnov EM, Robertson HM, Robinson GE, Werren JH, Sinha S: **Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances.** *PLoS Comput Biol* 2010, **6:**e1000652.

47. Rouault H, Mazouni K, Couturier L, Hakim V, Schweisguth F: **Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny.** *Proc Natl Acad Sci U S A* 2010, **107:**14615-14620.

48. Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, Brasefield JA, Zhu C, Asriyan Y, Lapointe DS, et al: **FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system.** *Nucleic Acids Res* 2011, **39:**D111-117.

49. Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL: **Morphological evolution caused by many subtle-effect substitutions in regulatory DNA.** *Nature* 2011, **474:**598-603.

50. Herrmann C, Van de Sande B, Potier D, Aerts S: **i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules.** *Nucleic acids research* 2012, **40:**e114.

51. Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J: **RSAT 2011: regulatory sequence analysis tools.** *Nucleic acids research* 2011, **39:**W86-91.

52. Payre F: **Genetic control of epidermis differentiation in Drosophila.** *Int J Dev Biol* 2004, **48:**207-215.

53. Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL: **Phenotypic robustness conferred by apparently redundant transcriptional enhancers.** *Nature* 2010, **466:**490-493.

54. Perry MW, Boettiger AN, Bothma JP, Levine M: **Shadow enhancers foster robustness of Drosophila gastrulation.** *Curr Biol* 2010, **20:**1562-1567.

55. Hong JW, Hendrix DA, Levine MS: **Shadow enhancers as a source of evolutionary novelty.** *Science* 2008, **321:**1314.

56. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci U S A* 2002, **99:**757-762.

57. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U: **Transcriptional control in the segmentation gene network of Drosophila.** *PLoS Biol* 2004, **2:**E271.

58. Chen H, Xu Z, Mei C, Yu D, Small S: **A system of repressor gradients spatially organizes the boundaries of Bicoid-dependent target genes.** *Cell* 2012, **149:**618-629.

59. Aerts S: **Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets.** *Curr Top Dev Biol* 2012, **98:**121-145.

60. Halfon MS, Zhu Q, Brennan ER, Zhou Y: **Erroneous attribution of relevant transcription factor binding sites despite successful prediction of cis-regulatory modules.** *BMC Genomics* 2011, **12:**578.

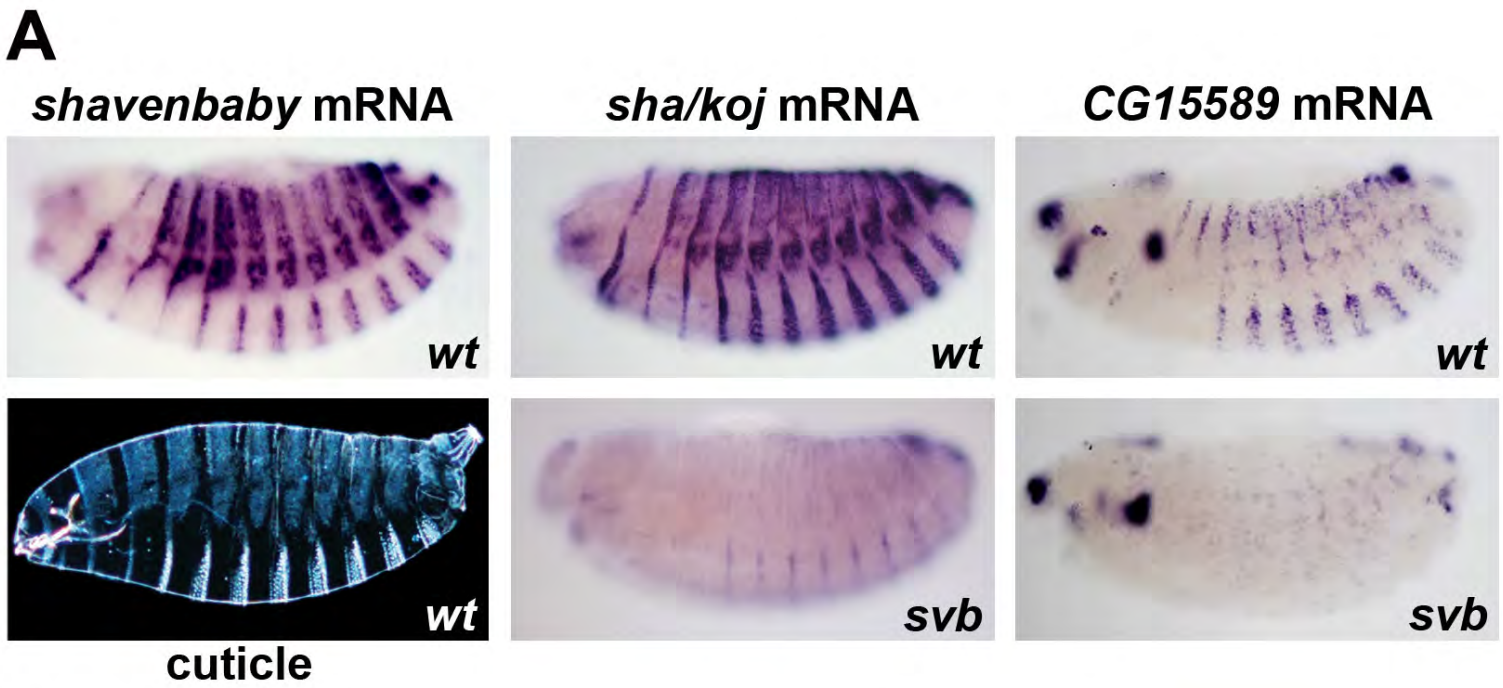
61. Walter J, Biggin MD: **DNA binding specificity of two homeodomain proteins in vitro and in Drosophila embryos.** *Proc Natl Acad Sci U S A* 1996, **93:**2680-2685.

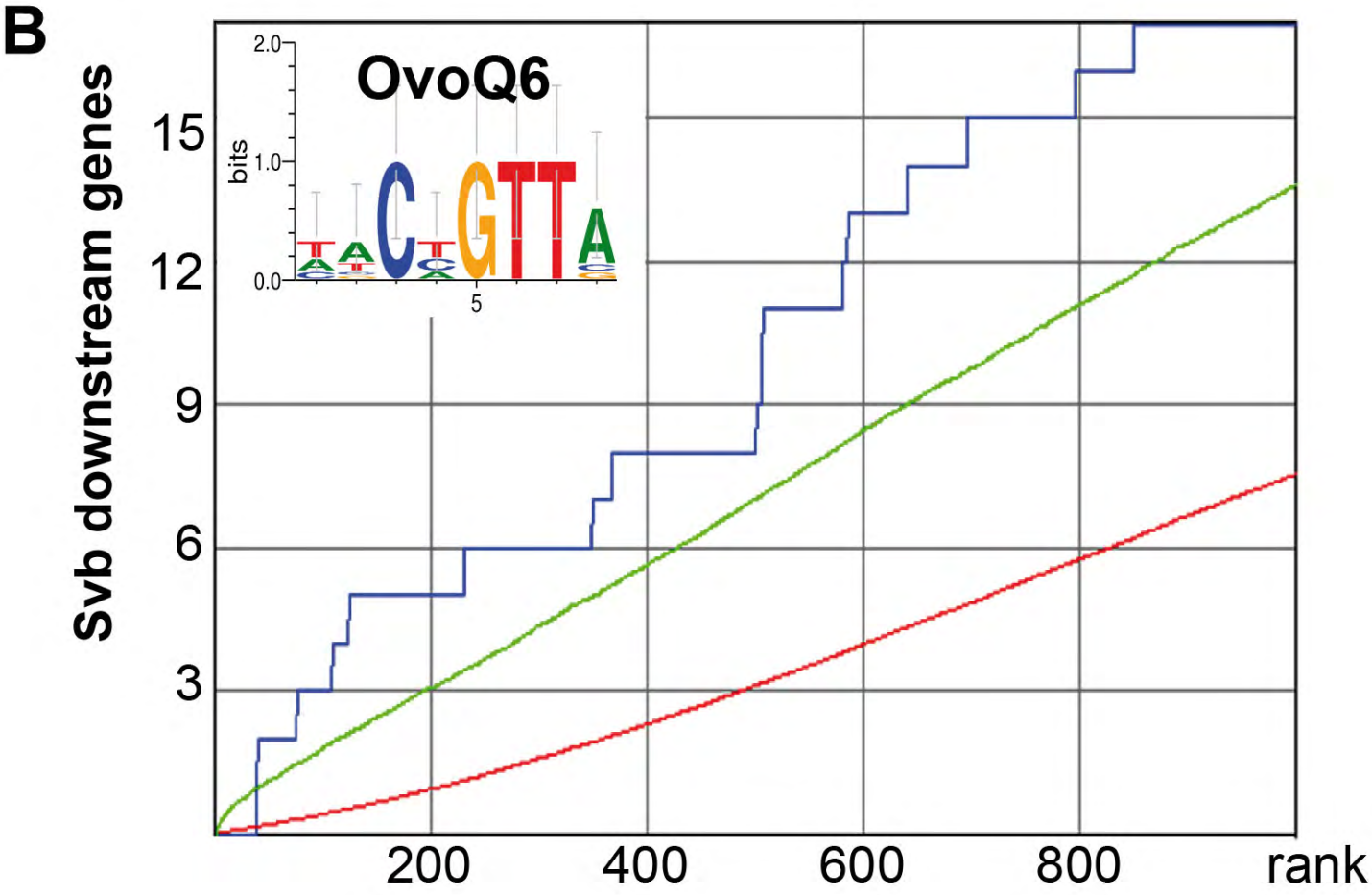
62. Ozdemir A, Fisher-Aylor KI, Pepke S, Samanta M, Dunipace L, McCue K, Zeng L, Ogawa N, Wold BJ, Stathopoulos A: **High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation.** *Genome Res* 2011, **21:**566-577.

63. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, Mann RS: **Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins.** *Cell* 2011, **147:**1270-1282.

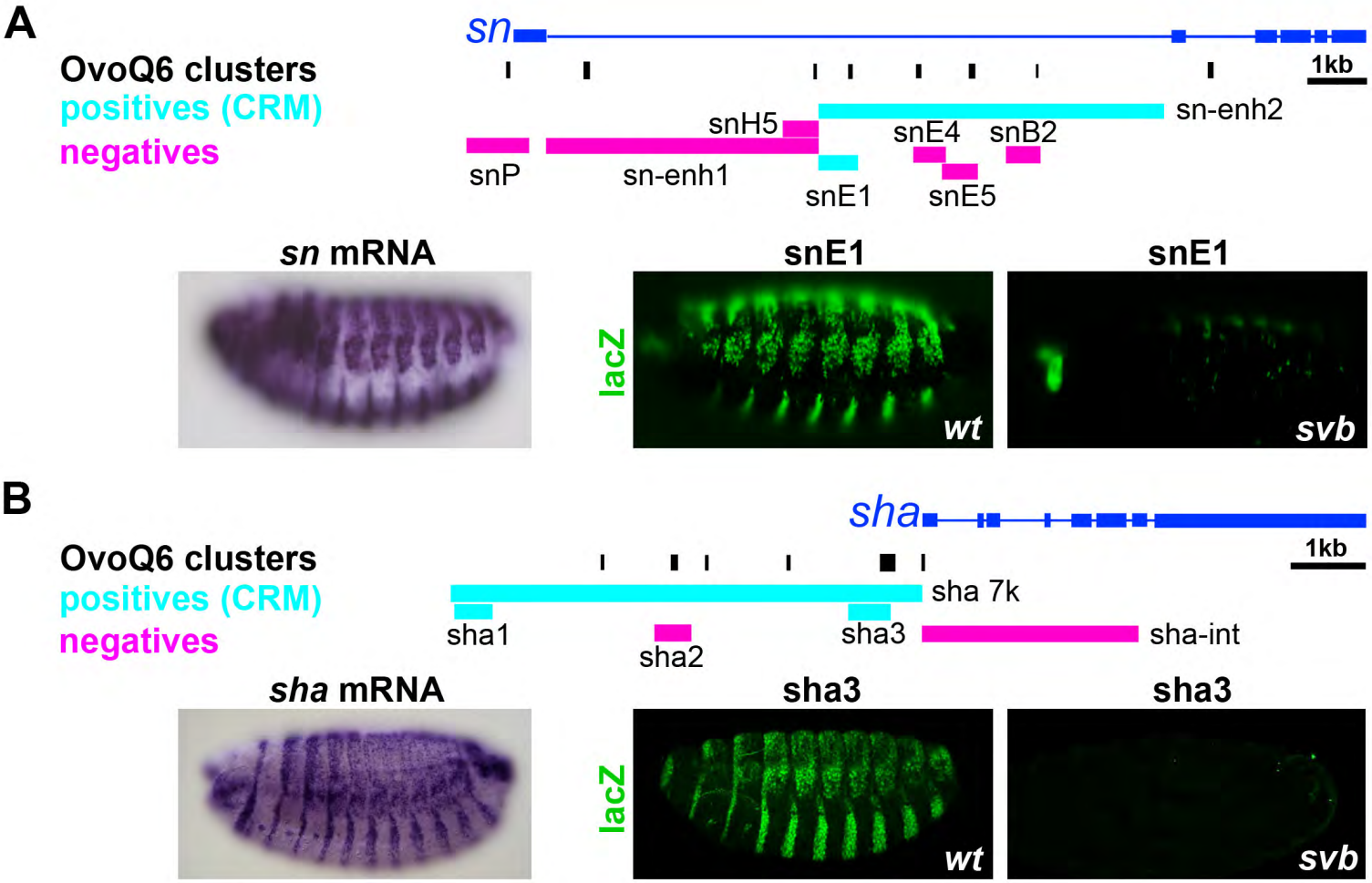
64. Sorge S, Ha N, Polychronidou M, Friedrich J, Bezdan D, Kaspar P, Schaefer MH, Ossowski S, Henz SR, Mundorf J, et al: **The cis-regulatory code of Hox function in Drosophila.** *EMBO J* 2012, **31:**3323-3333.

65. Jin H, Stojnic R, Adryan B, Ozdemir A, Stathopoulos A, Frasch M: **Genome-wide screens for in vivo tinman binding sites identify cardiac enhancers with diverse functional architectures.** *PLoS Genet* 2013, **9:**e1003195.



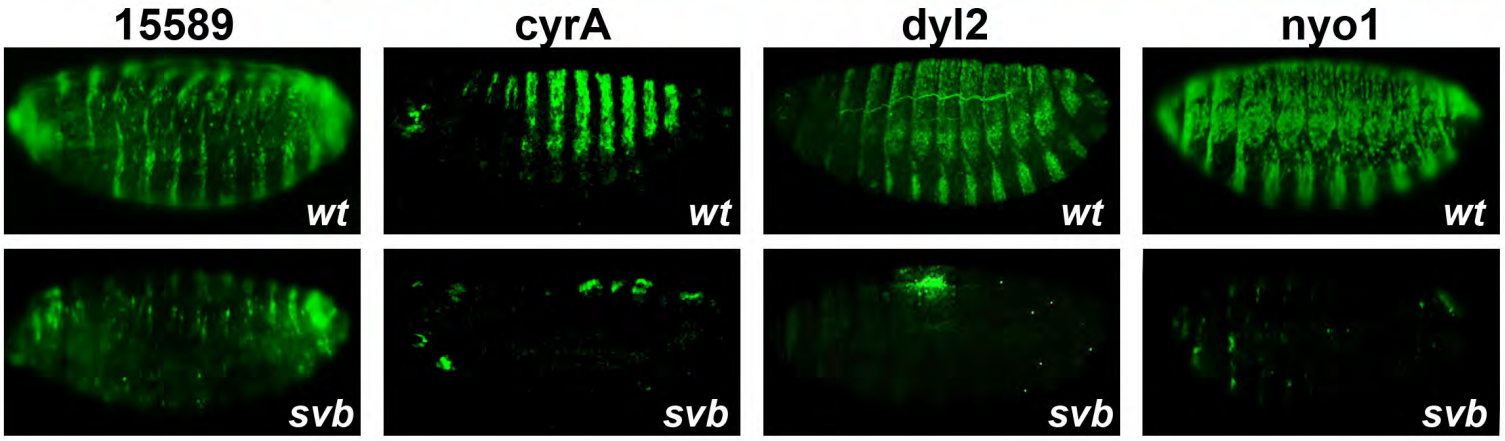


Menoret *et al.*; Figure 1

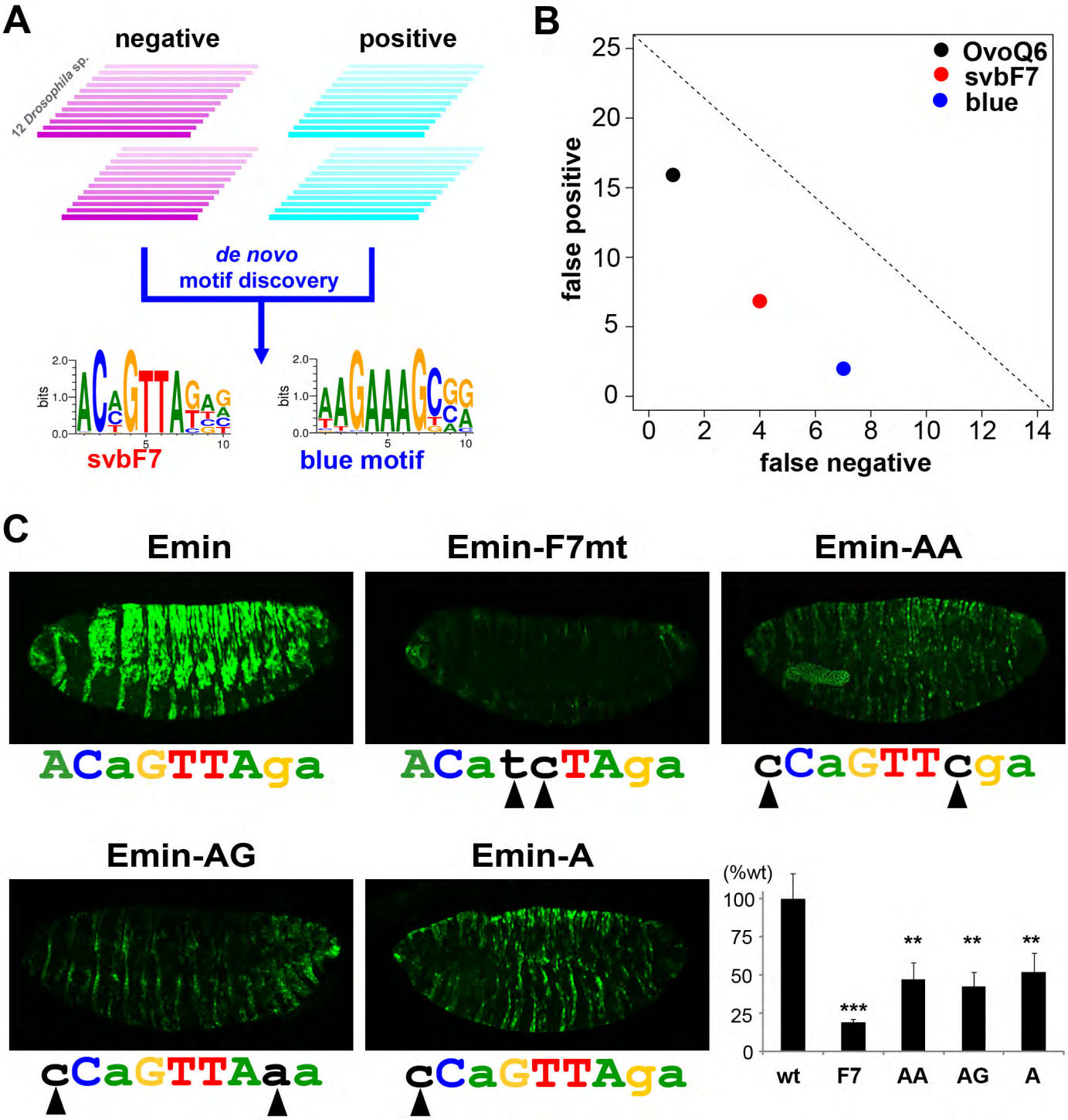


C

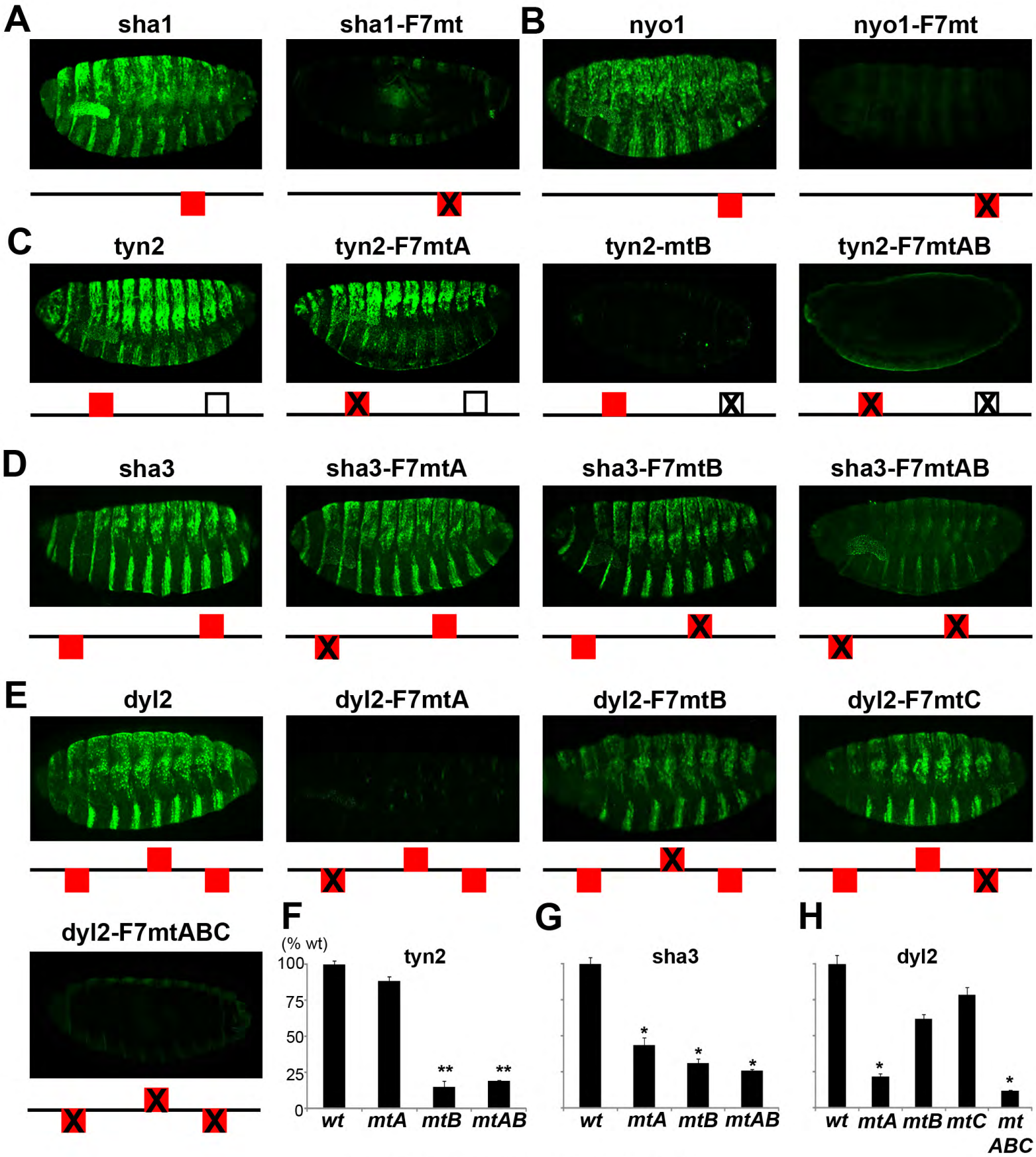
CisTargetX predicted regions	dyl1	15589	cyrA	snE5	dyl2	f2	sox21b	sn-enh1	sha3	snH5	snB2	nyo3	snE1	sha-int	f5	sha2	f4	cyrB	snP	nyo1	snE4
RANK	1	3	4	5	8	9	13	21	23	24	32	38	43	50	55	63	73	77	79	80	82
ACTIVITY	-	+	+	-	+	-	-	-	+	-	-	-	+	-	-	-	-	-	-	+	-



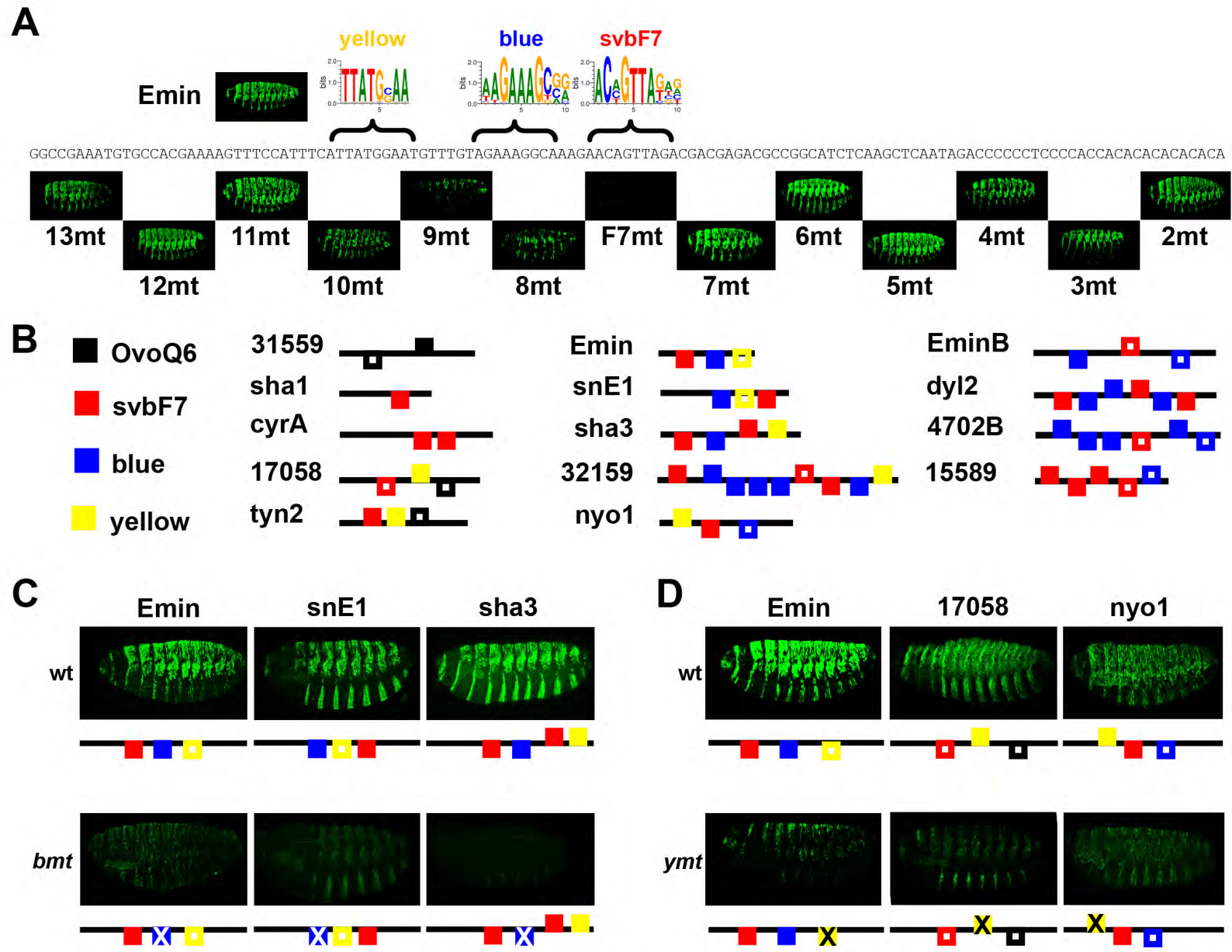
Menoret *et al.;* Figure 2



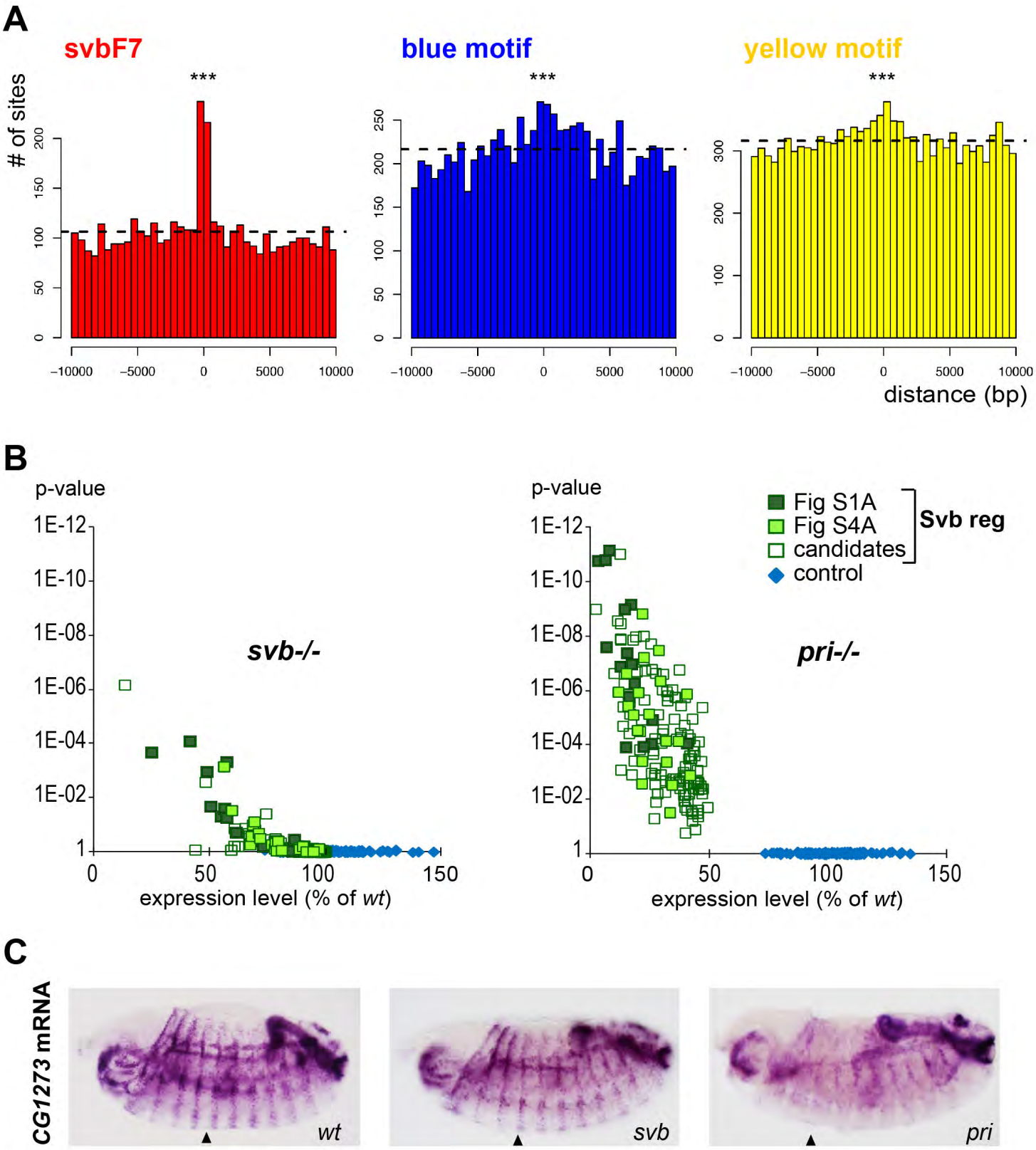
Menoret *et al.;* Figure 3



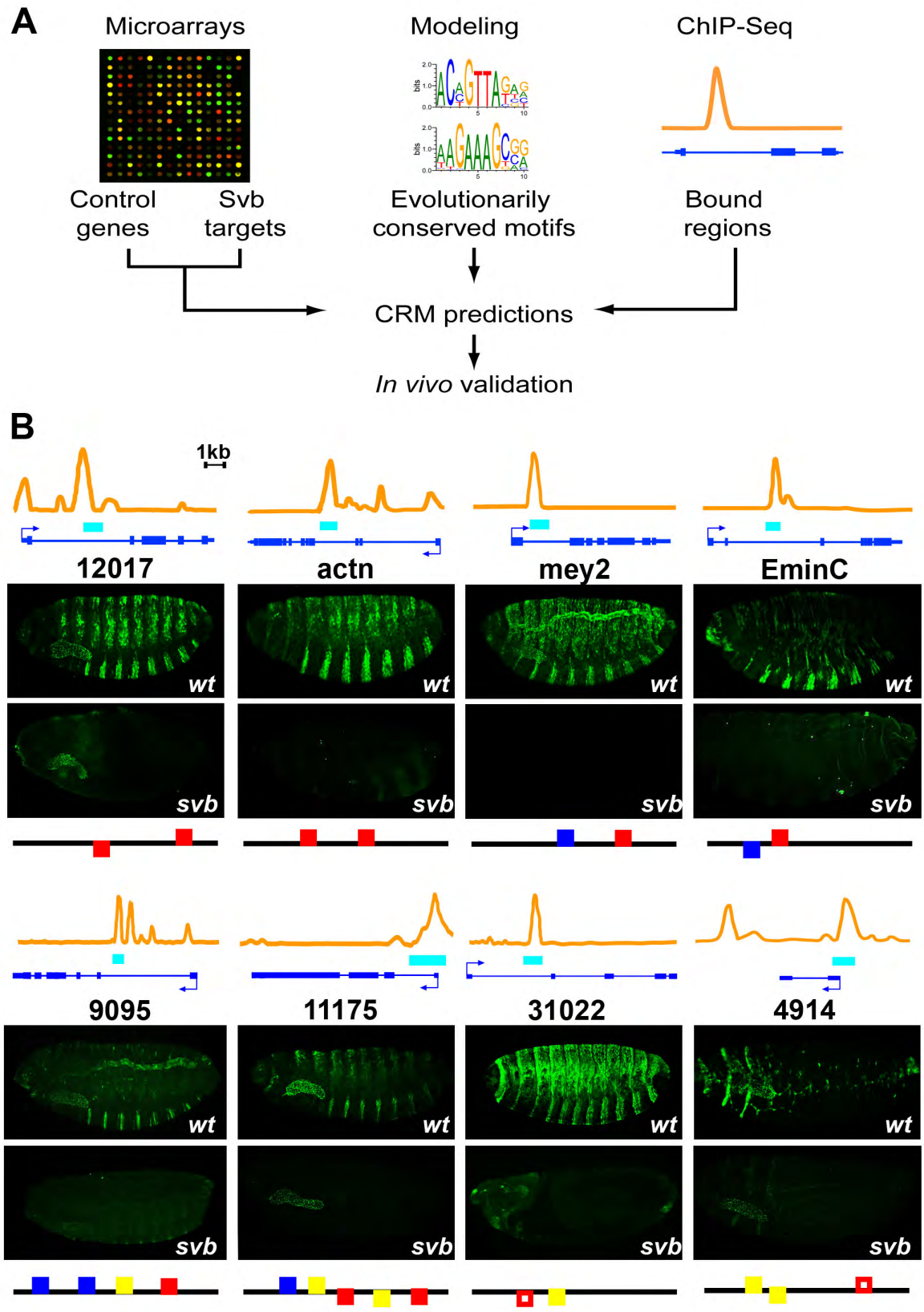
Menoret *et al.;* Figure 4



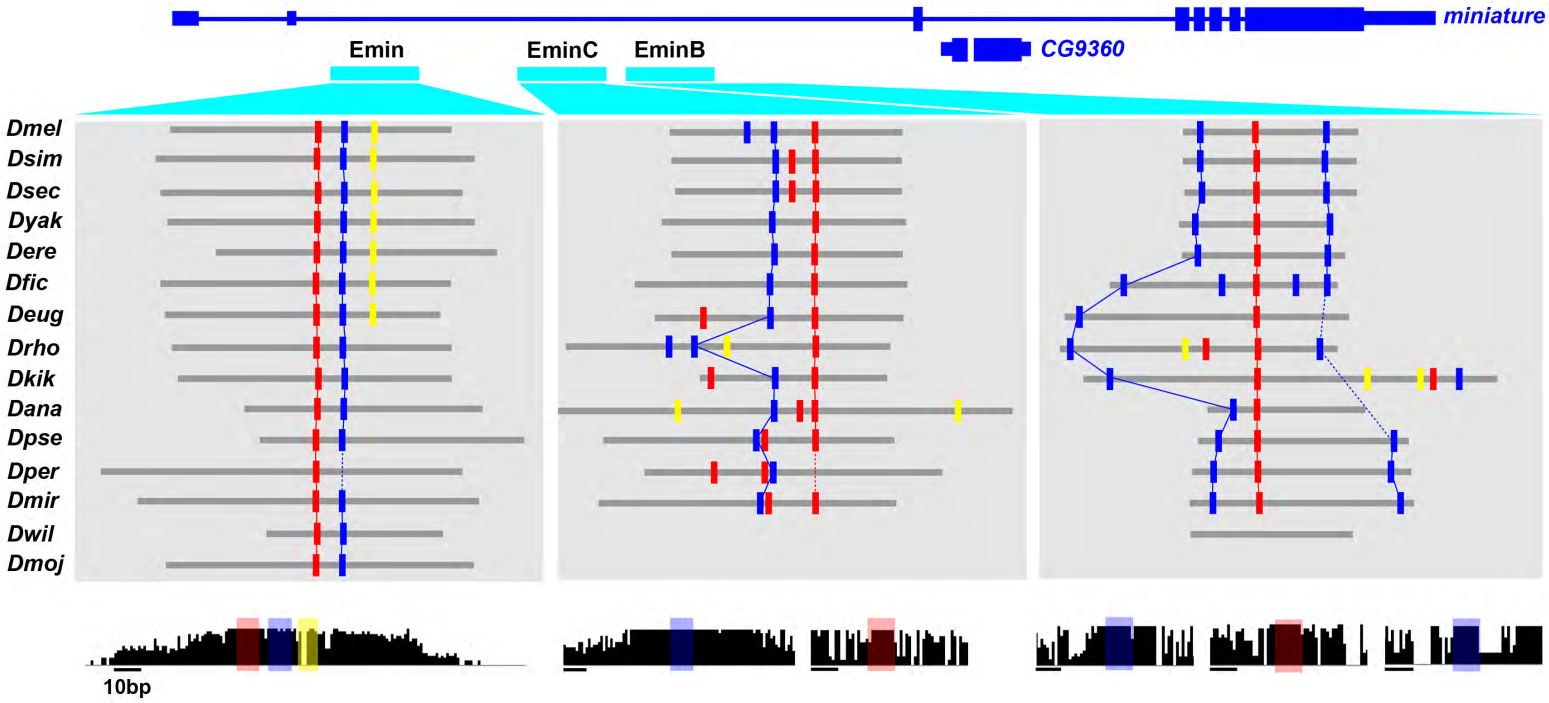
Menoret *et al.;* Figure 5



Menoret *et al.*; Figure 6



Menoret *et al.*; Figure 7



Menoret *et al.*; Figure 8

SUPPLEMENTARY INFORMATIONS

1. Genomic analyses and statistical Methods

De novo motif generation

We used the phylogeny-based *de novo* motif generation algorithm described in Rouault et al. (2010), available on the website (https://github.com/hrouault/Imogene/). The 14 positive CRMs were used as the training set for the algorithm and scanned for conserved motifs as described in (Rouault et al, 2010). The score threshold for motif generation, which sets the searched PWM information content, was varied from 7 to 13 bits in different runs of the algorithm, with a motif width set to 10 bp. In each run, the 5 highest scoring motifs were kept. This resulted in a large number of different motifs. In order to find the most discriminative ones, the 27 negative CRM were used as a negative set.

Positive and negative sets were used to evaluate the False Negative Rate (FNR) and False Positive Rate (FPR) for all motifs generated by the algorithm. For each motif, the two sets were scanned for conserved instances with a scanning threshold varied between 7 and 13 bits. For each threshold, FPR and FNR were computed as the proportion of Positive (resp. Negative) CRMs with at least one conserved instance for the motif with a score higher than the threshold. The best motifs, shown as red and blue dots, were selected based on the minimization of both FPR and FPR in a Pareto plot, as shown in Fig. 3B. These motifs were generated with a threshold of 10.1 bits and were scanned with optimal thresholds of 10.1 and 8.7 bits respectively.

Genome-wide ranking of enhancers and genes

In order to rank enhancers genome wide, we followed the method presented in Rouault et al, (2010). Coding sequences as well as the training set used for motif generation were masked. Conserved instances of *de novo* svbf7 and blue motifs at optimal threshold were then determined genome wide. Genomic fragment of 1Kbp were scored according to the additive Poisson score introduced in (Rouault et al., 2010) using the negative enhancers as a background set of intergenic fragments. Around each determined motif instance, the optimal scoring 1Kbp genomic fragment was defined as a putative enhancer. Each putative enhancer was associated to the nearest gene transcription start site. Each gene was attributed the highest score among its associated enhancers, or 0 if it had no associated enhancers.

Statistical analyses

To test for putative enrichment in a given motif between Svb-regulated and control set of genes (fig S2B), we used a Mann-Whitney U test using the transcribed region of each gene extended to 5 kb flanking sequences. A p-value was computed using the function wilcox.test from the R stats package. For motif *vs* ChIPseq cross-correlations (Fig. 6A & S5), we performed a c^2-test to disentangle cross-correlation signals from small number fluctuations. Correlation data were binned in 500bp elements in a +/-10 kb region around the center of each ChIP peak, resulting in k=40 bins. A c^2 was computed as the sum over the bins of the standardized counts S_i $(O_i - E)^2$ / E, where O_i represents the

observed count in bin i and E is the expected number of counts in a bin, taken to be uniform over the considered region. Finally, a p-value was computed as the probability that a c^2 statistics with k-1 degrees of freedom takes at least the observed value.

ChIP-seq analyses

Sequence data was analyzed using a virtual machine image on the Bionimbus cloud (http://www.bionimbus.org/) and aligned to the *D. melanogaster* genome using Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) (Langmead, 2009). Sequence density along the genome was visualized using wig files generated with SPP (Kharchenko, 2008) and sequence enrichment along the genome was defined by MACS with the following parameters: tag size=36, bandwidth=100, Pvalue=1e-5 (Zhang, 2008).
ChIP peaks were subjected to motifs detection using *i-cis*targetX (http://med.kuleuven.be/lcb/i-cisTarget/) (Herrmann et al, 2012) and Peak motif (RSA tools) (http://rsat.ulb.ac.be/peak-motifs_form.cgi) (Thomas-Chollier et al. 2012).

References

Hermann C, Van de Sande B, Potier D, Aerts S (2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. Nucleic Acids Res 2012 40(15): e114.

Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26: 1351-1359.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

Rouault H, Mazouni K, Couturier L, Hakim V, Schweisguth F (2010) Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. Proc Natl Acad Sci U S A 107: 14615-14620.

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res. 40(4): e31.

Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat Protoc 7(8): 1551-1568.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

2. List of transgenic constructs

name	*gene*	position	transgene
11175	*Rcd6*	chr2R 16519573 16521416	PhiC31
12017	*CG12017*	chr3L 3283270 3284580	PhiC31
12017-2	*CG12017*	chr3L 3279813 3281368	PhiC31
12063	*morpheyus*	chr3R 27320842 27321073	P-element
12063-2R	*morpheyus*	chr3R 27320842 27321073	PhiC31
14395-2	*CG14395*	chr3R 8494910 8496045	PhiC31
1499-1	*nyobe*	chr3R 27370295 27370415	P-element
15013-1	*dusky like*	chr3L 4299768 4299927	P-element
15013-2	*dusky like*	chr3L 4300420 4300715	P-element
15589	*CG15589*	chr3R 2027050 2028050	PhiC31
17058	*Peritrophin-A*	chrX 20113835 20115016	PhiC31
17058ymt	*Peritrophin-A*	chrX 20113835 20115016	PhiC31
31022	*PH4alphaEFB*	chr3R 26297285 26298528	PhiC31
31559	*CG31559*	chr3R 1977200 1978200	PhiC31
32159	*dsx-c73A*	chr3L 16435341 16436341	PhiC31
32356	*ImpE1*	chr3L 8370153 8371153	PhiC31
4702	*CG4702*	chr3R 7950428 7950586	P-element
4702B	*CG4702*	chr3R 7951250 7952250	PhiC31
4914	*CG4914*	chr3L 14670764 14672083	PhiC31
9095	*CG9095*	chrX 15046901 15047823	PhiC31
Actn	*actinin*	chrX 1925087 1927414	PhiC31
cyrA	*cypher*	chrX 8019397 8020397	PhiC31
cyrB	*cypher*	chrX 8017597 8018797	PhiC31
dyl1	*dusky like*	chr3L 4303968 4304768	PhiC31
dyl2	*dusky like*	chr3L 4298268 4299468	PhiC31
dyl2F7mtA	*dusky like*	chr3L 4298268 4299468	PhiC31
dyl2F7mtABC	*dusky like*	chr3L 4298268 4299468	PhiC31
dyl2F7mtB	*dusky like*	chr3L 4298268 4299468	PhiC31
dyl2F7mtC	*dusky like*	chr3L 4298268 4299468	PhiC31
dyl3	*dusky like*	chr3L 4305468 4306468	PhiC31
Emin	*miniature*	chrX 11650982 11651144	PhiC31
Emin10mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin11mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin12mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin13mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin2mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin3mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin4mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin5mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin6mt	*miniature*	chrX 11650982 11651144	PhiC31

Emin7mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin8mt	*miniature*	chrX 11650982 11651144	PhiC31
Emin9mt	*miniature*	chrX 11650982 11651144	PhiC31
EminA	*miniature*	chrX 11654097 11655097	PhiC31
EminAA	*miniature*	chrX 11654097 11655097	PhiC31
EminAG	*miniature*	chrX 11654097 11655097	PhiC31
EminB	*miniature*	chrX 11654097 11655097	PhiC31
Eminbmt	*miniature*	chrX 11650982 11651144	PhiC31
EminC	*miniature*	chrX 11652670 11654154	PhiC31
EminF7mt	*miniature*	chrX 11650982 11651144	PhiC31
Eminflkmt	*miniature*	chrX 11650982 11651144	PhiC31
Eminymt	*miniature*	chrX 11650982 11651144	PhiC31
f1	*forked*	chrX 17153478 17154756	P-element
f2	*forked*	chrX 17159378 17160246	P-element
f4	*forked*	chrX 17162096 17163096	PhiC31
f5	*forked*	chrX 17158996 17160096	PhiC31
mey2 (12063-2)	*morpheyus*	chr3R 27325605 27326605	PhiC31
Neyo	*neyo*	chr3R 25647300 25648300	P-element
nyo1	*nyobe*	chr3R 27384231 27384921	PhiC31
nyo1F7mt	*nyobe*	chr3R 27384231 27384921	PhiC31
nyo1ymt	*nyobe*	chr3R 27384231 27384921	PhiC31
nyo2	*nyobe*	chr3R 27381479 27382574	PhiC31
nyo3	*nyobe*	chr3R 27377275 27378274	PhiC31
sha-int	*shavenoïd*	chr2R 7216771 7220065	P-element
sha1	*shavenoid*	chr2R 7209659 7210257	P-element
sha1F7mt	*shavenoid*	chr2R 7209659 7210257	P-element
sha2	*shavenoïd*	chr2R 7212709 7213256	P-element
sha2-2R	*shavenoïd*	chr2R 7212709 7213256	PhiC31
sha3	*shavenoid*	chr2R 7215630 7216294	P-element
sha3bmt	*shavenoid*	chr2R 7215630 7216294	P-element
sha3F7mtA	*shavenoid*	chr2R 7215630 7216294	P-element
sha3F7mtAB	*shavenoid*	chr2R 7215630 7216294	P-element
sha3F7mtB	*shavenoid*	chr2R 7215630 7216294	P-element
sn-enh1	*singed*	chrX 7864407 7869657	P-element
snB2	*singed*	chrX 7873257 7873915	PhiC31
snE1	*Singed*	chrX 7869678 7870390	PhiC31
snE1	*Singed*	chrX 7869678 7870390	P-element
snE1bmt	*Singed*	chrX 7869678 7870390	P-element
snE1F7mt	*Singed*	chrX 7869678 7870390	PhiC31
snE4	*singed*	chrX 7871432 7872096	P-element
snE5	*singed*	chrX 7871996 7872659	P-element
snH5	*singed*	chrX 7868528 7868978	P-element
snP	*singed*	chrX 7862910 7864103	P-element

sox21b	*sox21b*	chr3L 14121641 14122852	PhiC31
tyn1	*trynity*	chrX 86343 87613	PhiC31
tyn2	*trynity*	chrX 77484 78384	PhiC31
tyn2F7mtA	*trynity*	chrX 77484 78384	PhiC31
tyn2F7mtAB	*trynity*	chrX 77484 78384	PhiC31
tyn2mtB	*trynity*	chrX 77484 78384	PhiC31

3. Microarray procedures

Biotinylated cRNA targets were prepared, starting from 200 ng of total RNA, using the MessageAmp™ Premier RNA Amplification Kit (Ambion CAT# AM1792), according to the manufacturer recommendations. Following fragmentation, 6.5 µg of cRNAs were hybridized for 16 hours at 45°C on GeneChip® Drosophila Genome 2.0 Array interrogating over 18,500 transcripts (Affymetrix, Santa Clara, CA). The chips were washed and stained using the GeneChip® Fluidics Station 450 and scanned using the GeneChip® Scanner 3000 7G according to Affymetrix recommendations. Raw data (.CEL Intensity files) were extracted from the scanned images using the Affymetrix GeneChip® Command Console (AGCC) version 3.2. CEL files were further processed with Affymetrix Expression Console software version 1.1 to calculate probeset signal intensities using the statistics-based Affymetrix algorithms MAS-5.0 with default settings and global scaling as normalization method. The trimmed mean target intensity of each chip was arbitrarily set to 100.

The control set of genes was defined by genes showing significant expression in wild type and showing irrelevant variations in *svb* and *pri* mutants (p-value >0.8). The table below summarizes their documented embryonic pattern, expression levels in *svb* and *pri* mutant conditions (%of *wt*) and if ChIP peaks are present within a +/-5kb window.

Control set of genes

gene	*CG Number*	*expression pattern*	*svb*	*pri*	*ChIP in 5kb*
Ald	CG7643	not epidermal	84,52	91,17	no
Atg18	CG7986	not epidermal	100,61	98,82	no
Bap55	CG6546	not epidermal	108,19	109,38	yes
bbx /// waw	CG1414	ND	101,98	109,42	yes
betaggt-I	CG3469	ND	110,78	130,95	yes
betaTub56D	CG9277	ep stripes	105,75	105,22	no
bip2	CG2009	not epidermal	146,43	130,23	yes
blow	CG1363	not epidermal	101,79	104,82	no
Bre1	CG10542	ND	108,66	134,50	no

bru-2	CG43065	ND	118,64	112,13	no
bsf	CG10302	ND	98,81	94,84	no
Cad87A	CG6977	ND	109,50	112,09	no
CBP	CG1435	not epidermal	102,78	91,76	yes
cenG1A	CG31811	not epidermal	108,36	102,03	no
CG10365	CG10366	not epidermal	100,39	101,34	yes
CG10731	CG10731	ND	97,43	96,07	yes
CG11877	CG11877	not epidermal	113,94	118,58	no
CG12006	CG12006	not epidermal	114,23	113,87	yes
CG12164	CG12164	ND	83,20	93,45	no
CG12375	CG12375	not epidermal	103,42	122,64	no
CG12404	CG12404	ND	104,22	90,36	yes
CG13284	CG13284	not epidermal	100,89	92,66	yes
CG1371	CG1371	not epidermal	115,46	100,11	yes
CG14229	CG14229	not epidermal	105,43	109,87	no
CG14442	CG14442	ND	111,93	120,36	no
CG14636	CG14636	ep stripes	112,31	88,23	no
CG15099	CG15099	ND	88,57	94,68	yes
CG17082	CG17082	ND	91,18	94,01	yes
CG18549	CG18549	not epidermal	98,92	90,20	yes
CG1965	CG1965	ND	90,48	108,38	yes
CG2249	CG2249	ND	97,42	83,94	no
CG2249	CG2249	ND	97,42	83,94	no
CG2918	CG2918	epidermal ubiquitous	103,01	88,58	no
CG31108	CG31108	not epidermal	103,94	105,91	no
CG32164	CG32164	not epidermal	98,31	125,93	no
CG32267	CG32267	ND	95,97	115,07	no
CG32676	CG32676	ND	90,95	93,01	yes
CG3305	CG3305	ND	88,29	79,96	yes
CG3493	CG3493	ND	126,41	109,19	yes
CG4210	CG4210	ND	91,21	118,02	yes
CG4841	CG4841	ND	110,54	95,91	no
CG5869	CG5869	ND	89,39	108,39	yes
CG5931	CG5931	not epidermal	123,35	92,97	no
CG6230	CG6230	ND	95,47	123,11	no
CG6406	CG6406	not epidermal	90,56	87,00	yes
CG6852	CG6852	ND	111,03	109,18	yes
CG7028	CG7028	not epidermal	101,44	109,71	yes
CG7852	CG7852	ND	106,54	113,74	no
CG8090	CG8090	ND	85,52	115,10	yes
CG8289	CG8289	not epidermal	101,82	109,37	no
CG8878	CG8878	not epidermal	103,04	96,89	yes
CG8928	CG8928	ND	91,78	113,34	no
CG8931	CG8931	not epidermal	110,33	111,09	yes
CG9293	CG9293	ND	99,22	92,58	yes
CG9715	CG9715	ND	119,34	97,89	yes
CG9776	CG9776	ND	107,81	92,47	no
CG9917	CG9917	not epidermal	126,61	102,54	no
Chd1	CG3733	ND	97,28	94,33	no
crp	CG7664	not epidermal	82,88	111,58	yes
dbr	CG11371	ND	139,91	126,84	no

drl	CG17348	ep stripes	115,07	78,43	yes
Fip1	CG1078	ND	104,96	87,54	no
gek	CG4012	ND	110,94	96,43	no
gp210	CG7897	ND	110,61	106,95	no
gry	CG17569	ND	97,01	89,65	no
hkl	CG10473	not epidermal	97,76	103,84	yes
kis	CG3696	ND	112,50	96,17	yes
Krn	CG32179	not epidermal	102,68	108,51	yes
kst	CG12008	ep stripes	127,31	103,42	yes
lack	CG4943	ND	109,29	111,22	no
MBD-R2	CG10042	not epidermal	99,57	106,58	no
mmy	CG9535	not epidermal	95,52	97,92	no
mRpL33	CG3712	ND	105,01	112,98	yes
mRpS11	CG5184	not epidermal	114,70	108,13	no
mRpS11	CG5184	not epidermal	114,70	108,13	no
msn	CG16973	ND	114,08	115,00	yes
Nat1	CG3845	not epidermal	108,49	82,47	no
RanBPM	CG42236	ND	124,37	103,44	no
Rap2l	CG3204	ND	108,16	107,35	yes
Rga	CG2161	not epidermal	93,53	108,09	yes
Rgl	CG8865	not epidermal	137,42	100,20	yes
RhoGAP1A	CG40494	ND	73,46	113,54	yes
robo	CG13521	ND	91,22	73,45	no
Rpb5	CG11979	not epidermal	81,38	91,59	no
RpS27	CG10423	not epidermal	106,16	76,13	no
RpS30	CG15697	ND	115,96	77,01	no
RpS30	CG15697	not epidermal	115,96	77,01	no
slik	CG4527	ND	90,39	88,72	yes
Ssdp	CG7187	not epidermal	111,89	87,21	no
Stlk	CG40293	not epidermal	105,66	87,23	no
Taf1	CG17603	not epidermal	107,83	109,83	no
tra2	CG10128	not epidermal	95,99	79,72	yes
Trim9	CG31721	not epidermal	96,29	91,91	yes
trr	CG3848	ND	80,15	106,49	yes
ttk	CG1856	ep stripes	97,50	80,48	yes
Ubp64E	CG5486	not epidermal	130,11	101,59	no
Ufd1-like	CG6233	not epidermal	112,22	110,15	no
ush	CG2762	ep stripes	100,46	98,55	yes
vsg	CG16707	not epidermal	113,76	107,51	yes
zormin	CG33484	ND	130,13	114,51	yes

4. Evolutionary conservation of the *tyn2* enhancer

tyn2 site A

```
D_mel GTATCTACCTACTAGTAGTTACCGTTACGCAGCTG--AAAGATGCCAAA
D_sim GTATCTACCTCCTAGTAGTTACCGTTACGTAGCTG--AAAGATGCCAAA
D_sec GTATCTACCTCCTAGTAGTTACCGTTACGCAGCTG--AAAGATGCCAAA
D_yak GTATCTACCTACTAGTAGTTACCGTTACGCAGATGCCGGAGAAAACA--
D_ere GTATCTACCTACTAGTAGTTACCGTTACGCAGATGCCAAAGAAAACA--
D_eug GTATCTACCTACTAGTAGTTACCGTTACGCAGCAA--GAAGATGCCACA
D_ele GTATCTACCTACTAGTAGTTACCGTTACGCAGCGA--GAAGATGCCACA
D_ana CTGCCTAGCTACTAGTAGATACCGTTACGCACACACCACACACA-CAC-
D_pse GAATCTACCTACTAGTAGTTACCGTTACGCAGCTACTATACCAG-CAC-
D_per GAATCTACCTACTAGTAGTTACCGTTACGCAGCTACTATACCAG-CAC-
D_mir GAATCTACCTACTAGTAGTTACCGTTACGCAGCTACTATACAAG-CAC-
D_wil ATTTTTACCTACTGGAAGTTACCGTTACGCAGCAACATTTCGTTTTA--
D_moj AGCTGCTGCGAGTAGTAGATACCGTTACGCAGCAACTTTTCAGCCAA--
D_vir GTCTGCTACGAATAGTAGATACCGTTACGCAGCAACTTTTCAGTTAA--
D_gri GAGTAGTAGTAGTAGTAGATACCGTTACGCAGCAACTTTTCAGTTAA--
```

tyn2 site B

```
D_mel AGTTCCTCGACTATCAG---ATACCCGTTACTCAACTGGAAGAGTGAAGG--
D_sec AGTTCCTCAACTATCCG---ATACCCGTTACTCAGCTGGACGTGTGATGC--
D_yak -TTGACTGGA--AAGTG-CATTTTATGTAGTTAATATAAACGAATAAGATG-
D_ere -TACACAGGA--AAGTG-CATTTTATGTAATTAATATAAACGAATAAGAT--
D_eug -TTGACTGGA--AAGTG-CATTTTCTGTTATTAATATATATGGTAGATAT--
D_ele TTTGACTGGA--AAGTG-C-TTTTCTGTTATTAATATATTTCTTCTATAT--
D_ana -TTGACTGGA--AAGTG-CATTTTCTGTTATTAATATCAA-GAAGTGGATT-
D_pse -TTGACTGGA--AAGTG-CATTTTCTGTTATTAATATAA--GAAATAGATGG
D_per -TTGACTGGA--AAGTG-CATTTTCTGTTATTAATATAA--GAAATAGATGG
D_mir -TTGACTGGA--AAGTG-CATTTTCTGTCATTAATATAA--GAAATAGATGG
D_wil --TGACTAAATCAATTGACAATTTCTTTTATTAAAAATGAAAAATTAA----
```

Figure S1A : Identification of 21 additional Svb-downstream genes

Genes displaying expression in subsets of epidermal were selected from the database of expression patterns, developed by the Berkeley Drosophila Gene Project (http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl). mRNA expression was compared between wild type (*wt*) and *shavenbaby (svb)* mutant embryos by *in situ* hybridization. These 21 genes show a reduced expression in the absence of Svb, as documented by ventral views, with the exception of CG12814, CG14395, CG15005, CG31559 and CG31973 representing laterals views and CG15022 a dorsal view. That the expression of these genes depends on *svb* activity was further confirmed by their up-regulation following ectopic expression of Svb in the epidermis (not shown).

Figure S1B : Epidermal genes independent of Svb

36 genes expressed in subsets of epidermal cells showing no significant modification of their expression in *svb* mutant embryos, when compared to *wt* control. This defines a set of epidermal genes used as negative control in motif discovery approaches.

Figure S1C : Motif prediction and CRM activity.

Top: Motifs predictions in Svb downstream genes and control epidermal genes, using cisTargetX (http://med.kuleuven.be/cme-mg/lng/cisTargetX/). The predicted motifs are ranked according to their enrichment within each set compared to all *Drosophila* genes, and their evolutionary conservation. Within the set of 36 epidermal genes that are independent of Svb (left), highly ranked motifs include binding sites associated with transcription factors involved in general epidermis differentiation, such as Grh, cEBP/Vri, but no Ovo/Svb-like motifs. In the set of 39 Svb downstream genes (middle), 4 of the top 5 motifs are related to the Ovo/svb binding site, all sharing the core sequence (CnGTT or AACnG in the reverse orientation). Upon their addition in the cisTarget library of motifs (right), svbF7 and blue motifs became the first and third most enriched motifs, respectively. The use of svbF7 also increased the accuracy of enhancer prediction when compared to OvoQ6, with three additional Svb-dependent enhancers (*32159, Emin, EminB,* in cyan) detected in the top100 cisTarget predictions, and 9 negative regions (pink) no longer predicted by cis-Target (*f2,*

sox21b, snH5, sha-intron, f5, f4, cyrB, snP & snE4). **Bottom:** Expression pattern of 6 additional trichome enhancers identified during initial stages of our study. These enhancers drive reporter expression in trichome cells (lacZ immuno-staining, brown), reproducing fully or partially endogenous expression of their respective genes, as assayed by *in situ* hybridization to mRNA (purple). Reporter expression was strongly reduced in *svb* mutant embryos, showing that the activity of these enhancers depends on Svb function. Tested regions were selected from different attempts of predictions, based on putative evolutionary footprinting (*EminB*), manual examination of OvoQ6-related motifs (*17058, 31559*) or an earlier version of cisTarget (version1, genome release4) for *4702B, tyn2* and *32159*. While *EminB* and *32159* become predicted by cisTargetX following the introduction of svbF7, other active enhancers do not, due to a lack of motif clustering and/or evolutionary conservation.

Figure S1D : comparison of the predictive efficiency of various Ovo/Svb related PWMs

Top: SvbF7, ovoQ6 as well as additional Ovo/Svb related PWMs (as extracted from the Fly Factor Survey database, http://pgfe.umassmed.edu/ffs/) were used with *i*-cistarget to analyse the set of 39 *svb* downstream genes. PWMs are ranked according to their enrichment score. Logo representation highlights differences in nucleotide composition and/or relative weight between PWMs. **Bottom:** Pareto plots comparing the efficiency the five Ovo/Svb-related PWMs in discriminating between the 14 functional enhancers and 25 negative regions using motifs conserved across *Drosophila* species (left) or all motifs present in *D. melanogaster* genomic reigons (right). SvbF7, and to a lesser extent OvoQ6, performs better than Ovo_FlyReg, ovo_SOLEXA or ovo_SANGER that detect more false negatives (x axis) and false positives (y axis).

Figure S2A : Architecture of cis-regulatory motifs within trichome enhancers

Graphs plot the distance measured between all possible combinations of homotypic pairs of svbF7 and blue motifs (F7-F7 and bm-bm, resp.) and of the distance between svbF7 and either a blue (F7-bm) or a yellow motif (F7-ym). These analyses did not revealed obvious

bias in the positioning of cis-regulatory motifs, as quantified by the absolute distance (bp) or relative to helical periodicity (expressed as the percentage of DNA helix rotation)

Figure S2B : Distribution of cis-regulatory motifs associated with Svb regulated genes.

Distribution of svbF7, blue or yellow motifs within the whole set of Svb-regulated genes (150 genes defined from microarrays) *versus* the set of control genes (100 genes from microarrays), as estimated by the number of detected motifs *per* gene. **Left panel**: The graph plots the number of evolutionarily conserved svbF7 and blue motifs detected in each set of genes. *** indicates a p-value <0,001, ** <0,01. **Right panel**: A significant enrichment for svbF7 alone, or in combination with blue (bm) or yellow motifs (ym) is detected in the set of Svb-regulated genes when compared to control genes. To avoid over-fitting, the positives sequences (CRMs) used in Fig. 3 for *de novo* motif discovery were masked prior analyses. The combination of svbF7 and blue motif exhibits higher selectivity (<5% FPR), albeit reducing sensitivity of detection. In addition, prediction with svbF7+blue or svbF7+yellow is higher (more sensitive) than with only svbF7+blue (or with only svbF7+yellow), indicating that a subset of Svb regulated-genes are predicted by the svbF7+blue combination, whereas others are predicted using the svbF7+yellow combination.

Figure S3 : Genes regulated by Svb as deduced from microarray profiling.

For microarrays analysis, we focused on genes showing significant levels of expression in wild type embryos, at the temporal stage examined. From this list of 5000 genes, 150 of them displayed down-regulation in *svb* mutant and in *pri* mutant embryos. Genes are ranked according to their expression levels in *svb* mutant embryos, expressed as the percentage of wild type levels. Levels of residual expression relative to *wt* are indicated for RNA samples extracted from *pri* and *svb* mutants. Further validation of candidate target genes was performed by *in situ* hybridization in embryos mutant for *svb* (see Fig. S4), or manipulated to drive ectopic *svb* expression. For each gene, the chart indicates known or putative function and protein domain, expression pattern in the epidermis and additional embryonic tissues. It

also summarizes the presence of associated svbF7, blue or yellow motifs. ChIP peaks (at two developmental stages) were associated with genes when located in a 5kb window upstream and downstream the transcribed region plus introns. Bona fide Svb-target genes are highlighted in green, tested genes that displayed no modifications of their expression pattern in modified *svb* genetic backgrounds are in grey.

Figure S4 : Experimental validation of novel Svb target genes identified from microarrays. Gene expression was assayed by *in situ* hybridization, comparing patterns observed in wild type (left panels) and *svb* mutant embryos (right panels). These 21 genes displayed a reduction in their mRNA levels in trichome cells in the absence of *svb*, while additional expression domains were unaffected, providing internal controls for specificity.

Figure S5 : Analysis of Svb-bound regions

Top: Cross-correlation between conserved svbF7 (red), blue or yellow motif instances and Svb ChIP peaks associated with either Svb regulated genes (left) or control genes (right). Plots show numbers of svbF7, blue and yellow motifs found in a 10kb window on each side of the center of peaks. **Bottom**: Histogram of p-values corresponding to cross-correlation tests between conserved svbF7 (red), blue or yellow motifs and Svb ChIP peaks, as defined from two independent ChIP-seq replicates, and their reproducibility analysis using the IDR package (https://sites.google.com/site/anshulkundaje/projects/idr).

Figure S6 : Motif analysis of ChIP peaks associated with Svb regulated or control genes.

Svb-bound sequences associated with Svb-regulated and control genes were subjected to *de novo* motif discovery, using the Peak Motifs computational pipeline from the Regulatory Sequence Analysis Tools package (http://rsat.ulb.ac.be). Enriched motifs are listed according to their rank and the corresponding logo build from *de novo* discovery is indicated. Each discovered motif was compared and aligned to known TF binding sites when showing

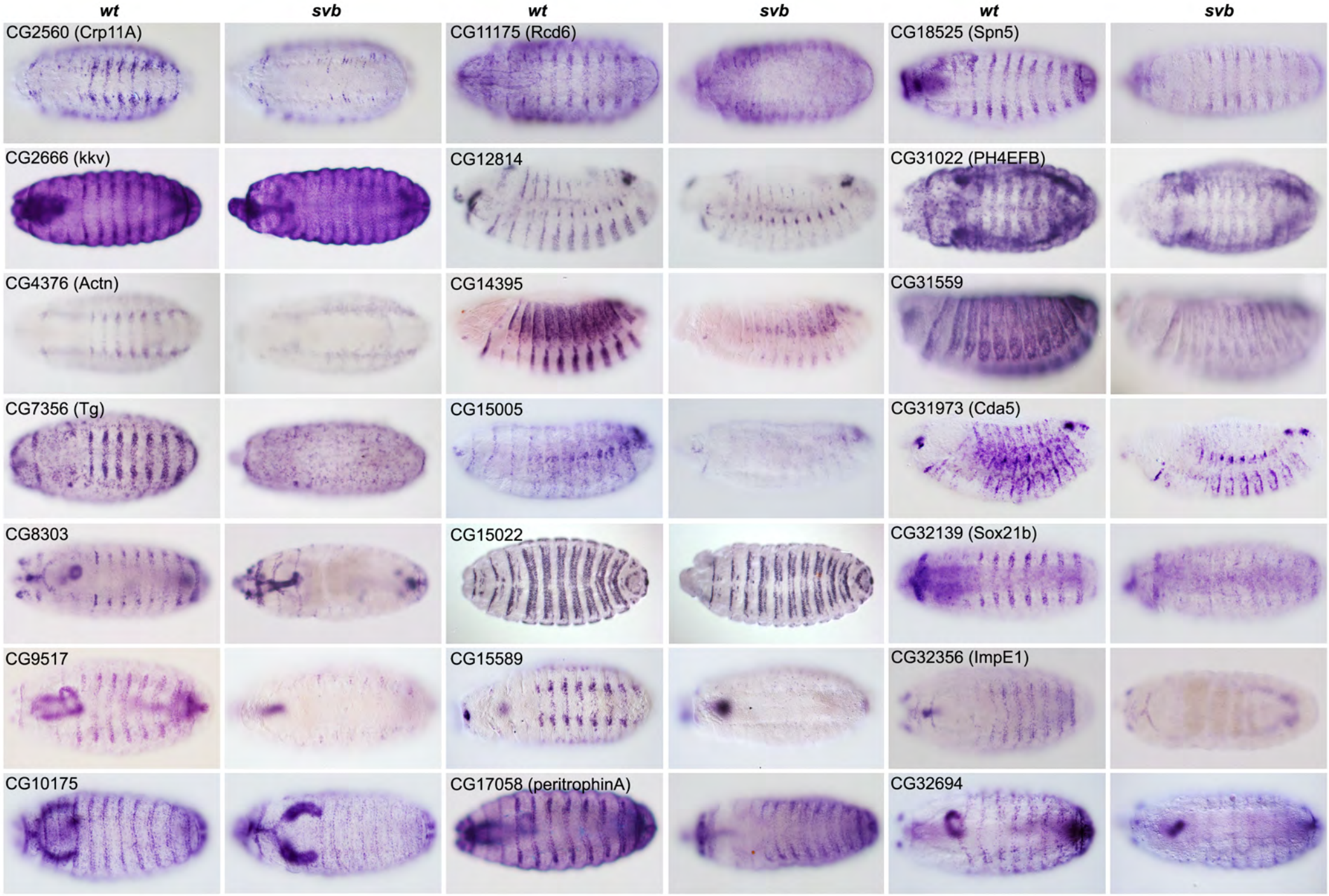
substantial overlap. Within peaks associated with Svb regulated genes, motif 3 (tACcGTTAs) extensively matches svbF7 (ACnGTTAg) and motif 9 shows limited similarity to the blue motif.

Figure S7 : ChiP-seq profiles of 18 Svb regulated genes

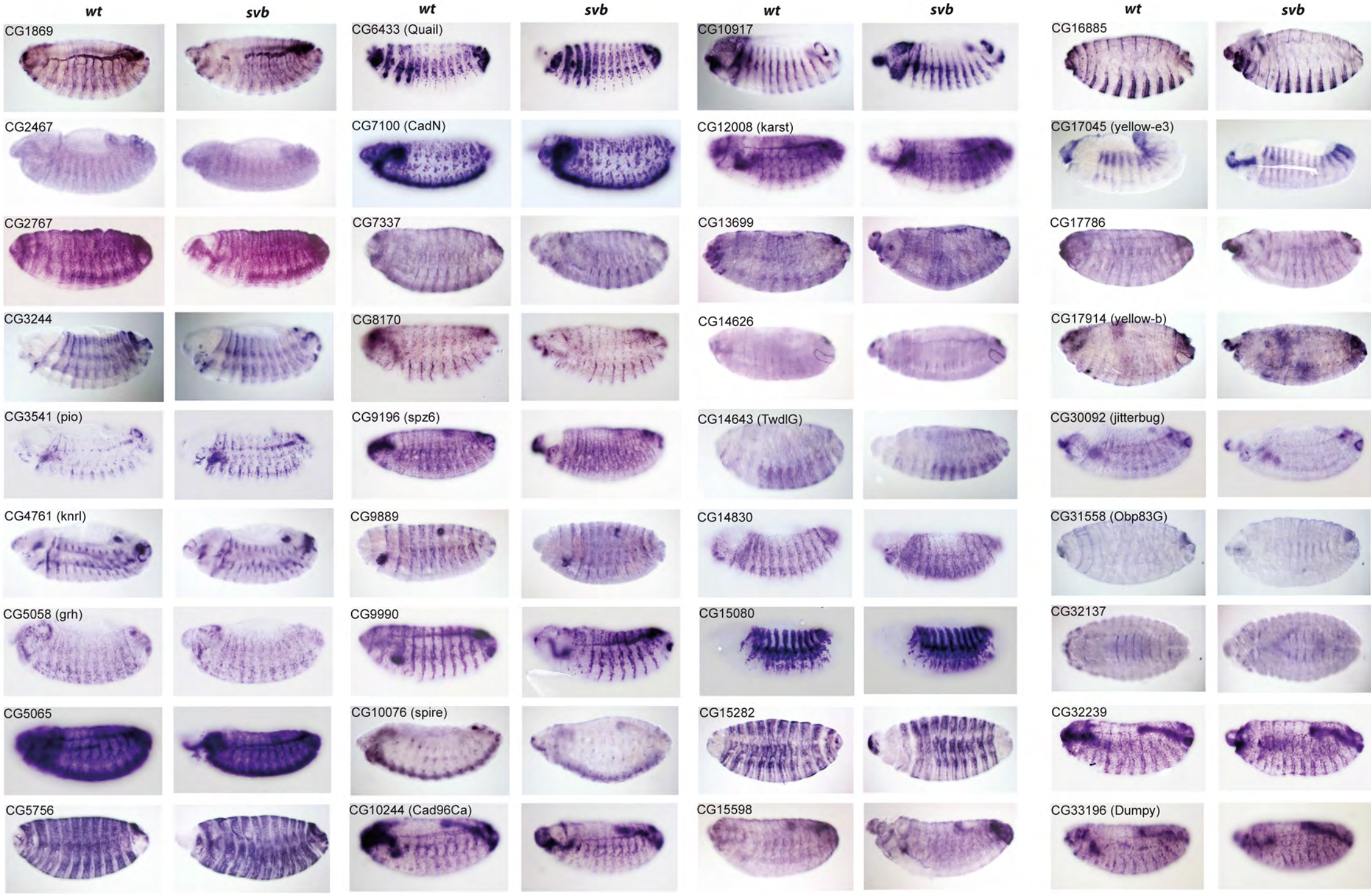
Screen shot views from the Integrated Genome Browser (http://www.bioviz.org/igb, Nicol & al, Bioinformatics 2009) of ChIP-Seq signals collected in the two independent replicates. ChIP-peaks called from MACS analysis are shown under each ChIP-seq profile. Conserved svbF7 (red), OvoQ6 (black), blue and yellow motifs are drawn as vertical bars. Enhancers (positives) are shown as cyan boxes, negative regions in pink.

Figure S8 : Evolution of the distribution of cis-regulatory motifs within trichome enhancers, across *Drosophila* species.

Schematic representation of the distribution of svbF7 (red) blue and yellow motifs for each enhancer region, across *Drosophila* species. For motif detection, individual sequences from *D. melanogaster* and each of the orthologous regions taken from the 11 additional *Drosophila* species were processed independently, using the same threshold for the three motifs. Orthologous regions were aligned with respect to the best-conserved svbF7 site. Cis regulatory motifs that are well conserved and traceable across species are connected by full lines. Motifs for which the pattern of conservation is inferred from a parsimonious guess are connected by dashed lines. Trichome enhancers were regrouped along those showing strong (A) or more relaxed (B) conservation in the positioning of cis-regulatory motifs.



Menoret *et al.;* Figure S1A



Menoret *et al.;* Figure S1B

cis-TargetX motifs detection

epidermal control genes

1 TIFDMEM0000099

2 SelexConsensus_grh

3 M00117-V-CEBPB_02

4 badis_lsgf3g_primary

5 M00109-V-CEBPB_01

39 Svb downstream genes

1 ACCGTTA

2 M00968-P-ATMYB77_01

3 M00993-V-TAL1_Q6

4 M01101-I- OVO_Q6

5 M00227-V-VMYB_02

39 Svb downstream genes + SvbF7 & BM

1 SvbF7

2 ACCGTTA

3 Blue Motif

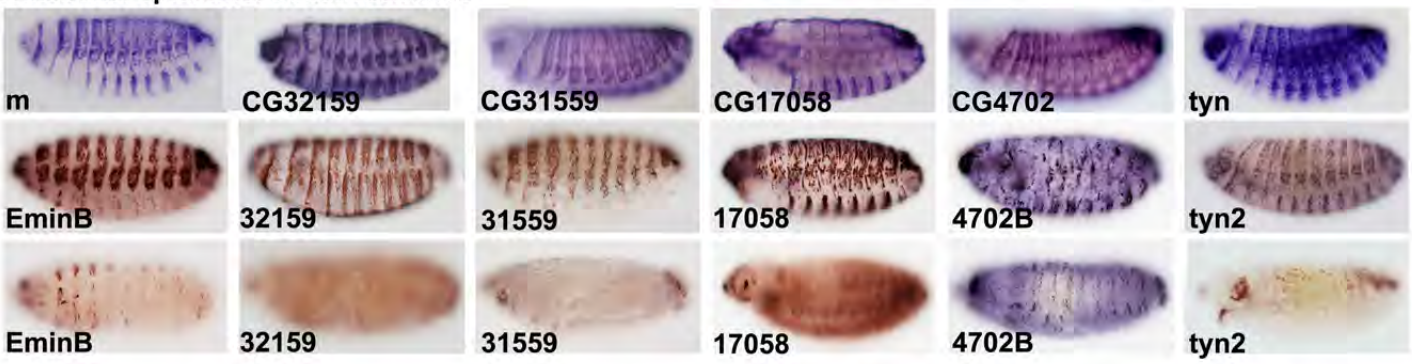
4 nAACwGTw

5 AATTAGCA

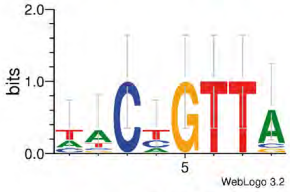
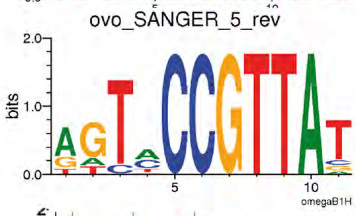
Ranking of genomic predicted regions

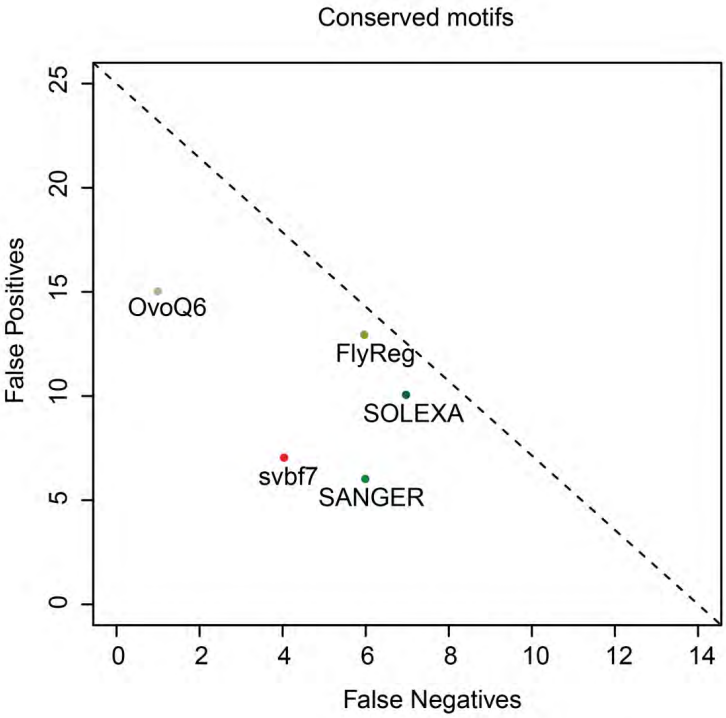
CisTargetX predicted CRMs	dyl1	15589	cyrA	snE5	dyl2	pmin	f2	32159	Emin	sox21b	sn-enh1	sha3	snH5	snB2	nyo3	snE1	sha-intron	EminB	f5	sha2	f4	cyrB	snP	nyo1	snE4
OvoQ6	1	3	4	5	8	ND	9	ND	ND	13	21	23	24	32	38	43	50	ND	55	63	73	77	79	80	82
svbF7	11	1	2	52	15	21	ND	24	26	ND	104	34	ND	58	7	38	ND	55	ND	10	ND	ND	ND	97	ND

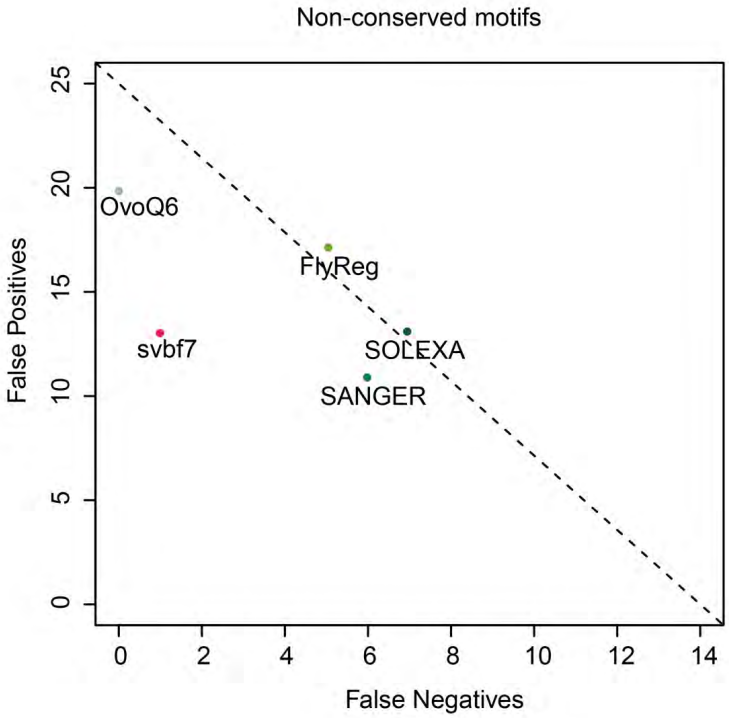
Additional epidermal CRMs identified



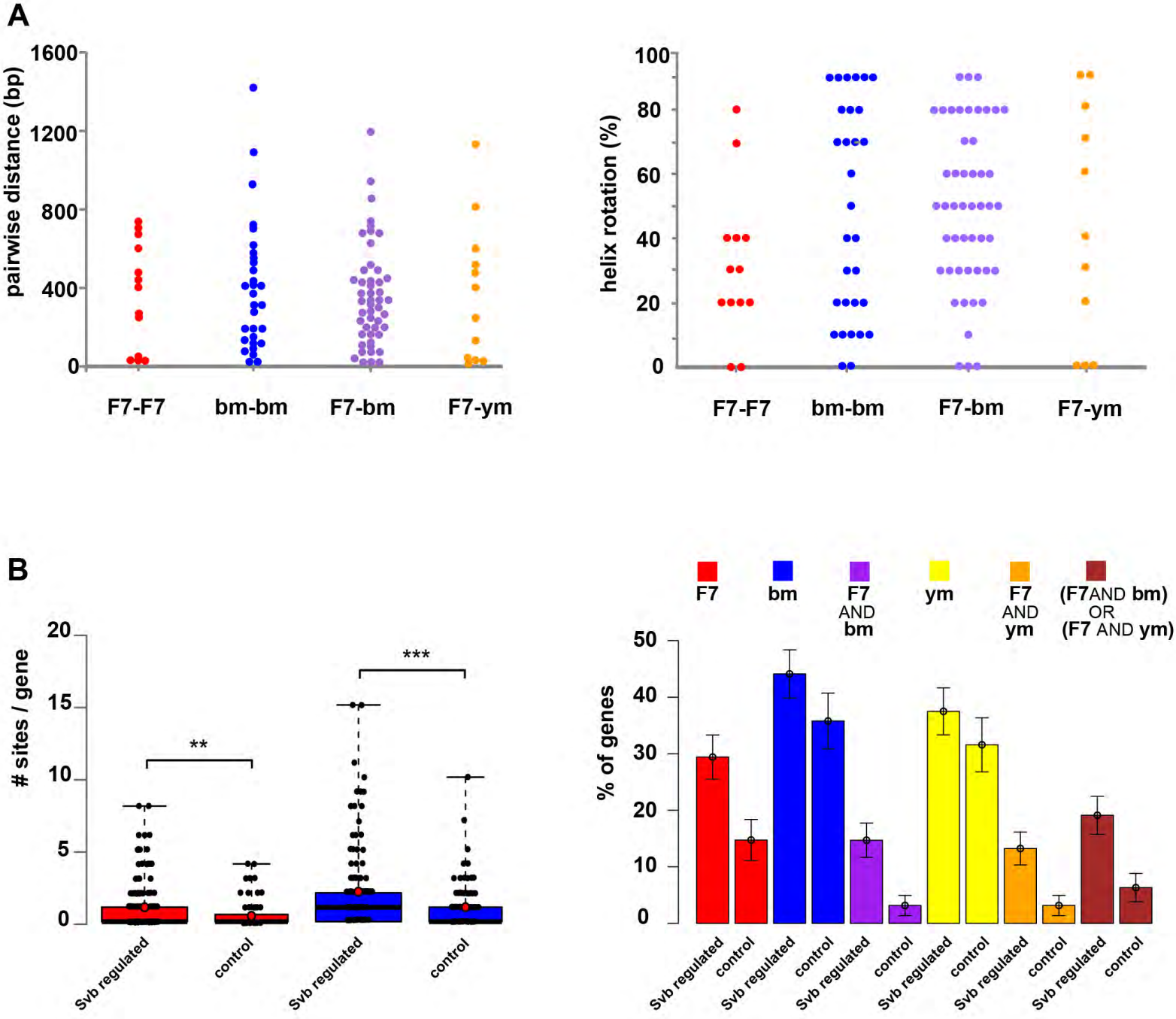
Menoret *et al.;* Figure S1C

rank	name	score	PWM
2	SvbF7	8.95	
4	OvoQ6	8.09	
5	ovo_SOLEXA	7.53	
6	ovo_SANGER	7.49	
9	ovo_FlyReg	5.56	





Menoret *et al.*; Figure S1D



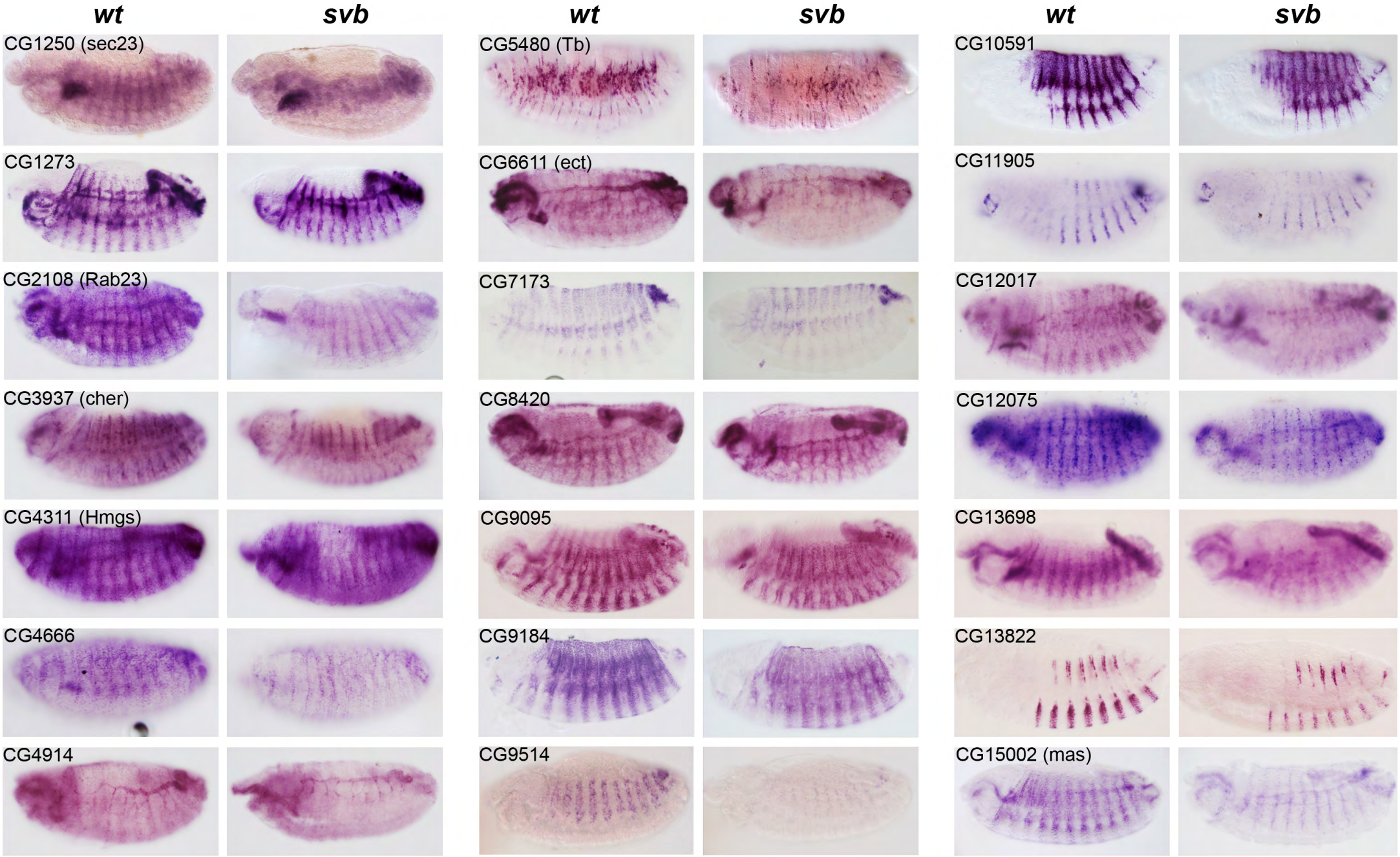
Menoret *et al.;* Figure S2

Gene symbol	Representative Public ID	Protein features/putative domains	epidermal expression	additional expression #	svb mutant microarrays (% of wt)	pri mutant microarrays (% of wt)	**validated in svb mutant (in situ)**	**validated in svb ectopic (in situ)**	**svbF7**	**blue motif**	**yellow motif**	ChIP peaks in 5kb (12-14h)	ChIP intensity (12-14h)	ChIP peaks in 5kb (8-10h)	**reference**
CG15370	CG15370	unknown	no	ubiquitous	13,1	2,1	no	ND	no	yes	no	yes	502	yes	this work
sha	CG13209	actin binding	stripes		24,6	6,6	yes	yes	yes	yes	yes	yes	92,92	yes	ref 1
CG14395	CG14395	PH like domain	stripes		41,2	3,0	yes	yes	yes	yes	yes	yes	135,174	yes	this work, FigS1
CG4386	CG4386	trypsin like protease	no	trachea	43,5	11,3	no	ND	no	yes	yes	no		no	this work
CG15818	CG15818	C-type lectin domain	no	midgut	47,9	17,8	no	ND	no	yes	no	no		yes	this work
m	CG9369	ECM	stripes		48,5	6,2	yes	yes	yes	yes	no	yes	173,768,69	yes	ref 1
CG32159	CG32159	Cuticle formation	stripes		50,1	12,6	yes	yes	yes	yes	yes	yes	56,789,69,54	yes	ref 2
nyo	CG1499	ECM	stripes	all gut	54,7	15,3	yes	yes	yes	yes	yes	yes	500	yes	ref 1
CG4914	CG4914	trypsin domain	stripes	hindgut	55,9	21,7	yes	ND	yes	yes	yes	yes	418,418	yes	this work, FigS4
neyo	CG7802	ECM	stripes	all gut	56,1	17,3	yes	yes	no	yes	no	yes	59,14	yes	ref 1
mey	CG12063	ECM	stripes		57,1	7,8	yes	yes	yes	yes	yes	yes	654	yes	ref 1
CG16798	CG16798	unknown	stripes		57,1	18,4	yes	yes	no	yes	yes	yes	140	yes	ref 1
CHOp24	CG3564	protein secretion	stripes	salivary gland	59,0	27,0	ND	ND	no	yes	no	no		yes	
CG11905	CG11905	unknown	stripes	trachea	59,3	20,1	yes	no	yes	yes	no	no		yes	this work, FigS4
CG14356	CG14356	unknown	no	foregut	60,3	27,1	ND	ND	yes	no	yes	no		no	
CG17211	CG17211	von Willebrand factor type C domain	ND	ND	60,5	26,4	ND	ND	no	yes	yes	yes	162	yes	
CG17780	CG17781	unknown	no	hindgut & anal pad	61,0	27,1	no	ND	no	yes	yes	no		no	this work
mwh	CG13913	actin binding	stripes		61,1	25,6	yes	yes	no	no	no	yes	333	yes	ref 1
CG30283	CG30283	trypsin like protease	stripes	atrium & adult eve PR	62,0	18,3	ND	ND	no	yes	no	no		yes	
CG4500	CG4500	bubblegum like domain	ND	ND	65,2	30,3	ND	ND	no	no	yes	yes	956	yes	
spz6	CG9196	Toll signalling pathway	stripes		66,9	16,5	no	no	no	yes	no	no		yes	this work, FigS1
CG1273	CG1273	unknown	stripes	all gut & trachea	67,2	20,8	yes	yes	yes	yes	yes	yes	98,98,52,84,	yes	this work, FigS4
CG6785	CG6785	unknown	ND	ND	67,7	21,4	ND	ND	no	no	no	yes	77	yes	
CG10591	CG10591	unknown	stripes		67,8	21,7	yes	yes	yes	no	yes	yes	86	yes	this work, FigS4
CG13698	CG13698	unknown	stripes	gut & salivary gland	68,4	29,3	yes	yes	yes	yes	yes	yes	51,219,1214	yes	this work, FigS4
CG9514	CG9514	GMC oxidoreductase	stripes		69,0	22,4	yes	yes	yes	yes	yes	yes	175,111	no	this work, FigS4
CG17562	CG17562	fatty acyl CoA reductase	no	oenocytes	69,2	25,3	no	ND	no	no	no	no		no	this work
f	CG5424	actin binding	stripes		70,3	26,0	yes	yes	yes	yes	yes	yes	106,122,59,136,57	yes	ref 1
CG9184	CG9184	unknown	stripes	corpus cardiacum	70,9	15,7	yes	yes	yes	yes	yes	yes	100,59,303,	yes	this work, FigS4
CG12075	CG12075	PH like domain	stripes	gut & salivary gland	71,5	33,4	yes	yes	yes	yes	yes	yes	139,253,135,54,50,186	yes	this work, FigS4
tyn	CG17131	ECM	stripes		72,9	22,2	yes	yes	yes	yes	yes	yes	291,147,200,66,67,99,252	no	ref 1
CG4678	CG4678	metallocarboxypeptidase	no	foregut & anal pad	73,4	36,3	ND	ND	yes	no	yes	yes	52	yes	
CG13616	CG13616	unknown	no	intestine & anal pad	73,4	12,7	ND	ND	no	yes	yes	no		no	
CG14756	CG14756	unknown	no	salivary gland	74,0	12,3	ND	ND	no	yes	no	no		no	
CG4666	CG4666	hot-dog domain	stripes	post spiracle	76,5	24,5	yes	yes	no	yes	yes	yes	66,136,	yes	this work, FigS4
CG4686	CG4686	unknown	ND	ND	77,0	33,7	ND	ND	no	no	no	no		no	
CG13082	CG13082	ketohexokinase	stripes	all gut	77,7	20,7	ND	ND	yes	yes	no	no		yes	
CG7173	CG7173	serine protease inhibitor	stripes		77,8	11,6	yes	yes	no	no	yes	yes	73,65,	yes	this work, FigS4
CG1632	CG1632	trypsin like protease	ND	ND	78,0	18,9	ND	ND	no	no	yes	yes	74,477,86,104,92	yes	
Orct	CG6331	Major Facilitator Superfamily	no	midgut & fat body	78,1	35,5	ND	ND	no	no	no	yes	70	no	
mas	CG15002	chymotrypsin, endopeptidase	stripes	foregut & hindgut & trachea	78,1	31,7	yes	yes	no	yes	no	yes	57	yes	this work, FigS4
CG5039	CG5039	unknown	ND	ND	78,6	12,6	ND	ND	no	yes	yes	yes	109	yes	
CG12017	CG12017	unknown	stripes		78,6	18,0	yes	yes	yes	yes	yes	yes	56,223,72,84,	yes	this work, FigS4
CG5873	CG5873	redox process	stripes	all gut & trachea	78,7	36,1	ND	ND	yes	yes	no	yes	281	yes	
CG8420	CG8420	unknown	stripes		79,3	15,0	yes	no	yes	yes	yes	yes	190,86,170,	yes	this work, FigS4
CG11200	CG11200	metabolic process	stripes	trachea	79,8	32,0	ND	ND	no	no	no	yes	104	yes	
Cyp301a1	CG8587	redox process	stripes	trachea	80,3	34,3	ND	ND	yes	no	no	no		no	
CG42331	CG6879	redox process	ND	ND	80,4	9,8	ND	ND	yes	yes	yes	yes	97,64,64,99	yes	
CG13822	CG13822	GILT domain, thiol reductase	stripes	lymph gland	80,8	34,3	yes	yes	yes	no	yes	yes	181	yes	this work, FigS4
CG12009	CG12009	chitin metabolic process	no	trachea	82,4	17,3	ND	ND	yes	yes	no	yes	53,503	yes	
CG8239	CG8239	isoprenoid biosynthesis	stripes	hindgut & anal pad	82,9	12,7	no	ND	no	no	no	no		no	this work
snRNP-U1	CG5454	mRNA splicing	no	ubiquitous	83,2	38,9	ND	ND	no	yes	no	no		yes	
CG1140	CG1140	ketone body catabolism	no	midgut & fat body	83,4	39,7	ND	ND	no	no	yes	yes	74	no	
dyl	CG15013	ECM	stripes	foregut & hindgut	85,3	16,1	yes	yes	yes	yes	yes	yes	57,58,174,85,498,528,285,57	yes	ref 1
Hr46	CG33183	ecdysone pathway	no	ubiquitous	85,5	39,9	ND	ND	yes	yes	yes	yes	47,54,262,108,460,60,190,70	yes	
vri	CG14029	tracheal system development	stripes	hindgut	85,6	30,5	no	ND	yes	yes	yes	yes	78,187,54	yes	
wus	CG9089	ECM	stripes	all gut & trachea	86,1	12,6	no	ND	no	no	yes	yes	188	yes	this work

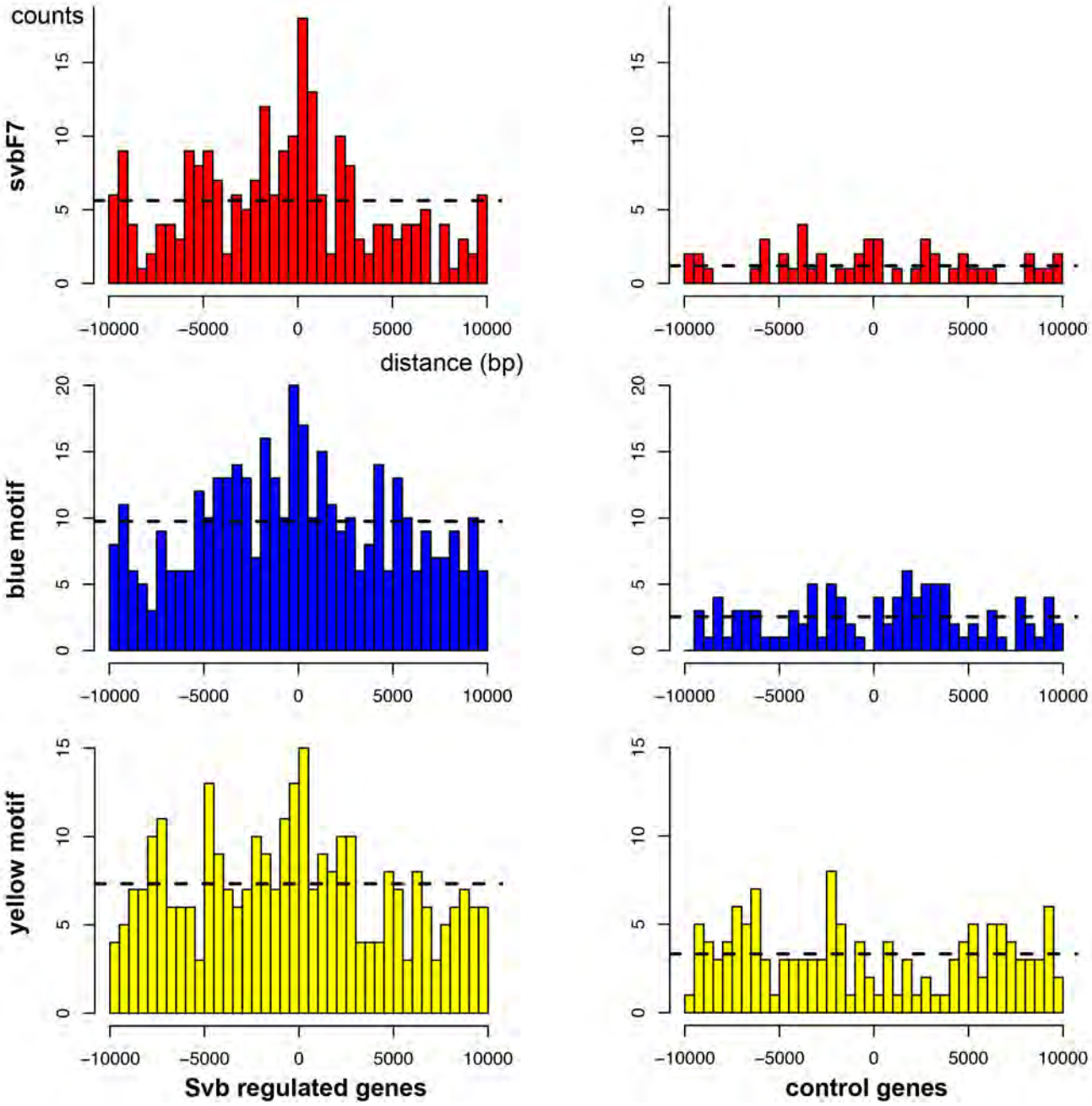
ImpE1	CG32356	ecdysone pathway	stripes	all gut	86,3	14,8	yes	yes	yes	yes	yes	yes	67,251	yes	this work, FigS1
Plip	CG10371	protein dephosphorylation	ubiquitous	midgut & hindgut	86,7	32,5	ND	ND	no	no	yes	yes	239	yes	
CG17672	CG11271	translation	no	ubiquitous	86,7	22,3	ND	ND	no	no	no	yes	573,102	yes	
CG32354	CG32354	unknown	stripes	all gut	86,8	42,3	ND	ND	no	yes	yes	no		yes	
CG8306	CG8306	redox process	stripes	foregut & hindgut	86,8	41,3	ND	ND	no	no	no	no		yes	
CG10585	CG10585	unknown	stripes	all gut	87,1	27,3	ND	ND	no	yes	yes	no		yes	
fw	CG1500	cell adhesion	no		87,2	30,5	ND	ND	no	yes	yes	yes	463,156	yes	
Obp99c	CG7584	sensory perception	no	fat body & amnioserosa	87,3	27,4	ND	ND	no	yes	yes	no		yes	
CG5525	CG5525	microtubule organization	no	muscle system & hindgut	87,4	31,5	ND	ND	no	yes	yes	no		no	
CG34007	Dm.2L.4959.0	unknown	ND	ND	87,5	25,6	ND	ND	no	no	no	yes	164	yes	
CG14470	CG14470	unknown	no	hindgut	87,5	39,1	ND	ND	no	no	no	no		no	
CG18249	CG18249	unknown	no	midgut & amnioserosa	87,6	28,4	ND	ND	no	no	no	yes	59	yes	
CG8386	CG8386	lateral inhibition	ubiquitous	salivary gland	87,9	41,5	ND	ND	no	no	no	no		no	
pwn	CG11101	EGF-like calcium binding	ND	ND	88,0	29,8	ND	ND	no	yes	yes	yes	51,51	yes	
PH4alphaSG1	CG31014	redox process	no	salivary gland	88,1	39,7	ND	ND	no	no	yes	no		no	
CG6415	CG6415	glycin catabolic process	no	fat body	88,2	37,1	ND	ND	no	yes	yes	yes	103	yes	
cher	CG3937	actin binding	stripes	muscle system	88,3	40,5	yes	no	yes	yes	yes	no		yes	this work, FigS4
CG2016	CG2016	unknown	stripes	all gut & trachea	88,4	22,9	ND	ND	no	yes	yes	no		no	
Tb	CG5480	Cuticle formation	stripes		88,6	21,5	yes	yes	no	yes	yes	yes	309,53	no	this work, FigS4
CG10932	CG10932	mitotic spindle organization	no	midgut	88,8	34,6	ND	ND	no	no	yes	no		yes	
scu	CG7113	ecdysone pathway	no	midgut	89,1	39,8	ND	ND	no	no	yes	no		no	
CG7860	CG7860	autophagic cell death	no	midgut & crystal cells	89,7	15,8	ND	ND	no	yes	no	no		yes	
CG9356	CG9356	unknown	no		89,7	39,8	ND	ND	no	no	no	no		yes	
Hmgs	CG4311	hydroxymethylglutaryl-CoA synthase	stripes	foregut & hindgut	90,0	28,7	yes	no	no	no	yes	yes	286	yes	this work, FigS4
CG11836	CG11836	proteolysis	ubiquitous		90,3	32,3	ND	ND	no	no	no	yes	111	no	
CG9503	CG9503	redox process	no	dorsal trunk	90,4	22,4	ND	ND	no	no	yes	yes	256	yes	
CG1837	CG1837	apoptotic cell clearance	no	ubiquitous	90,4	39,5	ND	ND	yes	no	no	yes	80,66	yes	
tw	CG12311	somatic muscle development	ND	ND	90,5	38,0	ND	ND	no	no	no	no		yes	
PH4alphaEFB	CG31022	procollagen dioxygenase	stripes	muscle system & plasmatocytes	90,6	40,5	yes	yes	yes	yes	yes	yes	540,78	yes	this work, FigS1
amd	CG10501	chitin metabolic process	ND	ND	90,6	38,1	ND	ND	yes	no	yes	yes	261,60,	yes	
Rab23	CG2108	GTPase, planar polarity	stripes	foregut	91,5	31,9	yes	yes	no	yes	no	yes	289	yes	this work, FigS4
Smn	CG16725	neuromuscular junction	no	gonad	91,5	37,6	ND	ND	no	no	no	no		yes	
Pros28.1	CG3422	protease	no	ubiquitous	91,6	28,1	ND	ND	yes	no	no	yes	204,185	yes	
CG9175	CG9175	unknown	no	midgut & hindgut& salivary gland	91,6	40,9	ND	ND	no	no	no	no		yes	
Rlc1	CG9378	translation	ND	ND	91,8	45,2	ND	ND	no	yes	no	no		yes	
Lip4	CG6113	lipid metabolism	no	amnioserosa	91,9	44,5	ND	ND	no	no	yes	yes	240	no	
CG7840	CG7840	lipid metabolism	no		91,9	35,2	ND	ND	yes	no	yes	no		no	
Gtp-bp	CG2522	protein secretion	stripes	midgut & hindgut& salivary gland	92,1	39,0	ND	ND	no	no	no	no		yes	
CG32250	CG32250	transport	ND	ND	92,2	41,7	ND	ND	no	no	no	no		no	
CG15506	CG15506	unknown	ND	ND	92,2	13,6	ND	ND	yes	yes	yes	yes	53,69	yes	
TRAM	CG11642	protein targeting to membrane	stripes	salivary gland	92,3	36,0	ND	ND	yes	no	yes	no		yes	
CG6704	CG6704	unknown	no	yolk nuclei	92,5	16,3	ND	ND	no	yes	yes	no		no	
CG17218	CG17218	tracheal system development	stripes	all gut & anal pad	92,5	40,8	ND	ND	no	no	yes	no		no	
CG4065	CG4065	unknown	no	midgut & muscle system	92,6	46,3	ND	ND	no	no	no	yes	322,66,	yes	
mRpL46	CG13922	unknown	no	midgut & muscle system	92,7	42,8	ND	ND	no	no	no	no		no	
CG6180	CG6180	unknown	no	midgut	92,7	44,0	ND	ND	no	yes	no	no		yes	
T-cp1	CG5374	protein folding	no	ubiquitous	92,9	41,1	ND	ND	yes	no	no	yes	72	no	
CG13585	CG13585	unknown	no	garland cell	93,0	45,3	ND	ND	no	no	yes	no		no	
nimB3	Dm.2L.8912.0	unknown	ND	ND	93,6	29,5	ND	ND	no	no	no	no		no	
CG2663	CG2663	transport	head	post spiracle	93,7	41,8	ND	ND	no	yes	yes	yes	208	yes	
CG11786	CG11786	unknown	no	dorsal trunk	93,8	27,0	ND	ND	no	no	no	yes	281	no	
rt	CG6097	synaptic activity	ND	ND	94,0	25,7	ND	ND	no	yes	no	yes	61	yes	
CG13627	CG13627	unknown	no	trachea	94,2	16,2	ND	ND	no	yes	yes	no		no	
Gmap	CG33206	protein targeting to Golgi	ND	ND	94,3	41,8	ND	ND	no	no	yes	no		yes	
CG6672	CG6672	transmembrane transport	ND	ND	94,4	42,0	ND	ND	no	yes	yes	no		yes	
CG4702	CG4702	unknown	stripes	all gut	94,5	16,9	yes	yes	no	yes	yes	yes	50,344	yes	ref 1
CG3831	CG3831	unknown	no	corpus allatum	94,5	37,3	ND	ND	yes	no	no	no		yes	

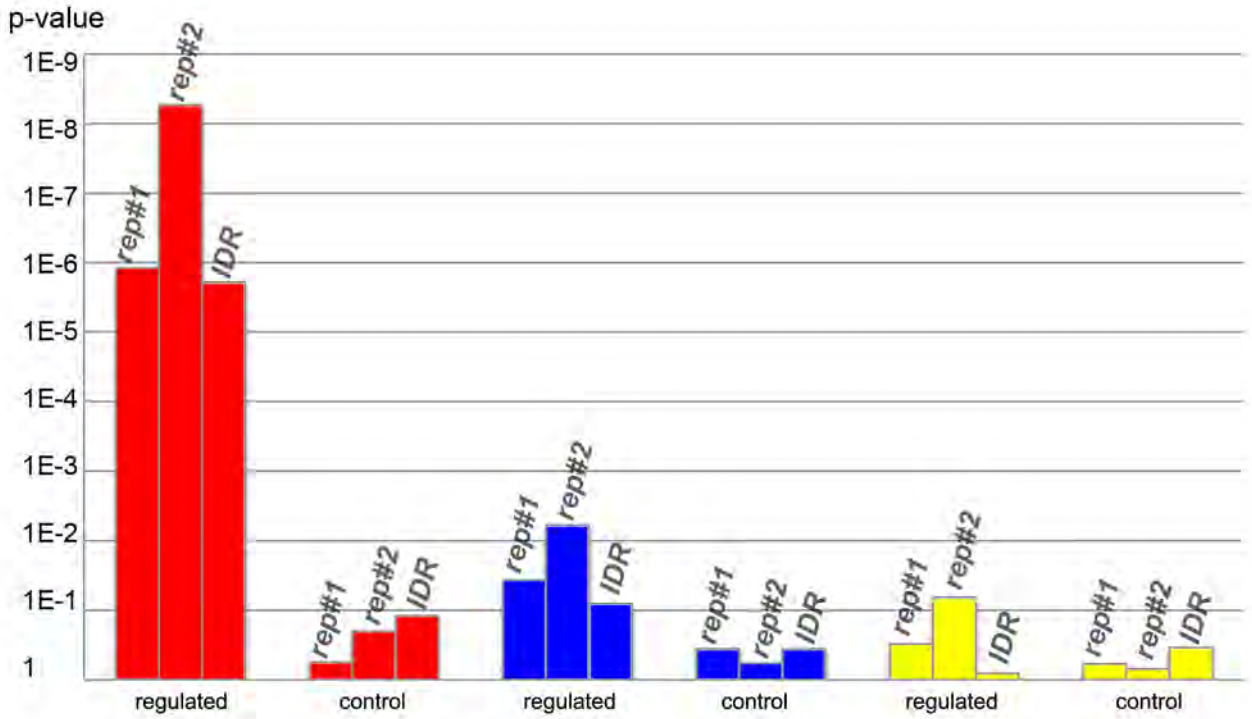
sec23	CG1250	secretory pathway	stripes	salivary gland	94,5	41,8	yes	ND	no	no	no	yes	136	yes	this work, FigS4
CG31559	CG31559	thioredoxin	stripes	foregut	94,8	14,3	yes	yes	no	yes	yes	yes	90,236,92	yes	this work, FigS1
CG11771	CG11771	proteolysis	no	all gut & muscle system	95,0	45,4	ND	ND	yes	no	yes	no		yes	
CG9095	CG9095	cell adhesion	stripes		95,4	36,9	yes	yes	yes	yes	yes	yes	89,265,200,111	yes	this work, FigS4
---	GM01028	unknown	ND	ND	95,6	42,4	ND	ND	no	no	no	yes	79	no	
CG32039	CG32039	unknown	no		95,7	26,7	ND	ND	no	no	yes	yes	122	yes	
CG1753	CG1753	cystein biosynthesis	no	midgut	96,0	26,3	ND	ND	no	yes	no	yes	154,93	yes	
CG4822	CG4822	unknown	ND	ND	96,3	43,6	ND	ND	yes	yes	no	yes	63	no	
CG8112	CG8112	unknown	no	yolk nuclei	96,8	30,2	ND	ND	no	yes	yes	yes	280	yes	
ect	CG6611	tube development	stripes	oesophagus & anal pad	96,9	19,7	yes	no	yes	yes	yes	yes	361	yes	this work, FigS4
CG15239	CG15239	unknown	stripes	salivary gland	97,0	19,4	ND	ND	yes	yes	yes	yes	166	yes	
CG9689	CG9689	unknown	stripes	oesophagus & post spiracle	97,1	43,7	ND	ND	no	yes	no	no		yes	
mRpL45	CG6949	translation & transport	no	ubiquitous	97,2	22,6	ND	ND	no	no	no	no		yes	
CG8213	CG8213	proteolysis	ND	ND	97,3	13,7	ND	ND	yes	yes	yes	no		yes	
CG2263	CG2263	phenylalanyl-tRNA aminoacylation	no	ubiquitous	97,4	46,9	ND	ND	no	yes	no	yes	109	yes	
Rpb8	CG11246	transcription	no	midgut	97,8	45,1	ND	ND	no	no	yes	yes	81	yes	
qua	CG6433	actin binding	stripes	all gut & lymph gland	97,8	43,1	no	no	yes	yes	yes	no		yes	this work, FigS1
CG11227	CG13630	proteolysis	ND	ND	97,8	47,0	ND	ND	yes	no	yes	no		no	
CG9205	CG9205	unknown	ND	ND	98,0	44,9	ND	ND	no	yes	yes	no		yes	
Nf-YA	CG3891	transcription & phagocytosis	no	ubiquitous	98,3	39,6	ND	ND	no	yes	no	yes	91	yes	
kar	CG12286	transmembrane transport	no	midgut & amnioserosa	98,3	42,8	ND	ND	no	no	yes	no		yes	
CG31717	CG31717	unknown	ND	ND	98,4	45,2	ND	ND	yes	yes	yes	no		no	
Past1	CG6148	endocytosis	no	midgut & salivary gland	98,4	46,6	ND	ND	no	yes	yes	yes	66	yes	
bw	CG17632	eye pigment biosynthesis	no	malpighian tubule & fat body	98,4	37,4	ND	ND	no	yes	no	no		yes	
mRpL51	CG13098	translation	no	midgut & muscle system	98,5	39,3	ND	ND	no	no	no	no		yes	
pk	CG11084	planar polarity	no	ubiquitous	98,6	45,5	ND	ND	yes	yes	yes	yes	190,178,125	yes	
CG5171	CG5171	trehalose biosynthesis	no	amnioserosa & yolk nuclei	98,9	37,8	ND	ND	no	yes	no	yes	351,159,153	yes	
CG13365	CG13365	unknown	no		98,9	43,2	ND	ND	no	yes	no	yes	518	yes	
CG5742	CG5742	neurogenesis	no	ubiquitous	98,9	43,8	ND	ND	no	no	yes	no		yes	
PKD	CG7125	intracellular signal transduction	ND	ND	99,0	46,5	ND	ND	no	yes	yes	yes	941,94	yes	
CG11127	CG11127	unknown	no		99,1	47,2	ND	ND	no	yes	yes	no		yes	
Fib	CG9888	centrosome organization	no	ubiquitous	99,1	49,2	ND	ND	no	no	no	no		yes	
CG30423	CG30423	unknown	ND	ND	99,1	43,0	ND	ND	yes	yes	yes	no		no	
CG3842	CG3842	redox process	stripes	foregut & anal pad	99,2	14,1	ND	ND	no	no	no	yes	213,160,74	yes	
CG15743	CG15743	phosphatidylinositol phosphorylation	no	salivary gland	99,9	44,9	ND	ND	no	no	yes	no		yes	

Menoret *et al.;* Figure S3



Menoret *et al.;* Figure S4





Menoret *et al.*; Figure S5

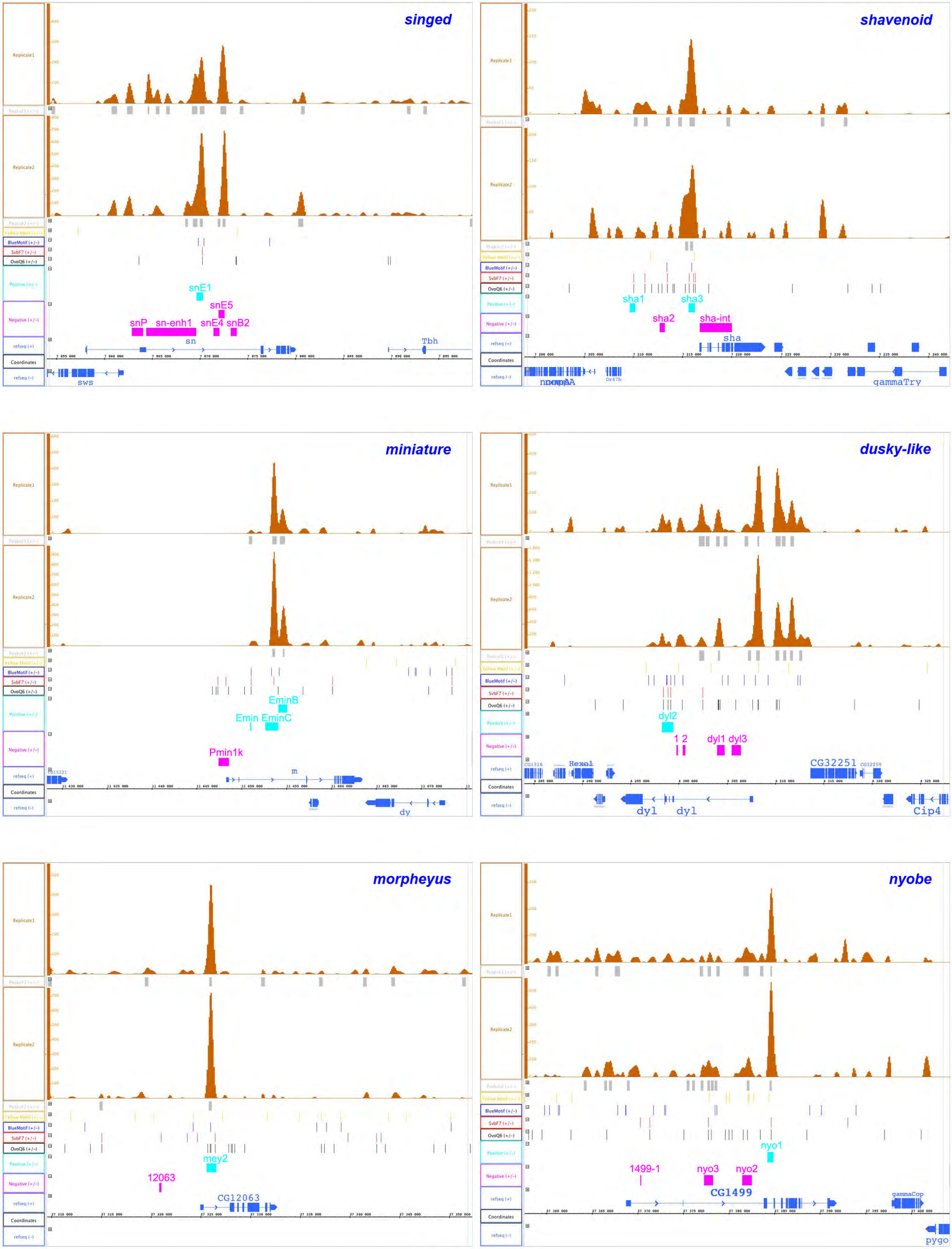
Svb regulated genes

	motif	logo	predicted factor (>70% aligned)
1	aCrCACACaCaC		Klu
2	atAwatAhATAtdTatwtat		ND
3	rktACCGTTAsck		ovo
4	srAGGCATGrg		ND
5	rgGGACTACwa		ND
6	wrCCGGwTAhv		ELK4, ovo odd, grh
7	yaaATATAwATATaTat		bab1
8	cayaCACACACACaC		klu
9	RSTGAAAAGCW		tll

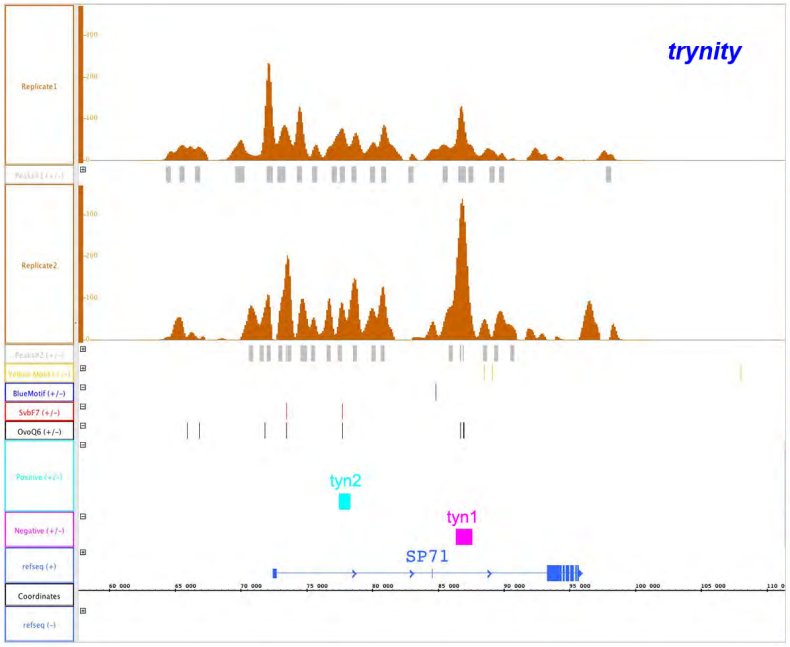
control genes

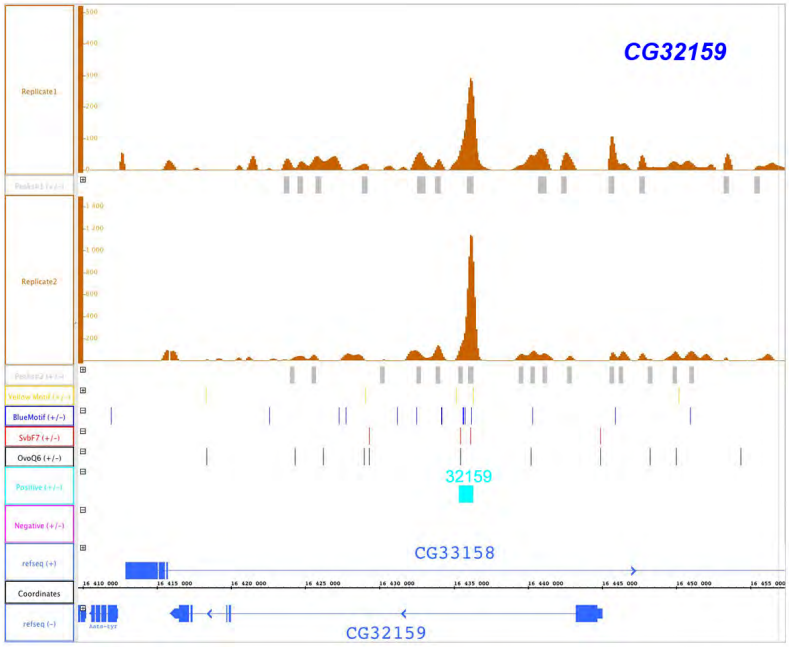
	motif	logo	predicted factor (>70% aligned)
1	tayrTATGTAyrt		ND
2	rsAAAAAyAAAak		jim
3	sraCAGCTGtys		NHLH1, sage
4	wwkCTAGTrCCbt		ND
5	rcGCGCGCGCgy		ND
6	raCAGCTGty		NHLH1, Myf, cato, sage
7	bataTACATACata		FOXL1
8	mmACAAmAACAaAcrc		ND

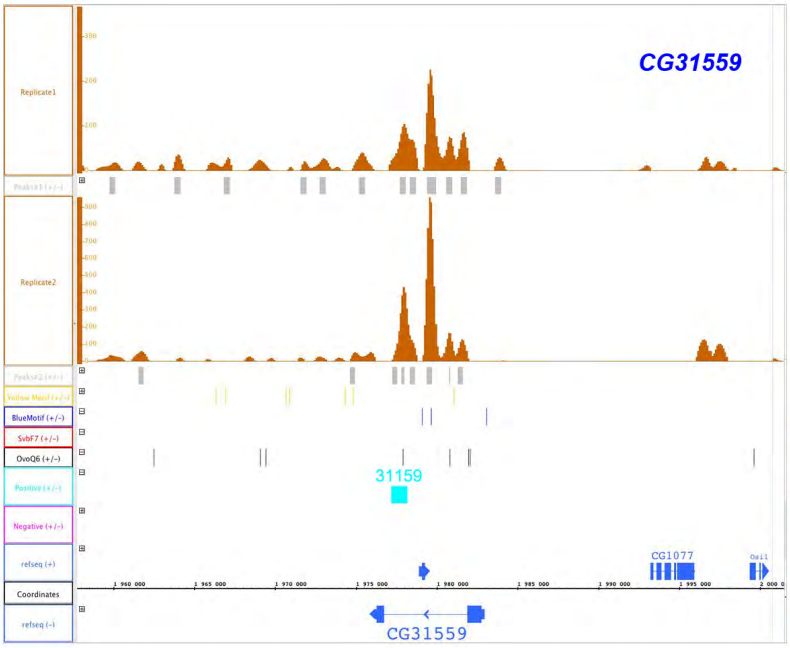
Menoret *et al.;* Figure S6

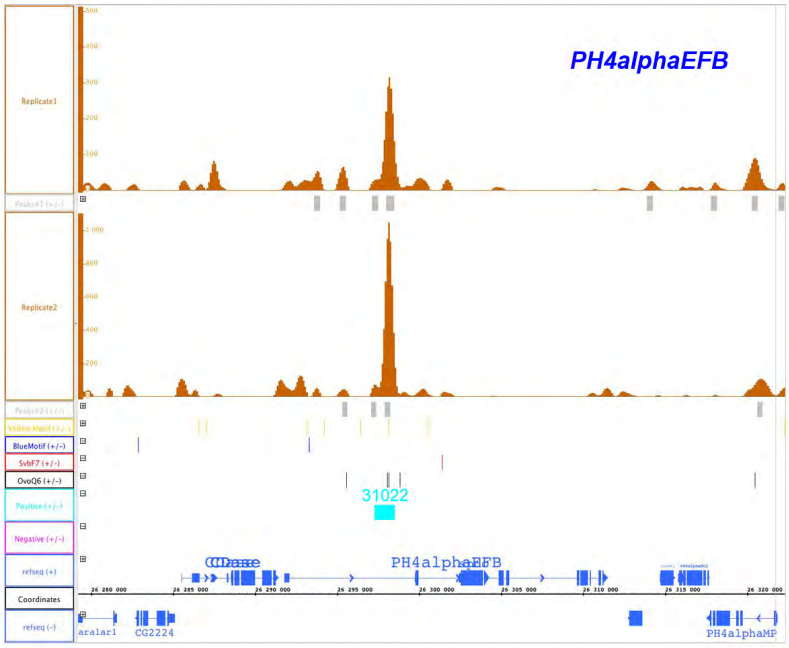


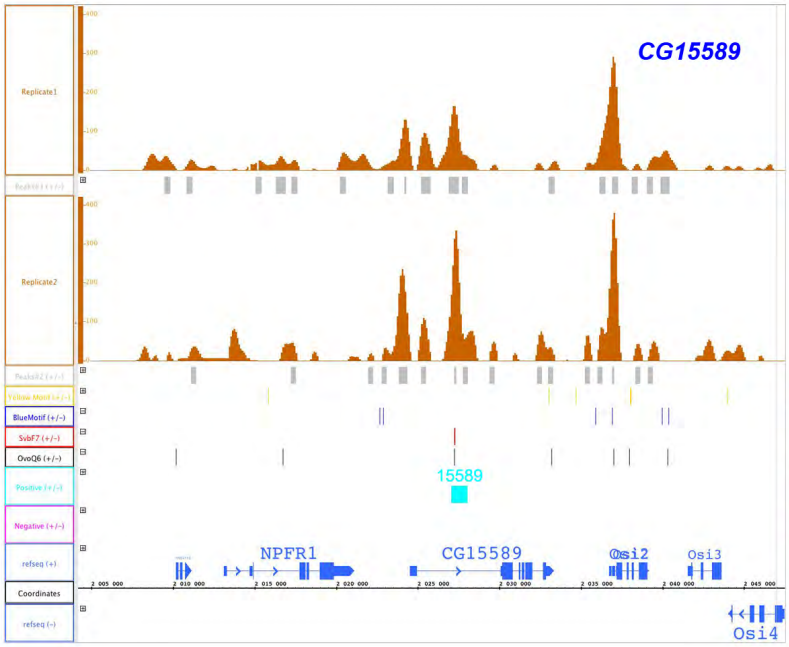
Menoret et al.; Fig S7A

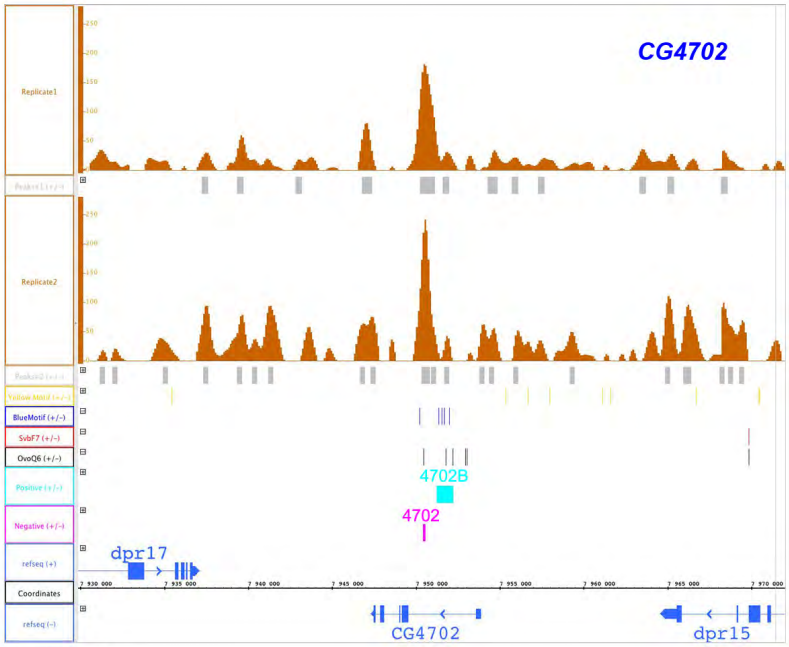






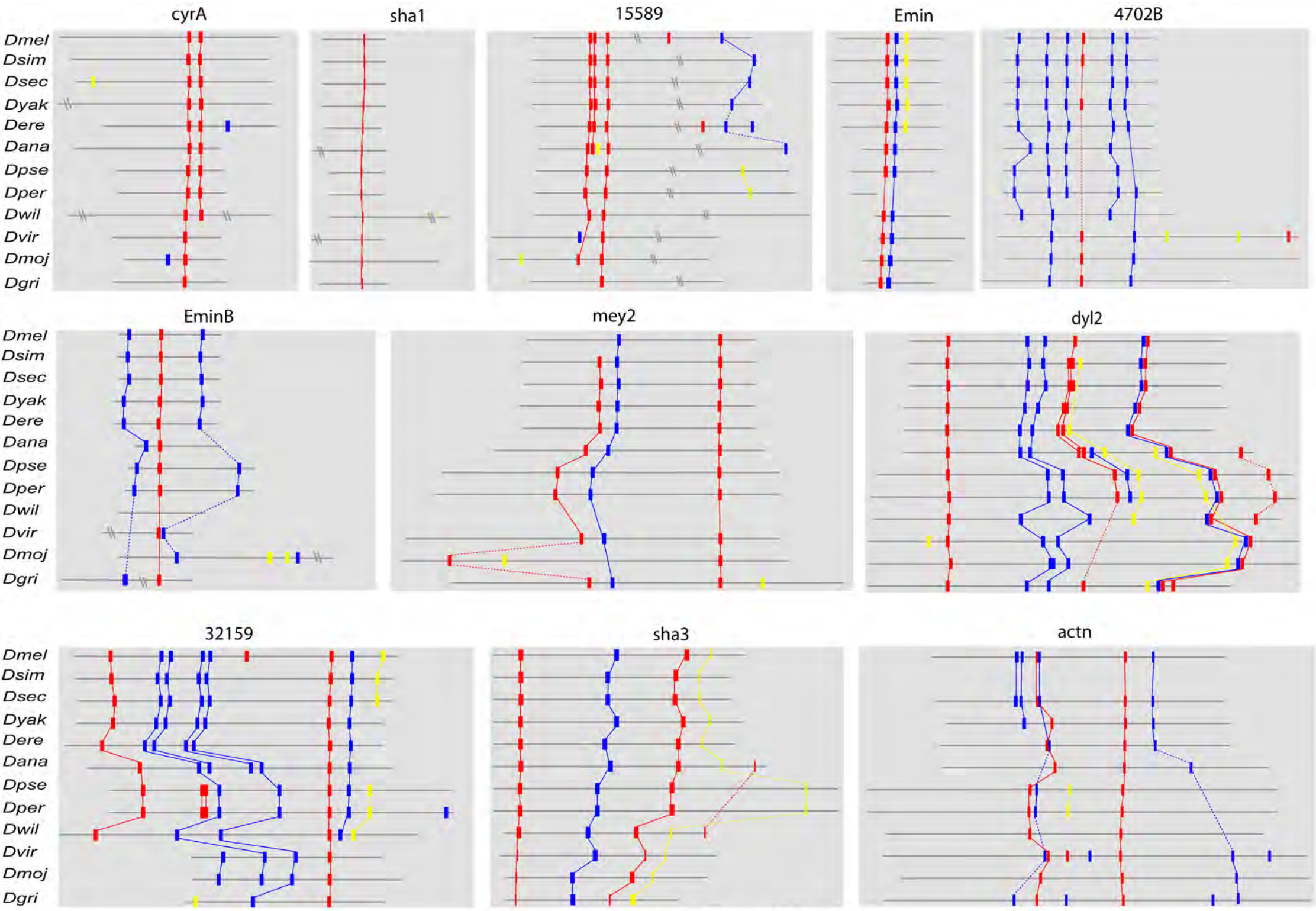




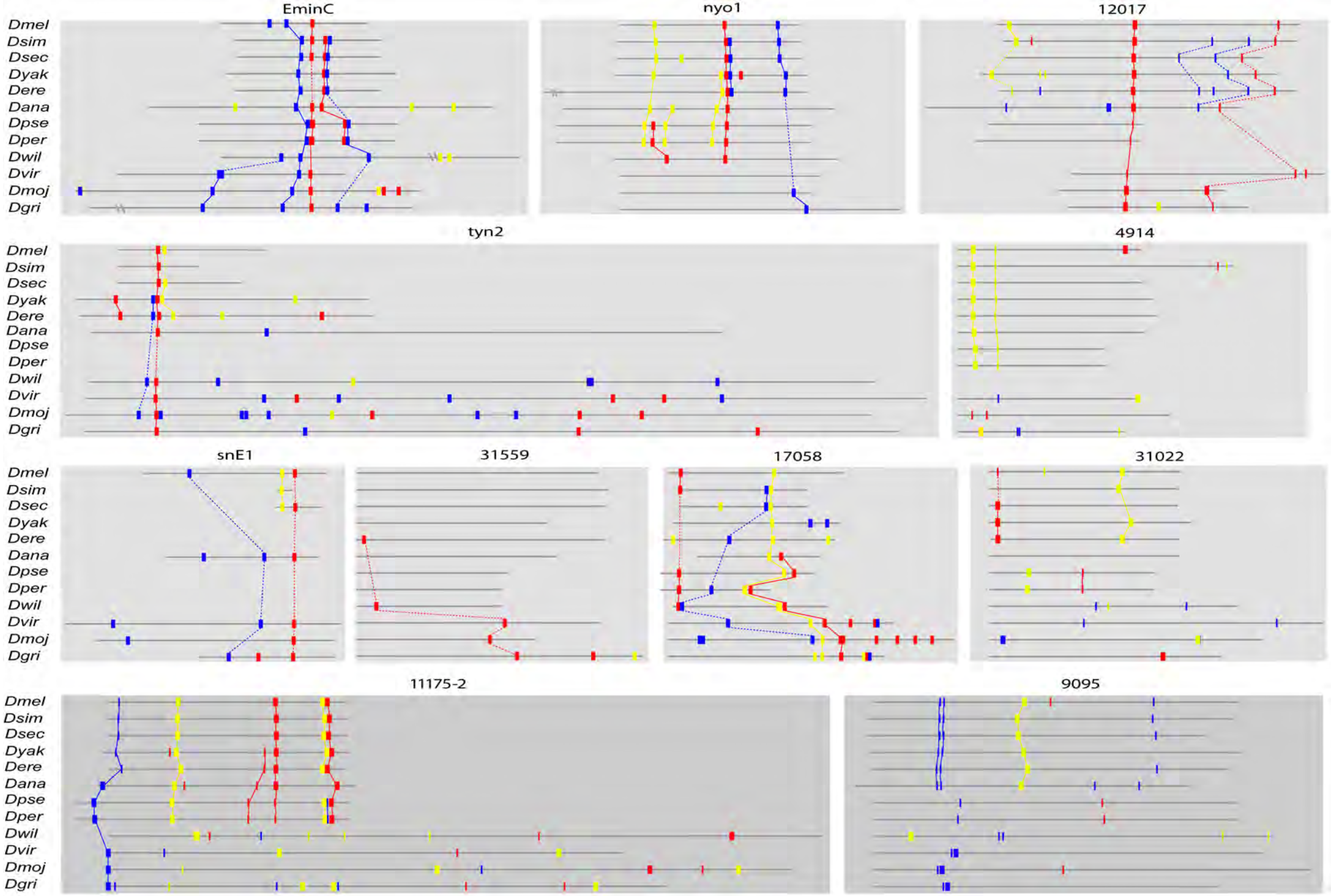


Menoret et al.; Fig S7B

Menoret et al.; Fig S7C



Menoret *et al.*; Figure S8A



Menoret *et al.*; Figure S8B

4.3 Conclusion et perspectives du chapitre 4

Imogene a été appliqué sur un ensemble d'apprentissage composé de 14 CRMs régulant la différenciation des trichomes chez l'embryon de Drosophile. Les différents paramètres (seuils de génération, de détection) ont été optimisés par une approche de Pareto visant à maximiser le nombre de CRMs positifs prédits par les motifs tout en minimisant le nombre de CRMs négatifs prédits parmi un ensemble de 25 CRMs n'ayant aucune activité au stade de développement considéré. Deux motifs ont été trouvé par cette approche : le motif svbF7 correspondant à *svb*, le régulateur maître de la différenciation des trichomes, et un nouveau motif, le motif bleu, que nous n'avons pas pu associer à un motif connu.

La validité de ces motifs a été montrée par mutagenèse (figures 3, 4 et 5 de l'article). Par ailleurs, ces motifs sont prédictifs des ChIP-seq de *svb* fonctionnels, c'est-à-dire ceux qui sont associés à un gène dont l'expression diminue chez les mutants *svb* (figure S5). Les CRMs fonctionnels possèdent une grande variété de grammaire des sites de fixation (figures 5 et 7 de l'article), un résultat similaire à celui obtenu par Zinzen et al. (2009) dans le cas de la différenciation de différents tissus chez l'embryon de drosophile. Plusieurs entrées (combinaisons de TFs) mènent à une sortie (motif d'expression du gène rapporteur) similaire : cette flexibilité est réminiscente du modèle *billboard* introduit en section 1.5.2. Néanmoins, bien qu'il soit clair que la grammaire des sites soit différente entre différents CRMs, cette grammaire semble relativement bien conservée au cours de l'évolution d'un CRM (figures 8, S8A, et S8B). Enfin, dans la plupart des cas on observe l'absence de *clustering* de motifs homotypiques sur les CRMs, bien que les motifs soient présents en plus grand nombre dans les loci des gènes régulés par rapport à des gènes non régulés (figure S2). Une explication possible est que 'abondance de sites dans l'environnement du CRM permet d'augmenter localement la concentration du TF pour faciliter son recrutement *in vivo* au niveau des CRMs possédant des sites de forte affinité.

Il serait à présent intéressant de caractériser plus en avant le motif bleu généré par l'approche *de novo*. Une possibilité serait d'utiliser la technique de simple hybride présentée en 1.4.2 afin d'identifier la protéine associé au motif bleu, en utilisant comme appât les protéines connues de la drosophile et comme proie le site consensus du motif bleu.

Chapitre 5

Étude de la différenciation musculaire chez la souris

5.1 Introduction à la myogenèse squelettique **219**
5.1.1 Les différentes étapes de la myogenèse 219
5.1.2 Les Facteurs de Régulation Myogénique et leurs cofacteurs 219
5.1.3 Les homéoprotéines Six 221
5.2 Intégration de données bioinformatiques et expérimentales sur le muscle **223**
5.2.1 Obtention d'une PWM optimale pour MEF3 223
5.2.2 Obtention de données à partir de la littérature et intégration au visualisateur de UCSC 226
5.3 Prédictions et validations de la régulation par Six **229**
5.3.1 Régulation de *Myod1* par les protéines Six chez l'embryon de souris 229
5.3.2 Régulation d'un lincRNA par Six dans le muscle adulte 231
5.4 Synergie entre Six et MyoD au cours de la myogenèse **265**
5.4.1 État de l'art sur la coopération Six/MyoD 265
5.4.2 Obtention de données d'expression Six+MyoD 266
5.4.3 Obtention de régions de régulation Six+MyoD 268
5.4.4 Croisement des données d'expression et des enhancers putatifs 269
5.4.5 Validation des enhancers MyoD+MEF3 272
5.4.6 Recherche de motifs et mutagenèses 273
5.5 Conclusion et perspectives du chapitre 5 . . . . . **280**

Introduction du chapitre 5

Dans cette dernière partie, nous avons cherché à tester expérimentalement les prédictions de l'algorithme Imogene chez les mammifères. Nous nous sommes pour cela intéressé à la différenciation musculaire, ou myogenèse, chez la souris. Contrairement au cas des trichomes introduit précédemment, nous n'avions initialement pas de CRMs tissu-spécifiques sur lesquels directement tester Imogene. Afin d'obtenir de tels CRMs et d'étudier la logique de régulation conférant la tissu-spécificité chez les mammifères, nous avons collaboré avec l'équipe de Pascal Maire à l'Institut Cochin, qui m'a accueilli et m'a permis de réaliser une partie des expériences présentées, avec l'aide de Iori Sakakibara.

Nous avons commencé par bâtir un atlas de la régulation de la myogenèse en regroupant diverses données bioinformatiques et expérimentales. Nous avons notamment centré notre attention sur les homéoprotéines Six1 et Six4 (*Sine Oculis Homeobox Homolog* 1 et 4, référées dans la suite par Six1,4), des facteurs de transcription centraux impliqués dans la régulation des stades successifs de la myogenèse. Ils activent en effet les TFs nécessaires à l'engagement de cellules pluripotentes dans la voie myogénique et à leur différentiation : Pax3 (*Paired Box* 3) ainsi que les Facteurs de Régulation Myogénique Myf5 (*Myogenic Factor* 5), Mrf4 (*Muscle-Specific Regulatory Factor* 4, aussi appelé Myf6), Myod (*Myogenic Differentiation*) et Myog (*Myogenin*). Nous avons d'abord réalisé un modèle PWM des sites MEF3 de fixation des protéines Six, que nous avons ensuite utilisé pour prédire des sites à l'échelle du génome. Nous avons par ailleurs récupéré de nombreuses données relatives à la différenciation musculaire : ChIP-seq de TFs et des marques épigénétiques des histones, données d'expression de type RNAseq, sites de fixation conservés que nous avons obtenus par analyse bioinformatique, etc. Ces données ont été regroupées sur le visualiseur de UCSC, permettant d'envisager facilement le contexte de régulation de certains gènes d'intérêt. Nous présentons plusieurs validations de prédictions obtenues par analyse bioinformatique avec ces données.

Ensuite, nous avons voulu utiliser cet atlas pour générer un ensemble de CRMs sur lesquels nous pourrions tester Imogene. Pour cela, nous nous sommes intéressé à l'action concertée ou *synergie* entre Six1,4 et le TF maître MyoD au cours de la différenciation musculaire. Nous avons trouvé un certain nombre de CRMs fixés par MyoD (données ChIP-seq), possédant un site MEF3 conservé chez les mammifères, et dont le gène le plus proche n'est activé qu'en présence de MyoD et de Six1,4. Parmi les gènes en question figurent *Myog*, TF requis pour la différenciation terminale, et des gènes structuraux comme *Tnnc1* (*Troponin C Type* 1) ou *Ttn* (*Titin*). Nous avons testé l'activité de ces CRMs putatifs et en avons trouvé 70% qui récapitulent l'expression du gène le plus proche lorsqu'ils sont testés par transfection avec un rapporteur Luciférase. Ceux-ci constituent alors un ensemble d'apprentissage adéquat pour tester Imogene. Différents corégulateurs ont ainsi été prédits, et nous présentons des tests expérimentaux de mutagenèse des sites de fixation correspondants.

5.1 Introduction à la myogenèse squelettique

5.1.1 Les différentes étapes de la myogenèse

La myogenèse correspond à la formation des tissus musculaires, ceux-ci étant regroupés en trois types majeurs : les muscles cardiaques, les muscle lisses et les muscles squelettiques. C'est la formation de ces derniers qui nous intéresse ici. Les muscles squelettiques sont composés de fibres musculaires polynucléées provenant de la fusion de progéniteurs musculaires appelés myoblastes. Ils sont sous le contrôle du système nerveux central, et composent l'un des organe majeur des vertébrés, concentrant $\sim$ 40% du poids corporel. La myogenèse squelettique débute relativement tôt au cours de l'embryogenèse : à 8.5 jours [17] après fécondation (ou E8.5 pour *Embryonic day* 8.5) chez l'embryon de souris sur un total de 18 jours embryonnaires. Elle a lieu au niveau des somites, structures périodiques situées au niveau des futures vertèbres (fig. 5.1a). Les étapes de fusion des myoblastes et de maturation des fibres s'étalent ensuite sur toute l'embryogenèse (fig. 5.1b), ainsi que dans le muscle adulte lors de la régénération musculaire (Parker et al., 2003).

5.1.2 Les Facteurs de Régulation Myogénique et leurs cofacteurs

D'un point de vue génétique, la myogenèse des vertébrés est coordonnée en partie par l'action de 4 Facteurs de Régulation Myogénique (MRF pour *Myogenic Regulatory Factors*) : MyoD, Myf5, Mrf4 et Myog, qui font partie de la famille de protéines *basic Helix-Loop-Helix* (bHLH) et se fixent sur les boîtes E (ou E-box) du type CANNTG. Ces MRFs sont activés successivement à travers une cascade de régulation génétique, et se régulent les uns les autres. Par exemple, Myf5, MRF4 et MyoD peuvent activer MyoD ; Myf5, MyoD et Mrf4 régulent l'expression de Myog ; et Myog peut activer l'expression de Mrf4 (Naidu et al., 1995). En outre, MyoD et Myog peuvent s'auto-activer (Thayer et al., 1989). Enfin, un certain degré de redondance a été observé entre Myf5, Mrf4 et MyoD au cours du développement embryonnaire de la souris, et l'analyse des embryons KO pour ces gènes a conduit à la conclusion qu'ils ont tous la capacité d'activer la myogenèse à partir de cellules embryonnaires pluripotentes et d'agir comme des gènes de détermination (Kassar-Duchossoy et al., 2004).

Ces MRFs sont la clé de voûte de la différenciation myogénique. Cependant, la régulation précise du réseau de gènes activés par les MRFs nécessite leur coopération avec d'autres facteurs de transcription. En particulier, la liaison de MyoD à l'ADN n'est pas prédictive d'une activité enhancer (Cao et al., 2010). Ainsi, Molkentin and Olson (1996) ont montré que l'activation transcriptionnelle par les MRFs est renforcée par la fonction des TFs à boîte MADS MEF2. Leur importance a ultérieurement été confirmée par les travaux de Blais et al. (2005) : en utilisant un grand nombre de gènes cibles des MRFs, les auteurs ont cherché dans leur région promotrice des sites de liaison sur-représentés pour les TFs issus de la base de données Transfac. Ils ont ainsi constaté que l'élément riche en A/T reconnu par MEF2 était parmi eux, et ont confirmé son recrutement à plusieurs de ces sites. Par ailleurs, un autre motif d'ADN trouvé comme spécifiquement enrichi parmi les promoteurs cibles des MRFs était l'élément MEF3 (consensus GAAACCTGA), le site de liaison des homéoprotéines Six. Cette séquence MEF3 était particulièrement abondante au sein des promoteurs des gènes dont l'expression était induite de manière significative au cours de la différenciation.

17. La mi-journée correspondant au fait que la fécondation a lieu la nuit.

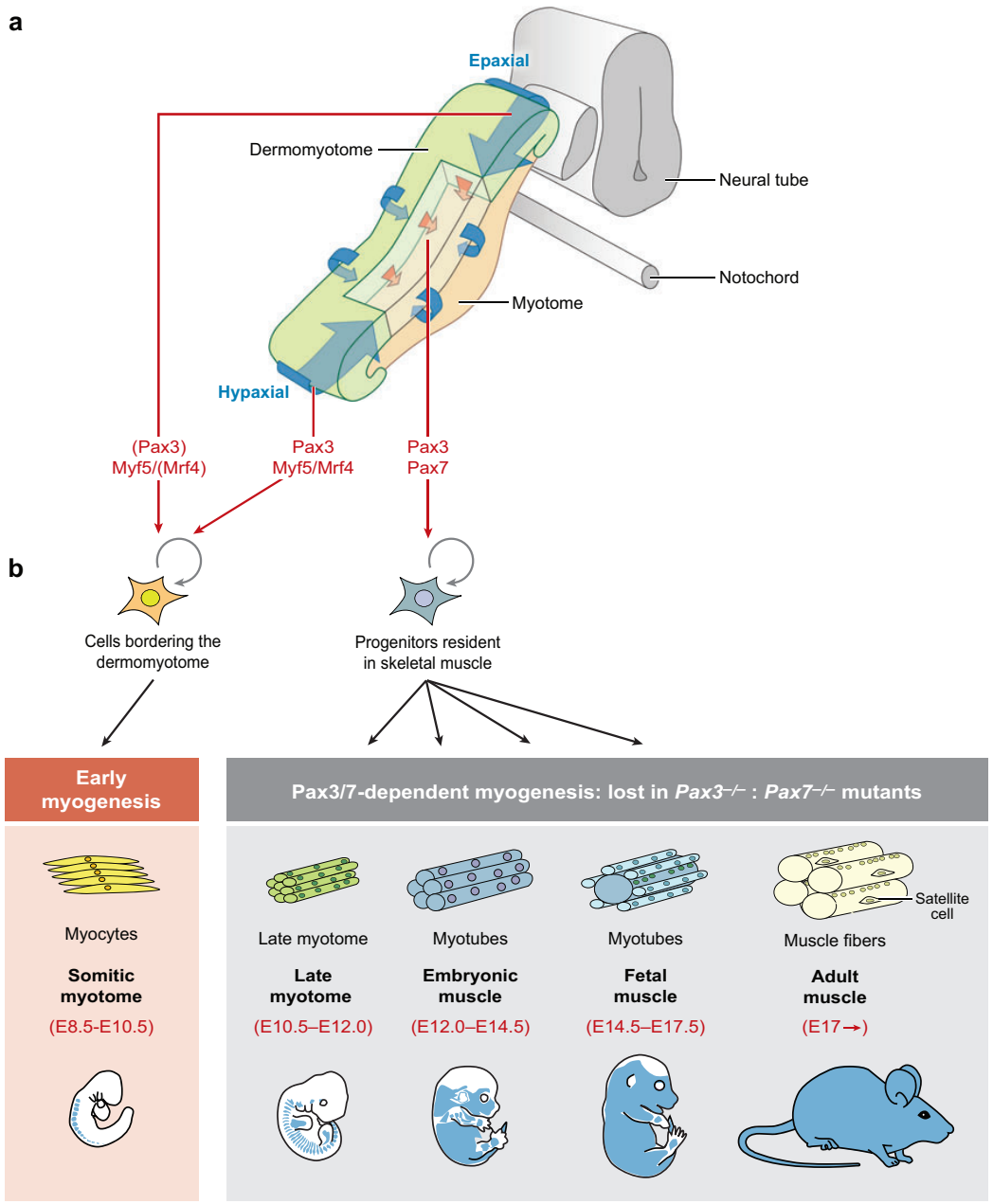


FIGURE 5.1 – **Formation du muscle squelettique chez l'embryon de souris.**
Figure tirée de Buckingham and Relaix (2007) illustrant les différentes étapes de la myogenèse. (a) Le dermomyotome épithélial d'un somite (vert) ainsi que le muscle squelettique du myotome (beige) sont d'abord formés par la délamination des cellules provenant des extrêmités du dermomyotome (flèches bleues) : c'est la première vague de la myogenèse. Ensuite, lors d'une deuxième vague, le dermomyotome central perd sa structure épithéliale et des cellules musculaires progénitrices s'intègrent au myotome (flèches rouges). Les TFs précurseurs de ces événements sont indiqués en rouge. (b) Schéma montrant les étapes de différenciation des progéniteurs musculaires, ainsi que les temps de développement associés (E : jour embryonnaire).

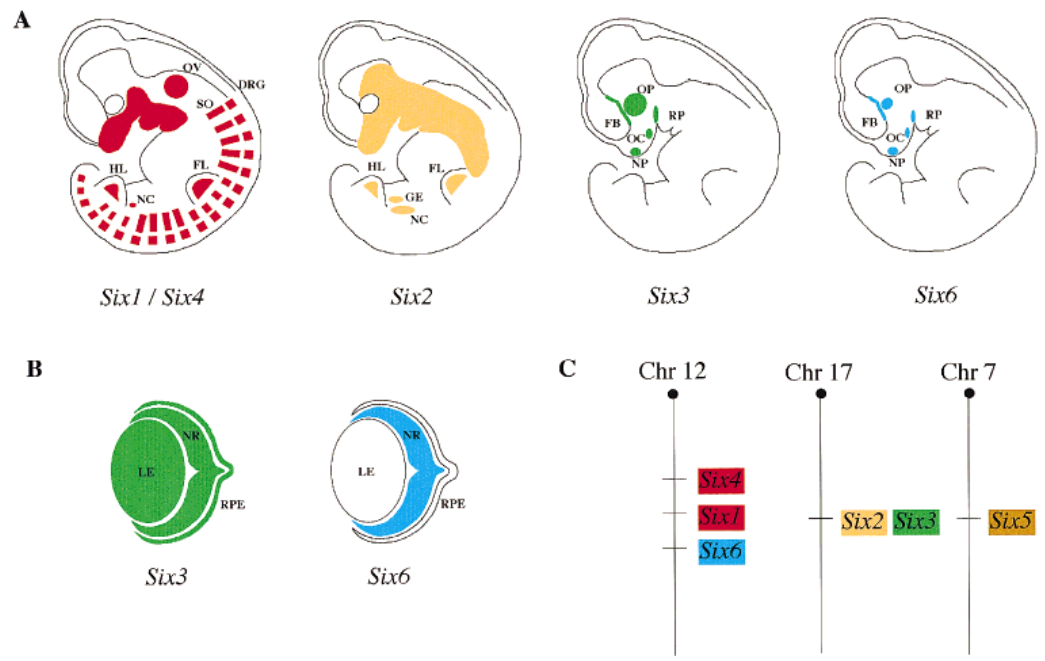


FIGURE 5.2 – **Motifs d'expression et localisation génomique des gènes de la famille Six.** Figure tirée de (Kawakami et al., 2000), montrant les motifs d'expression des gènes de la famille Six chez l'embryon (A) et dans l'oeil (B) à E13.5. Les couleurs similaires reflètent des motifs d'expression similaires. La localisation chromosomique des gènes est aussi montrée en (C).

5.1.3 Les homéoprotéines Six

Les homéoprotéines Six associées aux sites MEF3 possèdent un rôle étendu au cours de la myogenèse, et nous en présentons maintenant les détails. Parmi les 6 gènes appartenant à la famille des homéoprotéines Six (fig. 5.2), *Six*1 et *Six*4 ont reçu une attention particulière dans le contexte de la myogenèse squelettique, bien que d'autres membres comme *Six*2 ou *Six*5 semblent être en mesure de compenser partiellement leur perte (Relaix et al., 2013). Ces deux gènes se trouvent à proximité l'un de l'autre sur le chromosome 12 de la souris et sont séparés par 100kb de séquence intergénique. L'équipe de Pascal Maire a déjà identifié plusieurs fonctions clés de *Six*1 dans le lignage musculaire au cours de l'embryogenèse, par l'analyse de modèles de souris dépourvues de *Six*1 (Six1KO) et de Six1 et Six4 (Six1,4dKO), et par des expériences de ChIP. L'analyse des embryons Six1,4dKO (Grifone et al., 2005) à E10 (Niro et al., 2010) et E18 (Richard et al., 2011) a permis d'identifier des cibles génétiques directes et indirectes de Six1, 4, et a dévoilé plusieurs fonctions des homéoprotéines Six pendant le développement musculaire et la spécialisation des fibres musculaires.

Les principaux points sont les suivants (fig. 5.3) : Six1 et Six4 sont requis pour la genèse des progéniteurs myogéniques hypaxiaux (c'est-à-dire ventraux). En l'absence de Six1 et Six4, les cellules ventrales pluripotentes du dermomyotome somitique ne parviennent pas à adopter un destin de progéniteur myogénique et elles n'expriment ni *Pax3* ni les MRFs. Six1 et Six4 contrôlent directement l'expression de *Pax*3 (Grifone et al., 2007), et donc l'engagement des cellules pluripotentes dermomyotomales dans un destin de progéniteur myogénique. Plus tardivement, l'expression de *Myf*5 dans le membre est contrôlée par un enhancer « membre »

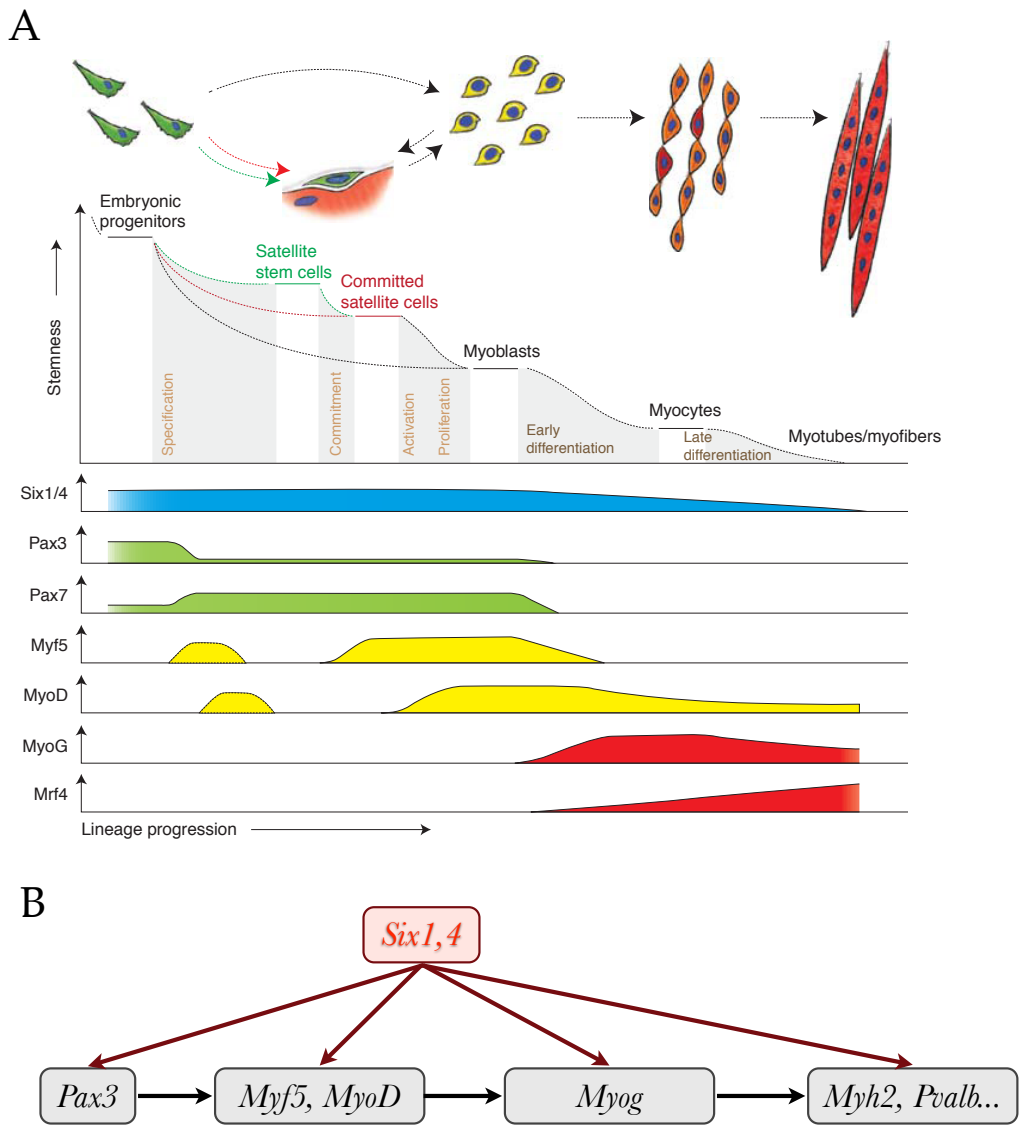


FIGURE 5.3 – **Rôle des protéines Six au cours de la myogenèse.**
(A) Figure tirée de Bentzinger et al. (2012). La myogenèse embryonnaire consiste en plusieurs étapes de différenciation musculaire : la genèse des progéniteurs embryonniques exprimant *Pax3*, leur spécification en myoblastes par l'induction des gènes de détermination myogénique comme *Myf5* et *Myod1*, la différenciation des myoblastes en myocytes avec l'expression de *Myog* et leur fusion en myotubes qui se spécialisent plus tardivement en fibres de type lentes (oxydatives) ou rapides (glycolytiques). On note que les protéines Six sont exprimées continuellement au cours de ces étapes (et aussi plus tardivement dans les fibres rapides).
(B) Les homéoprotéines Six interviennent dans la régulation de ces différentes étapes *in vivo* en régulant les TFs indiqués selon une cascade de boucles *feed-forward*.

lié par les protéines Six1 et Six4, et cette liaison est nécessaire pour permettre l'expression de Myf5 (Giordani et al., 2007). Par ailleurs, deux enhancers contrôlent spécifiquement l'expression de MyoD au cours de l'embryogenèse et au cours de la régénération du muscle adulte : le *Distal Regulatory Region* ou DRR (Asakura et al., 1995) et le *Core Enhancer* ou CE (Kucharczuk et al., 1999). La liaison de ces éléments par Six1 et Six4 a été confirmée par ChIP au cours de l'embryogenèse (Relaix et al. (2013), voir section 5.3.1) ainsi que dans des cultures de cellules satellites (Le Grand et al., 2012), qui sont des cellules souches du muscle adulte activées lors de la régénération musculaire. L'expression de Mrf4 n'est pas détectée chez les embryons Six1,4dKO à E10 (Grifone et al., 2007). Myog est aussi sous le contrôle direct des protéines Six (Spitz et al., 1998). Enfin, les protéines Six régulent des gènes spécifiques des fibres glycolytiques de type rapide (voir section 5.3.2). Par exemple, la mutation du site MEF3 au sein du promoteur de *Aldoa*, gène codant pour une enzyme impliquée dans la glycolyse, entraîne l'abolition de l'expression d'un transgène rapporteur dans des myofibres adultes (Spitz et al., 1997).

5.2 Intégration de données bioinformatiques et expérimentales sur le muscle

La première étape de notre collaboration avec l'équipe de P. Maire a été de construire un modèle de motif pour les sites de fixation MEF3 associés aux protéines Six1,4 chez la souris. Un tel motif, qui n'existait jusqu'alors pas, permet de réaliser des prédictions précises, et nous discuterons après de leur vérification. Par ailleurs, nous avons récupéré de nombreuses données expérimentales dans la littérature sur la différenciation musculaire (ChIP-seq, RNAseq, données épigénétiques...) et les avons intégrées au système de visualisation de UCSC (présenté en introduction, section 1.7.3). L'intégration de ces multiples données permet d'appréhender facilement et rapidement les phénomènes de régulation ayant lieu lors de la différenciation musculaire à l'échelle locale de gènes d'intérêt comme à l'échelle globale génomique en croisant les données.

5.2.1 Obtention d'une PWM optimale pour MEF3

Avant notre arrivée au laboratoire de P. Maire, il n'existait pas de PWM pour MEF3, et les prédictions étaient faites sur la base de la proximité au consensus GAAACCTGA issu du promoteur de Myog. Nous avons donc décidé de construire un motif pour les sites de fixation MEF3 associés aux protéines Six1,4. Pour ce faire, nous avons utilisé le fait qu'un certain nombre de sites de fixation avaient déjà été testés par retard sur gel (EMSA ou *Electrophoretic Mobility Shift Assay*), expérience permettant de détecter une interaction entre une protéine et de l'ADN, que ce soit dans la littérature ou au sein du laboratoire de P. Maire. Ces expériences ont permis de définir 32 sites positifs (table 5.1) et 14 sites négatifs (table 5.2). Les sites positifs ont ensuite été partagés en un ensemble d'apprentissage composé de 14 séquences pour lesquelles les séquences orthologues chez les autres mammifères étaient disponibles, et en un ensemble « test » indépendant de 18 sites.

Nous avons alors cherché la PWM apprise sur l'ensemble d'apprentissage qui distinguait le mieux les sites positifs de l'ensemble test des sites négatifs. Ces derniers ayant été choisis en fonction de leur resemblance avec le site consensus de MEF3 sur le promoteur de *Myog*, le fait de trouver une PWM distinguant sites positifs des négatifs permet véritablement d'améliorer l'approche heuristique. Nous avons généré deux types de PWMs : une PWM « référence »

Séquences positives d'apprentissage	
Myog	GAAACCTGA
Pax3	GAAATCTAA
Myf5	GTAACTGGA
Myod-DRR	GAAACCGGA
Mlc-hox	GTAATTTAA
Atp2a1-1	GTAACTGGA
Atp2a1-2	GTAACCTGA
Tnnc2	GAAATTTAA
Nrk	GCAAGGCGA
Sarcalumenin	GGAACTTGA
Myl1-2	GAAATTGAA
Ifitm3	GTAATTTGA
Mybph-2	GAAATCTGA
Myeov2	GAAACTTGA

Séquences positives test	
Aldoa	GAAACCTGA
Ato	GTCATTTGA
Kcne1l	GATAACGGA
Myf5	GAAATTTAA
Pax3-1	GAAATGTAA
Pax3-2	GTTACTGGA
Pvalb	GTAACCTGA
Utrophine	GTCACCTGA
Na-K-ATPase	GCAACCTGA
Myf5-2	GCAACCTGA
Myf5-sat	GAAATCTGA
Lbx1	GCCACCTGA
MCK	GACACCCGA
IgfBp5	GCAATTTGA
Troponine-C	GTAACCTGA
Sarcalumenin	GGAACTTGA
Nrk	GCAAGGCGA
Tp4	GCAAGCAGA

TABLE 5.1 – **Sites MEF3 positifs.**
Le tableau de gauche correspond aux 14 sites conservés chez les mammifères que nous avons utilisés pour l'apprentissage de la PWM MEF3, et le tableau de droite aux sites utilisés comme un ensemble test indépendant de Vrais Positifs dans la courbe ROC de la figure 5.4.

Séquences négatives	
myf5-1	GACAGTGGA
myf5-2	GTAACCTCA
myf5-3	GTAACTGGG
pax3-1	GGAACTTGA
pax3-2	GTATTAATA
pax3-3	GGATAAAGA
pax3-4	GCTAATTGA
pax3-5	GAAAGATTA
pax3-6	GCTCTCTGA
pax3-7	GAGCCCTGA
pvalb-1	GCACAATGA
pvalb-2	GCAGGCTGA
unknown-1	GCAATCTGA
unknown-2	GAGTCCTGA

TABLE 5.2 – **Sites MEF3 négatifs.**
Sites négatifs utilisés comme Faux Positifs dans la courbe ROC de la figure 5.4.

apprise sur les 14 séquences d'apprentissage, et une PWM « avec évolution » apprise avec Imogene à partir des alignements des séquences d'apprentissage avec leurs orthologues. Le seuil de génération des PWMs a été varié entre 7 et 12 bits, et les deux modèles d'évolution (*Felsenstein* et *Halpern-Bruno*) ont été utilisés. Pour chaque PWM, une courbe ROC (pour *Receiver*

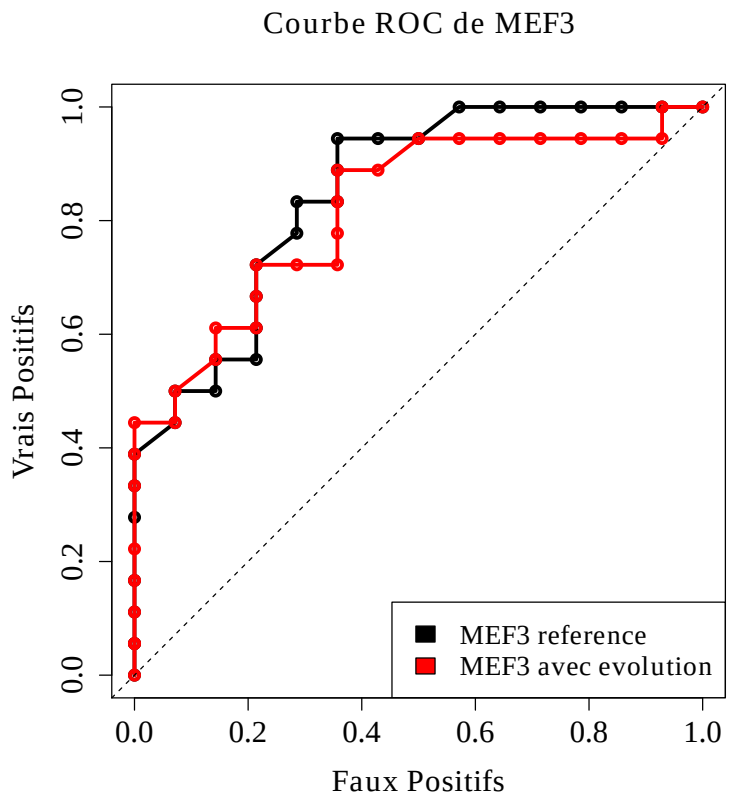


FIGURE 5.4 – **Courbe ROC de MEF3.**
Nous comparons la capacité de classification de sites MEF3 positifs ou négatifs (par EMSA) pour différentes PWMs : chaque point représente le taux de Vrais Positifs et de Faux Positifs prédits au-dessus d'un seuil de détection donné. Les PWMs ont été obtenues à partir de 14 sites positifs différents des sites utilisés dans la classification, soit en utilisant les sites orthologues chez d'autres mammifères et un modèle d'évolution (courbe rouge) ou en utilisant juste les sites de l'espèce de référence (courbe noire). Dans chaque cas, les paramètres (seuil de génération fixant la valeur du pseudo-count, modèle d'évolution) ont été variés et seule la meilleure courbe ROC est montrée. Le cas avec évolution est meilleur à haut seuil (inflexion initiale pour $S_s > 6.5$bits) que le modèle sans évolution. Les paramètres associés sont : $S_g = 8.7$ bits et modèle Halpern-Bruno pour le cas avec évolution, $S_g = 12$ bits pour le cas référence.

Operating Characteristic) peut être construite indiquant pour différents seuils de détection S_s la proportion de séquences positives détectées dans l'ensemble test (Vrais Positifs) et la proportion de séquences négatives détectées (Faux Positifs ou FPs). Nous montrons en figure 5.4 les meilleures courbes ROC obtenues pour les cas référence et avec évolution. Les deux courbes sont relativement similaires et permettent de détecter 45% des sites positifs sans détecter aucun négatif pour un seuil de l'ordre de $S_g = 7$ bits. La PWM avec évolution (fig. 5.5) montre une légère amélioration du signal à haut seuil (petits FPs) par rapport à la PWM référence, et nous avons donc utilisé celle-ci dans la suite de notre travail.

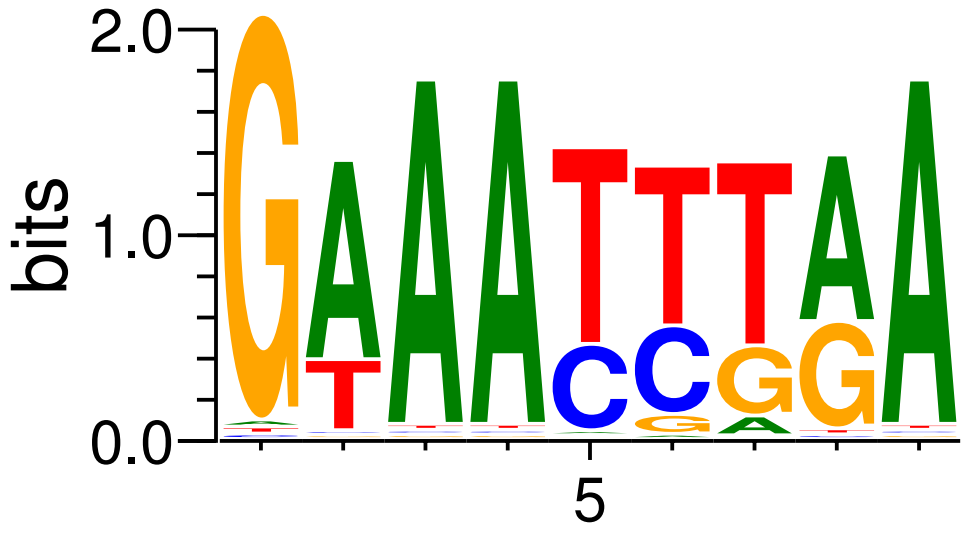


FIGURE 5.5 – **PWM MEF3 obtenue avec le modèle d'évolution.**

5.2.2 Obtention de données à partir de la littérature et intégration au visualisateur de UCSC

La littérature regorge de données publiées obtenues dans différents modèles musculaires, que ce soit des cultures de cellules C2C12, lignée cellulaire de myoblastes de souris couramment utilisée pour modéliser la différenciation musculaire, ou dans des tissus provenant de la dissection de muscles squelettiques chez l'embryon ou chez l'adulte. Néanmoins, il n'existe pas d'outil centralisant ces données spécifiques au muscle. Nous avons donc récupéré ces données et les avons intégrées sur l'outil de visualisation Genome Browser de UCSC afin de faciliter l'analyse des événements de régulation au cours de la différenciation musculaire (fig. 5.6).

Parmi ces données figurent d'abord des expériences issues du projet ENCODE (présenté dans l'introduction en section 1.7.4). Celles-ci sont de plusieurs types et ont été obtenues par des laboratoires différents. Il y a notamment des donnés RNAseq, quantifiant précisément la transcription d'ARN dans différentes conditions (C2C12 à Caltech, muscle squelettique à University of Washington ou UW), des données d'hypersensibilité à la digestion par DNAseI (voir 1.4.3) en muscle adulte (UW) ainsi que diverses données de ChIP-seq de TFs en C2C12 (Caltech). Parallèlement, nous avons obtenu plusieurs données à partir de la littérature. Par exemple, nous avons recueilli les données ChIP-seq de Asp et al. (2011) pour 10 marques épigénétiques sur les histones de cellules C2C12 cultivées en milieu de prolifération (MB pour myoblastes) et en milieu de différenciation (MT pour myotubes). Nous avons aussi recueilli des données ChIP-seq pour divers TFs impliqués dans la différenciation musculaire : le TF maître MyoD en C2C12 MB et MT (Cao et al., 2010), les TFs Pax3 et Pax7 marquant respectivement les progéniteurs embryonnaires et les cellules satellites, ou progéniteurs adultes, obtenus dans des myoblastes primaires (Soleimani et al., 2012), le TF Smad3 associé à la voie TGF-β en MT (Mullen et al., 2011), le TF Msx1 permettant le recrutement du groupe Polycomb

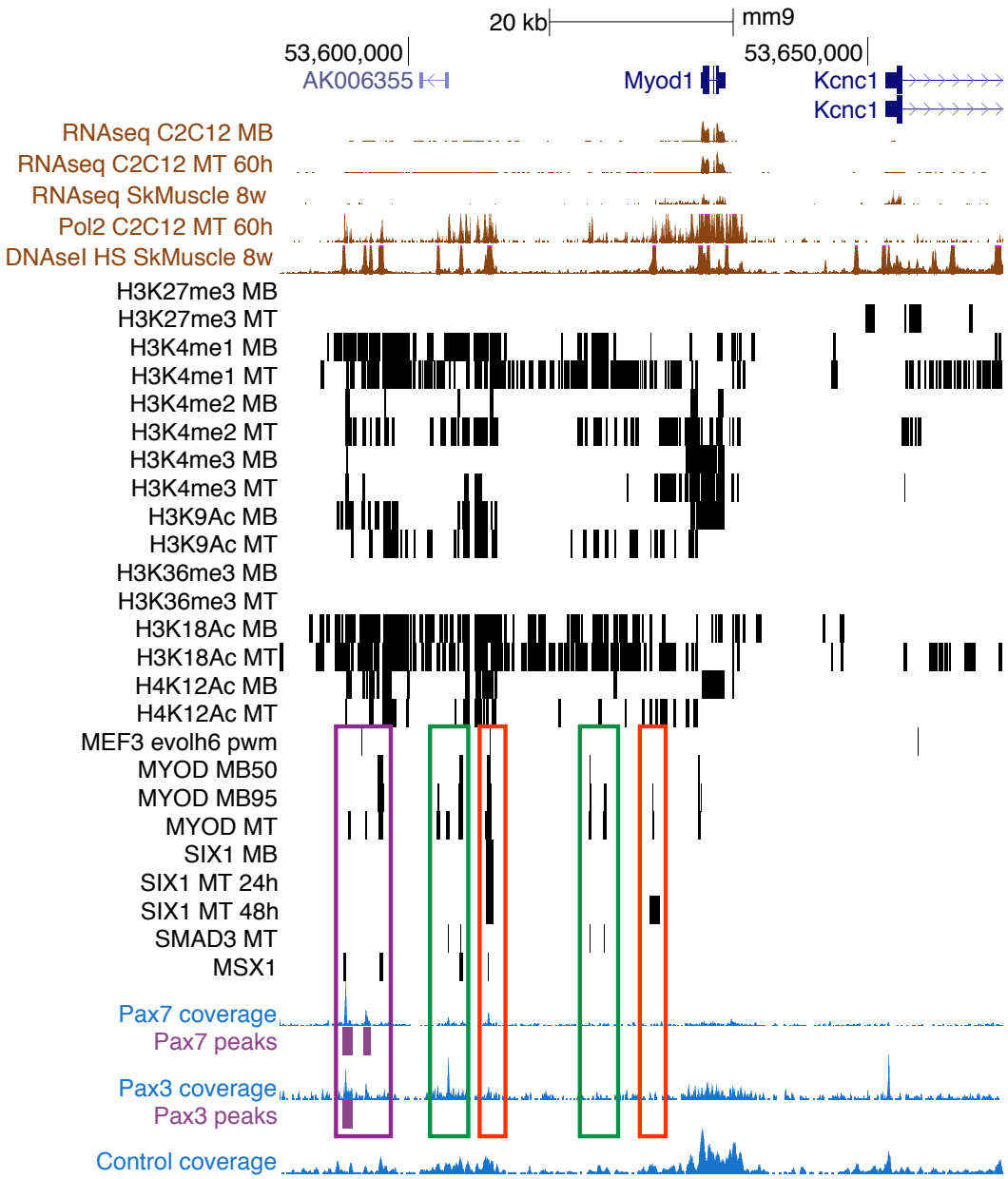


FIGURE 5.6 – **Visualisation de données musculaires sur UCSC.**
Capture d'écran du navigateur de UCSC (UCSC Genome Browser, `http://genome.ucsc.edu`) montrant le locus de *Myod1* et incluant un certain nombre des données que nous avons recueillies. L'intégration des données permet de conduire rapidement plusieurs analyses sur la régulation génétique (voir texte principal). Les cadres colorés indiquent des régions de régulation potentielles fixées par MyoD (cadre pourpre avec Pax3 et Pax7, vert avec Smad3), ainsi que le *Core Enhancer* CE (cadre rouge de gauche) et le *Distal Regulatory Region* DRR (cadre rouge de droite). L'ensemble de ces enhancers constitue un *super enhancer* (voir 1.5.5 et Whyte et al. (2013)).

associé aux marques répressives H3K27me3 en C2C12 MB (Wang et al., 2011), ou encore le TF Sox6 impliqué dans l'inhibition des gènes spécifiques des fibres lentes lors de la différenciation terminale en myoblastes primaires à E18 (An et al., 2011). Par ailleurs, certains TFs n'ont pour le moment été étudiés que par des expériences de ChIP-on-chip à la résolution et l'étendue limitée : le TF Gli associé à la voie de signalisation *Sonic hedgehog* (SHH), dont les données ont été obtenues dans le bourgeon de membre d'embryons à E11.5 (Vokes et al., 2008), ou encore le TF Six1 en C2C12 MB et MT (Liu et al., 2010). Dans ce dernier cas, les données ne couvrent que 17% du génome et possèdent une faible résolution ($\sim$ 1kb) : elles sont donc seulement complémentaires aux prédictions obtenues en utilisant la PWM MEF3 introduite précédemment, car elles confirment la fixation de Six1 sur certains des sites prédits (mais n'infirme pas l'absence de fixation des protéines de la famille Six sur les autres sites).

Nous présentons en figure 5.6 un exemple montrant la visualisation d'une partie des données recueillies dans le cas du locus de *Myod*1, gène associé aux protéines MyoD. Des informations sur la position génomique et les transcrits présents au sein du locus sont d'abord données dans la partie supérieure. Puis les pistes suivantes indiquent diverses expériences, avec la nomenclature suivante : MB pour myoblaste (avec parfois le pourcentage de confluence des cellules en culture), MT pour myotube (avec parfois le temps passé en milieu de différenciation), SkMuscle 8w pour muscle squelettique adulte à 8 semaines post-natales et GM pour *Growth Medium*. Les premières pistes de données (en brun) sont issues du projet ENCODE, les pistes suivantes (en noir) indiquent les coordonnées des pics ChIP-seq des marques épigénétiques des histones ainsi que de divers TFs. Les pistes bleues pour Pax3 et Pax7 indiquent les données brutes du nombre de séquences de ChIP-seq par nucléotide, les segments pourpres indiquant les pics. On montre aussi le contrôle de cette expérience (input) auquel les données sont comparées pour définir les pics. Ainsi, le pic des données brutes de Pax3 dans la partie droite de la piste n'est pas détecté comme un pic authentique car les données contrôle y possèdent aussi un pic.

Prises ensemble, ces données permettent d'appréhender la régulation du gène *Myod*1 au cours de la différenciation musculaire. D'abord, *Myod*1 est transcrit au cours de la différenciation des cellules C2C12 (détection d'ARN par RNAseq, traces de PolII marquant la transcription) mais son expression est plus faible en muscle adulte, soit longtemps après la différenciation. Le gène est dans une région de chromatine ouverte (DNAseI HS), et c'est aussi le cas pour plusieurs régions alentour. Les pics de DNAseI HS sont notamment corrélés aux pics de PolII ainsi qu'à d'autres TFs comme MyoD [18], Six1, Msx1, Pax3 ou encore Pax7, et à des marques epigénétiques « activatrices » comme les méthylation d'histone H3K4me1, H3K4me2 et H3K4me3 ou l'acétylation H3K9Ac, qui sont toutes renforcées au cours de la différenciation (passage de MB à MT). Le fait que l'on observe des pics de PolII hors des régions transcrites est dû à l'interaction entre des régions de régulation contenant plusieurs TFs et le promoteur du gène régulé (ici *Myod1*) fixant la polymérase. Ces données pointent donc vers l'existence de multiples régions de régulation à la chromatine ouverte et interagissant avec le promoteur de *Myod*1. Plus précisément, on trouve de gauche à droite : une région fixant Pax3, Pax7, Msx1 et MyoD en différenciation tardive qui pourrait par exemple être impliquée dans la différenciation des cellules satellites (cadre pourpre), plusieurs régions fixant Smad3 et MyoD pouvant peut-être servir à intégrer des signaux TGF-β (cadres verts), ainsi que deux régions régulatrices connues (cadres rouges), l'une (Kucharczuk et al., 1999) fixant MyoD, Six1 et Msx1 tout au long de la différenciation et possédant un site MEF3 conservé (le CE de MyoD, impliqué

18. Il est effectivement connu que *Myod1* s'autorégule (Thayer et al., 1989)

dans la différenciation des progéniteurs embryonnaires), et l'autre (Asakura et al., 1995) fixant Six1 et MyoD en MT (le DRR, impliqué chez l'embryon ainsi que dans le muscle adulte). Ces enhancers semblent donc avoir différent rôles dans l'activation de *Myod*1 et la mise en place de la myogenèse. On notera pour finir que l'ensemble de ces régions de régulation, à forte densité de ChIP-seq pour le TF maître MyoD et associées à des marques épigénétiques extensives, a été défini ailleurs comme un *super-enhancer* (voir 1.5.5 et Whyte et al. (2013)).

L'intégration des données expérimentales et bioinformatiques liées à la régulation du phénotype musculaire permet ainsi de visualiser rapidement la régulation possible d'un gène et de proposer des hypothèses de mécanismes sous-jacents, qu'il s'agit ensuite de valider expérimentalement. Nous avons pu partager la session UCSC contenant ces différentes données avec nos collaborateurs au sein de l'équipe de P. Maire, qui ont maintenant la possibilité de croiser et interpréter rapidement un grand nombre de données publiées ainsi que de données obtenues au cours de cette thèse.

5.3 Prédictions et validations de la régulation par Six

Plusieurs prédictions réalisées grâce à la PWM MEF3 ont été réalisées et validées expérimentalement, nous en présentons ici deux majeures.

5.3.1 Régulation de *Myod1* par les protéines Six chez l'embryon de souris

Nous avons d'abord réalisé des prédictions de sites MEF3 sur deux enhancers connus de *Myod1*, le *Core Enhancer* CE et le *Distal Regulatory Region* DRR, qui ont été publiées dans l'article Relaix et al. (2013) paru dans *Plos Genetics*. Les prédictions ont été réalisées avec un seuil de 7 bits [19] correspondant au seuil de distinction optimal de sites MEF3 positifs et négatifs introduit en figure 5.4. Nous avons ainsi détecté 2 sites MEF3 sur le CE et 1 sur le DRR que nous montrons en figure 5.7. Sur le CE, les deux sites recoupent en partie des éléments essentiels à l'expression *in vivo* du CE humain dans les cellules musculaires myotomales de la souris (Kucharczuk et al., 1999). Nous montrons par ailleurs d'autres sites détectés dans des travaux précédents (L'honoré et al., 2010; Kucharczuk et al., 1999) : les sites Pitx2 et Pax3 dans le CE, ou encore les E-box dans le CE et le DRR.

La validation de ces prédictions est présentée en figure 5.8. Tout d'abord, il est montré par retard sur gel que les protéines Six1 et Six4 se fixent effectivement sur les séquences prédites (fig. 5.8B). Dans cette expérience, des oligonucléotides marqués par la radioactivité contenant le site MEF3 consensus GAAACCTGA du promoteur de Myog sont incubés avec les protéines Six1 et Six4 synthétisées *in vitro* (colonne 1). On ajoute ensuite des oligonucléotides non marqués en excès d'un facteur 60 ou 300 et contenant le site MEF3 de Myog (contrôle positif, colonnes 2 et 3), les sites MEF3 du DRR (colonnes 4 et 5) et du CE (site 1 : colonnes 6 et 7, site 2 : colonnes 8 et 9), ou le site NFI du promoteur de Myog (contrôle négatif, colonne 10, excès

19. Il faut noter que le seuil utilisé pour afficher les sites MEF3 conservés dans le navigateur de UCSC en figure 5.6 est plus élevé (10.5 bits). En effet, à l'échelle génomique, il y a une très grande quantité de sites négatifs, et il faut donc aller vers les petits taux de Faux Positifs (ou les hauts seuils de détection) pour assurer un filtrage du bruit suffisant : par exemple, on trouve un site non conservé tous les $2^7 \sim 128$ bp pour un seuil de 7 bits, contre $2^{10.5} \sim 1500$ pour un seuil de 10.5 bits. En ajoutant le critère de conservation, ces quantités sont plus petites, et le seuil de 7 bits paraît raisonnable pour des prédictions sur des séquences de petite taille ~ 200 bp.

Mef3, E-box, Pitx, Pax3

Myod1 CE

```
GTCATTGAGA GCTAGGCAGG GGGACACCCT GGAGCACCCC ACAACATGAG
CCCCACAGCA TTTGGGGGCA TTTATGGGTC TTCCTATAAA CTTCTGAGAC
AGTAATTTTA TCCTGCTTCT TTCGGCCAAG TATCCTCCTC CAGCAGCTGG
TCACAAAGCC AGTTAATCTC CCAGAGTGCT CAGCTTAAAA CCCGTGACTC
ACAACACAGC CAGTTGGGGG AAGGGGACAG CCGCCTCCAA ACGTGGCGCC
CAGAGTCAGC TGTTCCTGGG TCTTCTCCGG TTTCTCTAGC TCAGGCCTAG
GGCTGGGGCC TCTTCCTTCC TTCCTGGAGT CC
```

Myod1 DRR

```
TTTCATCCTC CAGTCCTTCA GCCCCCTAGA CCCAAGCCAG CCATGCAGCC
CGCAGTAGCA AAGTAAGAGG CCACAGGTCC AGACTGGGTA GGGCAGAGGT
GCCTGAGGCT TGGGGCAGGT GCTAGTTGGA TCCGGTTTCC AGAGGCTATA
TATATATAAA GGCTGCTGTT TCCCCATGGT GCAACCACCC CAGAGGCCTA
```

FIGURE 5.7 – **Prédiction de sites MEF3 sur les séquences régulatrices CE et DRR de *Myod1*.**
Séquences correspondant aux régions régulatrices CE et DRR de *Myod1*. Trois sites MEF3 ont été prédits à des scores respectifs de 7 bits pour le premier du CE (CE1) et 11.1 bits pour le deuxième du CE (CE2) et celui du DRR. Le seuil de 7 bits correspond au seuil minimal trouvé en fig. 5.4 comme discernant positifs de négatifs. Les sites E-Box, Pitx et Pax3 indiqués correspondent à des sites déjà validés. Les séquences soulignées correspondent aux boîtes LS4 et LS15 de l'expérience de *Linker-Scanner Mutagenesis* du CE humain réalisée par Kucharczuk et al. (1999) et sont requises pour l'expression dans les lignées musculaires.

de 300). Ensuite, il est montré que la mutation des sites MEF3 abolit l'expression d'un rapporteur *in vivo* de l'activité des enhancers par transgenèse. Pour cela, deux constructions ont été réalisées, contenant le CE, le DRR ainsi que le promoteur PRR (*Proximal Regulatory Region*) de *Myod1* en amont du gène rapporteur LacZ, avec ou sans mutations au niveau des 3 sites MEF3 prédits (fig. 5.8C). Ces constructions ont été introduites dans des embryons par transgenèse transitoire, et ceux-ci sont prélevés à E12.5 puis l'expression de LacZ est révélée par coloration au X-Gal (fig. 5.8D). Au total, 6 transgènes *wild-type* sur 10 expriment le rapporteur LacZ [20] avec un motif d'expression stéréotypé (3 d'entre eux sont montrés sur la figure). Dans le cas des transgènes mutés, 3 sur 8 expriment très légèrement LacZ, ils sont tous montrés sur la figure. Enfin, des sections réalisées au niveau de la tête et du thorax montre que l'expression du rapporteur du transgène *wild-type* récapitule le motif d'expression du gène *Myod1* endogène, alors que pour les transgènes mutés l'expression est très rare. Ainsi, les protéines MyoD, Mef2, Pitx, etc., présentes dans les cellules myogéniques sont incapables d'activer les régions régulatrices mutées dans les sites MEF3, suggérant qu'elles sont incapables de se fixer à l'ADN.

20. Cette variation d'expression est notamment due au fait qu'il est difficile de contrôler le nombre de transgènes insérés dans le génome. Les différents embryons montrés en figure 5.8D,E sont représentatifs des différents taux d'insertions obtenus.

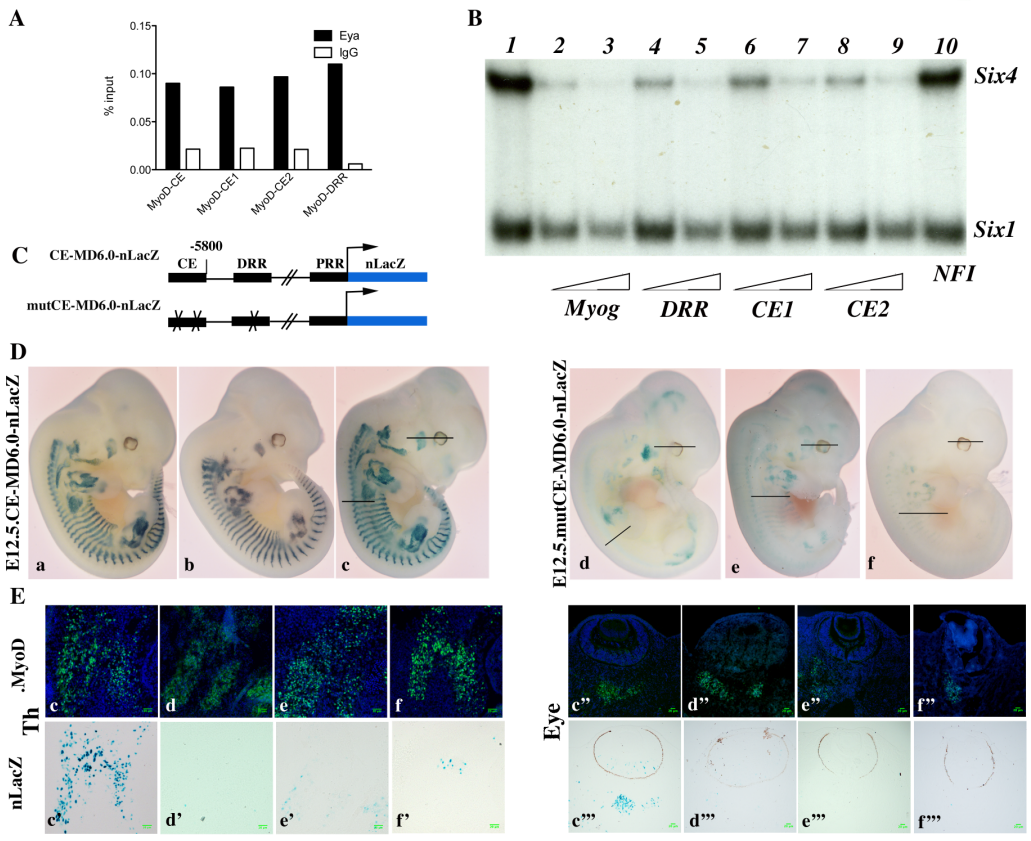


FIGURE 5.8 – **Validation des sites MEF3 prédits dans les éléments régulateurs de *Myod1*.** Figure présentée dans l'article Relaix et al. (2013). Plus de détails sont donnés dans le texte principal. (A) Expérience de ChIP montrant que les protéines Eya, cofacteurs de Six1 et Six4, sont recrutées au niveau du CE et du DRR. (B) Expériences de retard sur gel (EMSA) montrant que Six1 et Six4 interagissent avec les sites prédits sur les enhancers. (C) Schéma des transgènes wild-type et mutés en sites MEF3. (D) Coloration au X-Gal d'embryons à E12.5 porteurs de transgènes des séquences wild-type (a, b, c) et mutantes (d, e, f), montrant l'abolition de l'expression du gène rapporteur lorsque les sites MEF3 sont mutés. (E) Analyse de l'expression de MyoD par immunohistochimie et du rapporteur LacZ par coloration X-Gal dans les coupes indiquées par un trait noir chez les embryons c, d, e et f au niveau du thorax (Th) et des yeux (Eye). On note l'absence d'expression du transgène muté dans les cellules exprimant MyoD.

5.3.2 Régulation d'un lincRNA par Six dans le muscle adulte

Nous présentons maintenant un travail réalisé en collaboration avec Iori Sakakibara de l'équipe de P. Maire.

- **Contexte général**

Les protéines Six possèdent un rôle important dans la spécialisation des fibres musculaires de type rapide (ou glycolytiques) lors de la différenciation tardive (Niro et al., 2010). De plus, dans le muscle adulte, Six1 est exprimé de manière plus importante dans les noyaux des fibres adultes que dans ceux des fibres lentes. Le phénotype rapide est marqué par l'expression de

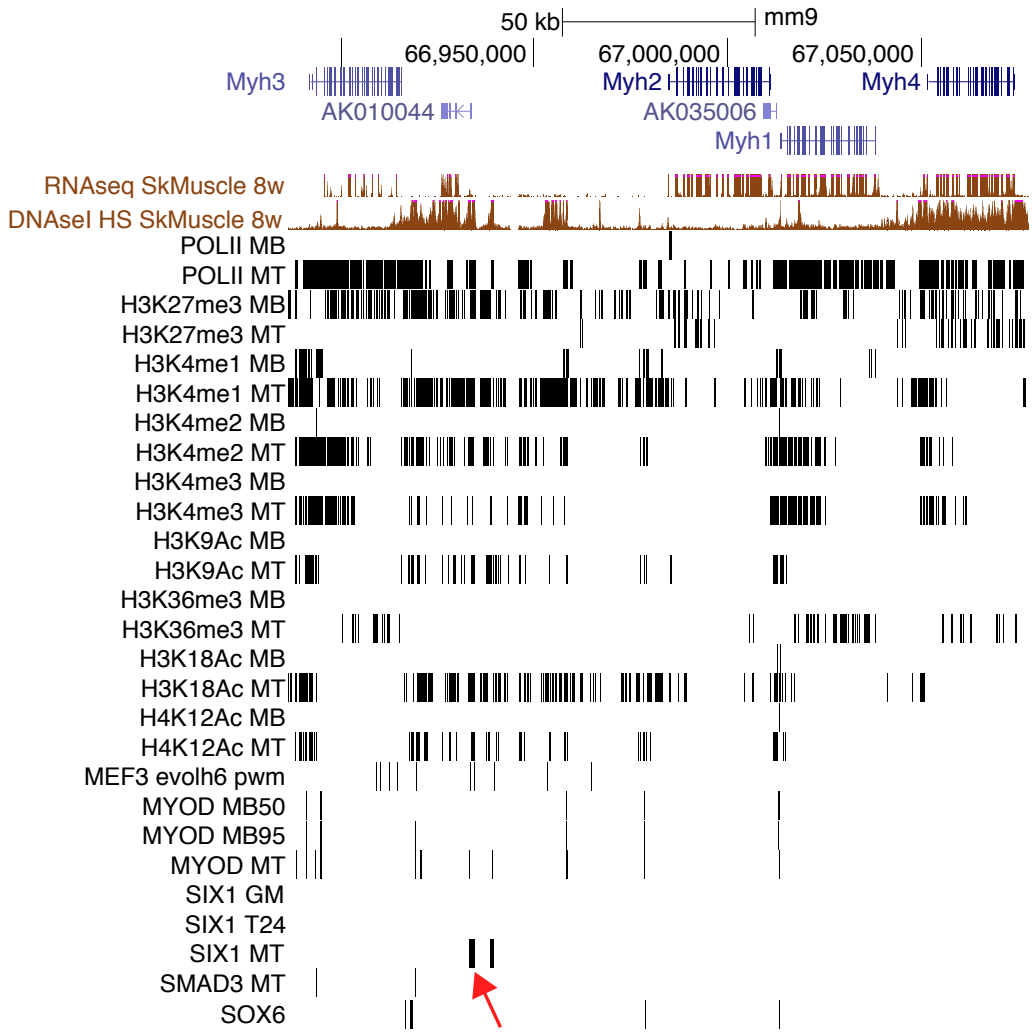


FIGURE 5.9 – **Visualisation du locus des myosines sur UCSC.**
Le locus des myosines s'étend sur 300 kb sur le chromosome 11 de la souris. Au sein de ce locus ce trouve un long ARN non codant (linc-RNA), AK010044 ou linc-MYH. Ce lincRNA est transcrit dans le muscle squelettique (RNAseq), où il se trouve au sein d'un région de chromatine ouverte (DNAseI HS). La région est marquée par une forte régulation épigénétique lors de la différenciation des C2C12 : apparition de marques d'activation (H3K4me et diverses acétylations) et disparition de marque inhibitrices (H3K27me3), ces changements étant corrélés avec des traces étendues sur le locus de PolII. Linc-MYH est entouré de régions de régulation putatives contenant les TFs MyoD, Six1, Smad3 et Sox6. Ici, nous nous sommes concentrés sur la région directement en amont du TSS de lincMYH (indiquée par une flèche rouge), contenant des traces ChIP-seq pour MyoD et Six1 en MT, ainsi que plusieurs sites MEF3 conservés à haut score.

gènes « rapides », dont les troponines *Tnnt*3, *Tnni*2 et *Tnnc*2 ou les myosines *Myh*2, *Myh*1 et *Myh*4. La question concernant le rôle exact des protéines Six lors de cette spécialisation est encore ouverte .

Nous avons concentré notre étude sur le « cluster myosines » s'étendant sur 300 kb dans le chromosome 11 de la souris, et contenant les myosines *Myh*3 (embryonnaire), *Myh*2 (rapide de type 2A), *Myh*1 (rapide de type 2X), *Myh*4 (rapide de type 2B), *Myh*8 (périnatale) et *Myh*13 (extra-oculaire). Nous montrons en fig. 5.9 une partie de ce locus. Nous nous sommes demandés si les protéines Six pouvaient réguler directement les myosines de type rapide au niveau de ce locus. La concentration des gènes de myosine en un même locus est réminiscente du cluster des gènes de la β-globine. Dans ce cluster, les différents gènes de la β-globine sont régulés par une même région régulatrice, appelée « Locus Control Region » ou LCR (Palstra et al., 2008). Nous avons donc cherché à savoir si un tel LCR pouvait exister dans le cas du cluster myosine.

Nous avons ainsi centré notre attention sur une région de chromatine ouverte fixée par Six1 et MyoD en cellules C2C12 et contenant plusieurs sites MEF3 (indiquée par une flèche rouge dans la figure 5.9). Cette région est située à proximité des myosines de type rapide, et se trouve en amont du TSS d'un long ARN non codant ou lincRNA (pour *long intergenic non-coding RNA*) que nous avons nommé linc-MYH. Les linc-RNAs ont reçu beaucoup d'attention ces dernières années du fait des rôles variés qu'on leur a découverts, notamment au niveau de la régulation de la chromatine (Rinn and Chang, 2012). Nous nous sommes donc posés deux questions : est-ce que cette région de régulation peut servir de LCR aux myosines de type rapide ? Et peut-elle servir à activer linc-MYH, qui pourrait alors avoir un rôle dans le processus de spécialisation ?

- **Article**

Dans l'article qui suit, il est d'abord montré que cette région régulatrice fixée par Six1 sert de LCR aux myosines rapides. Ainsi, elle interagit *in vivo* avec les promoteurs des myosines rapides. Elle active par ailleurs des gènes rapporteurs sous le contrôle des promoteurs de *Myh*2, *Myh*1, *Myh*4, et cette activité est abolie lors de la mutation des sites MEF3. Par ailleurs, il est montré que ce LCR est capable d'activer linc-MYH. Le rôle de linc-MYH est étudié par *knock-down* dans un muscle adulte de type rapide par électroporation d'un shRNA. Ce *knock-down* résulte en une augmentation significative de l'expression des gènes associés au phénotype lent, comme *Tnnt*1, *Tnni*1 et *Tnnc*1. Ainsi, en se fixant sur le LCR, Six1 jour un double rôle : localement, il permet l'activation les myosines de type rapide, et à l'échelle du génome, il bloque le programme d'expression de gènes lents *via* l'activation de linc-MYH, permettant *in fine* la mise en place du programme rapide.

Six Homeoproteins and a linc-RNA at the Fast MYH Locus Lock Fast Myofiber Terminal Phenotype

Iori Sakakibara[1, 2, 3], Marc Santolini[4], Arnaud Ferry[2, 5], Vincent Hakim[4], Pascal Maire[1, 2, 3, *]

[1]INSERM U1016, Institut Cochin, Paris, 75014, France.

[2]CNRS UMR 8104, Paris, 75014, France.

[3]Université Paris Descartes, Sorbonne Paris Cité, Paris, 75014, France.

[4]Laboratoire de Physique Statistique, CNRS, Université P. et M. Curie, Université D. Diderot, École Normale Supérieure, Paris, 75005, France.

[5]Université Pierre et Marie Curie-Paris6, Sorbonne Universités, UMR S794, INSERM U974, CNRS UMR7215, Institut de Myologie, Paris, 75013, France.

[*]Correspondence author: pascal.maire@inserm.fr.

Tel: 33 (0) 1 44 41 24 13/16

Running title: *linc-MYH* locks fast muscle phenotype.

Keywords: long non coding RNA, Six1, muscle fiber type, myosin heavy chain, linc-MYH, LCR

Abstract

Thousands of long intergenic non-coding RNAs (lincRNAs) are encoded by the mammalian genome. However, the function of most of these lincRNAs has not been identified *in vivo*. Here, we demonstrate a role for a novel lincRNA, *linc-MYH*, in adult fast-type myofiber specialization. Fast myosin heavy chain (*MYH*) genes and *linc-MYH* share a common enhancer, located in the fast *MYH* gene locus and regulated by Six1 homeoproteins. *linc-MYH* in nuclei of fast-type myofibers prevents slow-type and enhances fast-type gene expression. Functional fast-sarcomeric unit formation is achieved by the coordinate expression of fast *MYHs* and *linc-MYH,* under the control of a common Six-bound enhancer.

Introduction

Adult skeletal muscles are composed of slow and fast myofiber subtypes which selectively express the genes required for their specific contraction activity and metabolic properties [1–4]. These properties are acquired at the end of fetal development and during the neonatal period, when mixed skeletal myofibers expressing a panel of embryonic, fast and slow genes develop a specific slow or fast phenotype. The formation of efficient fast sarcomeric units, and Ca^{++} cycling and excitation/contraction/relaxation coupling in fast-myofibers, is achieved through the coordinate control of fast *Myhs* and associated fast sarcomeric genes (including *Tnnt3, Tnni2*, *Tnnc2, Atp2a1* and *Pvalb*) (Schiaffino et Regiani, 2011; Greising et al, 2012). Myofibers can be classified by their MYH expression profile: slow-type myofibers in mice express *MYH7* (also known as *MYHCI*, *β* or *slow*), and fast-myofibers express *MYH2* (*MYHCIIA*), *MYH1* (*MYHCIIX*) or *MYH4* (*MYHCIIB*). Fast *Myh* genes found in developmental and adult stages *(Myh3, Myh2, Myh1, Myh4, Myh8* and *Myh13*) are organized as a cluster within a 300kb region on mouse chromosome 11 [5]. The spatio-temporal expression of one specific fast *Myh* gene at the *Myh* locus is reminiscent of the

organization and expression of Globin genes at the beta-globin locus [6]. However, we are yet to investigate potential enhancers or the *Myh* locus control region (LCR) that could be responsible for sequential and specific *Myh* gene expression in myofibers. The coordination of fast-type and slow-type gene expression in fast myofibers is not currently understood. Distinct intramyofibrillary calcium transients, evoked by slow tonic motor neuron firing, induce a cascade of downstream signaling involving Calcineurin and CamK. This results in the activation of selective transcription activators and repressors in slow myofibres. However, the signaling pathways operating in distinct *MYH2*, *MYH1* and *MYH4* myofiber subtypes, which coordinate the activation of the other fast-type genes and the repression of slow-type genes, is less well understood [1]. Better knowledge of the mechanisms controlling muscle specialization and plasticity is important to enable the understanding and modulation of muscle adaptations in pathophysiological conditions.

Six homeoproteins are major myogenic transcription factors which directly bind to DNA sequences (called MEF3s) to control myogenesis [7,8] and the genesis of fast-type myofibers during embryogenesis [9,10]. In adult skeletal muscle, Six1 accumulates at a higher level in the nuclei of adult fast myofibers than in of slow myofibers. Forced expression of Six1 and its Eya1 cofactor in slow myofibers causes adult slow-twitch oxidative fibers toward a fast-twitch glycolytic phenotype [11]. Animals with a *Six1* KO present severe muscle hypoplasia and die at birth [12]. This prevents the *in vivo* analysis of the adult phenotype and the ability to investigate the direct or indirect involvement of Six1 in the spatio-temporal control of the expression of genes in the fast *Myh* cluster.

The mammalian genome encodes thousands of long intergenic non-coding RNAs (lincRNAs) which have multiple functions [13,14]. Some accumulate in the cytoplam as miRNAs decoys [15,16]. Others accumulate in the nucleus and participate to gene regulation through chromatin remodeling and epigenetic modifications [14,17,18]. Here, they may act as cis [19]

or trans [20] transcriptional activators, as transcriptional repressors [21,22] or through DNA-RNA triplex formation [23,24].

In this study we identify a new lincRNA, *linc-MYH*, and the mechanism of its control of adult muscle fast fiber-type specification *in vivo*. We demonstrate a three-element genetic partnership, where a LCR under the control of the myogenic homeoprotein Six1 functions as a regulatory hub to control fiber phenotype. In this partnership, the LCR positively controls the expression of both the adjacent fast *Myh* gene cluster and *linc-MYH*, suppressing slow-type gene expression and facilitating fast fiber-type specialization.

Results and Discussion

Six1 binds directly to the enhancer/LCR of the *Myh* genes cluster.

We suggest that Six1 could be directly involved in the control of the expression of fast *Myh* genes as higher levels of this transcription factor accumulate in the nuclei of adult fast myofibers than in slow myofibers. We used computational analysis to locate MEF3 sites (at the fast *Myh* locus) to investigate how *Six1* could control the expression of fast *Myh* isoforms. Six clustered MEF3 sites are conserved across human, rat and mouse genomes in an intergenic region located 50 kb upstream of the *Myh2* gene (Figures 1A and S1) and 4kb upstream of a lincRNA (2310065F04Rik); we refer to this as *linc-MYH* (Figures 1A and S2). Six1 binding at these MEF3 sites was demonstrated *in vivo* by ChIP experiments with Six1 antibodies on adult fast gastrocnemius plantaris (GP) and tibialis anterior (TA) muscles (Figure 1*B*), and confirmed for five of these sites by EMSA assays (Figure S3). We asked whether this *Myh* intergenic region could constitute an enhancer element, controlling the spatio-temporal expression of *Myh* genes in this locus. A 2kb DNA fragment of this region, including the six identified MEF3 sites and 1kb of DNA fragments upstream of the transcription start site of fast-type *Myh2*, *Myh1* and *Myh4* genes, was isolated. The putative

enhancer was ligated to each *Myh* promoter using luciferase pGL3 basic plasmids to generate pGL3-Enhancer-*Myh2/1/4* constructs. We mutated all six MEF3 sites present in the enhancer, and named these reporters pGL3-mutEnhancer-*Myh2/1/4,* to test the involvement of Six binding in enhancer activation of the *Myh2*, *Myh1* and *Myh4* promoters. Luciferase activity was tested after the electroporation of reporter plasmids in adult TA muscles. The luciferase activity of pGL3-Enhancer-*Myh2/1/4* was between seven and twelve times higher with either of the promoters, than with pGL3-*Myh2/1/4*. Enhancer activity was not observed in plasmids with MEF3 mutated sites associated with either of the *Myh* promoters (Figure 1C). To determine *in vivo* interactions between the enhancer and each *Myh* gene, we performed chromatin conformation capture (3C) assays of adult EDL muscles. These experiments revealed that the enhancer interacts with the promoter of *Myh2/1/4* genes in native chromatin of EDL myonuclei (Figure 1D). The strongest interactions were observed with the *Myh1* and *Myh4* promoters, consistent with the expression profile of these two genes in EDL muscles. The data demonstrates that the identified conserved cis-element acts as an enhancer for the *Myh* locus, and that MEF3 sites are essential for its enhancer activity *in vivo.*

Loss of *Six1* impairs fast muscle genes and *linc-MYH* expression during post-natal development.

To further characterize the role of *Six1* in the control of fast *Myh* gene expression, we bred $Six1^{flox/flox}$ mice with transgenic mice expressing CRE recombinase under the control of the human skeletal actin (HSA) promoter and obtained $Six1^{flox/flox}$*;HSA-CRE* conditional knockout mice (hereafter named *cSix1 KO*) [25,26]. We analyzed the expression of fiber type specific genes in the back muscles of wild-type control mice and *cSix1 KO* mice at embryonic day 18.5 (E18.5) and at several post-natal stages (two weeks (P2W), four weeks (P4W) and eight weeks (P8W)) animals (Figure 2), as muscle fiber fast-subtype specialization can be studied from the end of embryogenesis [9]. *Six1* mRNA was not detectable in back muscles of

cSix1 KO mice (Figure 2). The expression of fast-type genes (*Myh4, Tnnt3*, *Tnni2*, *Tnnc2* and *Pvalb)* increased during postnatal development in control mice but that of slow-type genes (*Myh7, Tnnt1*, *Tnni1*, *Tnnc1* and *Sln*) decreased. The *linc-MYH* RNA was detected after birth in muscle samples and its expression increasing in line with that of *Myh4* (Figure 2). The induction of fast-type genes and *linc-MYH,* and the suppression of slow-type genes, were impaired in *cSix1 KO* mice. Expression of *linc-MYH* was reduced by between three and five times in *cSix1 KO* mice during postnatal development (Figure 2). These data show that *Six1* induces *linc-MYH* and fast-type genes during postnatal development.

***Six1* deficiency impairs adult muscle fast phenotype.**

We analyzed 12 week-old *cSix1 KO* mice to further characterize the role of *Six1* in adult muscle. *Six1* mRNA and protein were not detectable in GP enriched with fast-myofibers or soleus (SOL) muscle enriched with slow-myofibers (Figure 3A and B), and fatigue resistance of TA muscle was 35% higher (Figure 3C) in the *cSix1 KO* mice. We used immunohistochemistry to analyse the composition of MYH7, MYH2 and MYH4 in *cSix1* mutant myofibers. TA muscles had a higher percentage of fibers containing MYH7 and MYH2-, but a lower percentage of fibers containing MYH4 (Figures 3D and S4). We found consistent results during qPCR analysis of *Myh* mRNA: high levels of *Myh*7 and *Myh*2 mRNA, and lower levels of *Myh*4 mRNA levels, were observed (Figure 3E) in the fast TA muscles of *cSix1 KO*. Expression levels of other specific fast and slow-type genes were also tested. We found downregulation of fast-type genes (*Tnnt3*, *Tnni2*, *Tnnc2* and *Pvalb*) and between a five and to 25 times increase in the levels of slow-type genes (*Tnnt1*, *Tnni1*, *Tnnc1* and *Sln*) (Figure 3E). The expression of *linc-MYH* expression was lower in the adult TA of *cSix1 KO* mice, than in control mice (Figure 3E). Our results indicate that the *Six1* homeoprotein can control the phenotype of fast skeletal myofibers in adult animals.

***linc-MYH* is exclusively expressed in adult fast-type muscles.**

We found that *linc-MYH* is expressed in fast-type skeletal muscles (GP, TA and EDL), but not in SOL, brain, kidney, heart or fat tissues, an expression pattern which parallels that of the fast-fiber *Myh4* (Figure 4A). This suggested that *linc-MYH* is only expressed following robust nuclear accumulation of Six1, like in the nuclei of *MYH4* myofibers [11], and that the less robust nuclear accumulation of Six1 observed in SOL myonuclei did not induce *linc-MYH* expression. We used luciferase reporter transfection assays (as described previously) to test the requirement for Six binding on the *MYH* enhancer to activate *linc-MYH* expression. These transient transfection assays, performed in adult TA, show that the *MYH* enhancer activates *linc-MYH* expression in a Six-dependent manner (Figure 4B). *lincRNAs* can localize in cytoplasm [16] or as a single focus [19] or multiple foci [20] in nuclei. We performed fluorescent *in situ* hybridization (FISH), using *linc-MYH* antisense RNA and isolated myofibers from fast EDL, to analyze *linc-MYH* localization in skeletal muscle fiber. Intranuclear localization of *linc-MYH* was observed, with approximately 10 *linc-MYH* foci per nucleus (Figure 4C).

***linc-MYH* coordinates fiber-type gene expression.**

Following these observations, we hypothesized that *linc-MYH* could act in trans [17] to control gene expression in fast myofibers. To test this theory, we used electroporation to introduce a shRNA against *linc-MYH* (sh*linc-MYH*) in TA muscle and analyzed the transfected samples after fourteen days. This method yielded the efficient knockdown of *linc-MYH*, with a 90% reduction of its expression (Figure 5A). RNA samples from sh*linc-MYH* transfected adult TA were analyzed by Affymetrix microarrays (Figure 5B and Table S1), and validated by qPCR experiments, to identify the consequences of *linc-MYH* knockdown and understand its mode of action. The expression of *linc-MYH* was significantly lower in the absence of *Six1* but, *Six1* expression was not affected by the absence of *linc-MYH.* Knockdown of *linc-MYH* produced a phenotype with robust gene expression phenotype; this

downregulated the expression of several fast genes, (including *Tnnt3*, *Tnni2* and *Pvalb*), and upregulated the expression of numerous slow genes (such as *Sln*, *Tnni1*, *Tnnc1* and *Tnnt1*) (Figure 5B). Contrary to what was observed in the muscles of *cSix1* KO mice, slow *Myh*7 expression level did not change. Following *linc-MYH* knockdown, expression levels of the neighboring genes *Myh2* and *Myh1* did not change but the expression of the more distant *Myh4* gene was downregulated; this suggests a specific requirement of *linc-MYH* for *MYH4* expression. We found a strong qualitative and quantitative correlation in the expression of specific genes between *linc-MYH* knockdown and *cSix1* mutant myofibers. The expression of slow muscle genes was between 3 and 10 times greater in *linc-MYH* knockdown samples, and between 5 and 25 times greater in c*Six1*KO samples, than in the wildtype. This suggested that *linc-MYH* lies downstream of *Six1* in the Six myogenic pathway and helps to repress slow muscle genes in fast myofibers. The downregulation of all fast-type genes (other than *Myh*4), and the upregulation of slow-type genes, was weaker in the *linc-MYH* knockdown than in the *Six1*cKO line. *Six1* may control several inhibitory pathways, including the *linc-MYH* pathway, to prevent slow-type genes expression in adult fast myofibers. During fetal development, at a stage where *linc-MYH* expression is not yet activated, *Six1/4* increases the nuclear accumulation of the slow muscle repressors Sox6 and HDAC4 to repress slow muscle gene expression [9,27,28]. In accordance with this, the expression of the slow genes *Myh7, Sln*, *Tnni1* and *Tnnt1* is upregulated in the muscle-specific *cSox6* mutant [29]. This demonstrates that *linc-MYH* and Sox6, both lying downstream of Six1, directly participate in the downregulation of *Sln*, *Tnni1* and *Tnnt1* in fast myofibers. However, the repression of slow *Myh7* in fast myofibers has a Six1-Sox6 dependent, but *linc-MYH* independent, repression mechanism. In this study, we observed that levels of fast muscle gene expression decreased by between a factor of 2 and 3 in the *linc-MYH* knockdown, with the highest decrease found for *Myh*4 expression. Their expression decreased by a factor of 1.3 to 2.5 in c*Six1*KO, with

the highest decrease in levels of *Pvalb* expression (Figures 4E and 5A). The presence of Six4 and Six5 proteins in adult myofibers [11], which have the same DNA binding specificity as Six1, could compensate its absence in c*Six1* KO animals and enable the activation of downstream fast muscle targets. In this case, *linc-MYH* expression could be preferentially dependent upon Six1, rather than on Six4 or Six5.

Transcriptomic analysis of *cSix1* and *linc-MYH* knock down.

We compared the networks of genes under the control of *linc-MYH* and of Six1 homeoprotein in adult muscles by the transcriptomic analysis of c*Six1* and *linc-MYH* knockdown lines (Figure 5A and Table S2). We found that the six genes whose expression was the most increased in the *linc-MYH* knockdown were also significantly upregulated in *cSix1 KO* muscles (Figure S5). Two genes, *Ankrd1* and *Peg10*, were more severely upregulated in the *linc-MYH* knockdown line (10 and 8 times, respectively) than in c*SIX1* mutant myofibers (by 2.8 and 1.5 times, respectively). These non slow-type genes may be exclusively repressed by *linc-MYH* in adult fast myofibers as there was stronger downregulation of *linc-MYH* accumulation after its knockdown than in c*Six1* mutant myofibers. The accumulation of *linc-MYH* transcripts in the nuclei of fast myofibers seem to facilitate the regulation of a network of genes that drive myofiber specialization via the same pathway as *Six1* and downstream of this transcription factor.

Conclusion. We identified a novel mechanism for the specialization of the fast-myofiber subtype. The long intergenic non-coding RNA *linc-MYH* and fast *MYH* genes, both of which are essential for myofiber specification, share a common enhancer which is regulated by Six1 homeoproteins. The RNA *linc-MYH* specifically accumulates in nuclei of adult fast myofibers. Its function, as revealed by *in vivo* knockdown and transcriptome-wide analysis, is to prevent slow-type muscle gene expression and increase fast-type muscle gene expression in fast-type myofibers. *linc-MYH* was found to downregulate genes associated

with slow muscle contractile properties like the slow genes *Tnn* and *Sln* (a known repressor of Serca1/Atp2a1 protein [30,31] involved in Ca^{++} reuptake by the sarcoplasmic reticulum). These genes, which belong to the muscle contractile machinery and are repressed in adult fast myofiber, are positively controlled by Six1 in myogenic C2 cells [32]. This suggests that their expression in adult fast myofiber may be restricted by an additional level of regulation involving the *Six1-linc-MYH* axis. As a result of these observations we suggest that Six1 controls the acquisition of fast-type myofiber mechanical properties by binding to a single enhancer region of the fast *Myh* locus. It promotes the coordinated expression of fast *Myhs* and that of a strong repressor of genes controlling slow contractile properties. The modulation of Six activity (depending on fiber-type) facilitates changes in the expression levels of the fast genes *Myh* and *Tnn*; these are required for the formation of efficient sarcomeric units and the appropriate Ca^{++} cycling and excitation/contraction/relaxation coupling. The enhancer element therefore connects distinct regulatory hubs to achieve ultimate muscle fiber specialization. *linc-MYH* functions as an end-of-the-chain control element, conveying the state of fast *Myh* LCR activity to repress slow-type specific genes and coordinate a finer level of regulation.

Materials and Methods

ChIP experiments. GP and TA muscles of 2 months old mice were minced with scissors just after sampling and fixed in 1% formaldehyde for 10 minutes. Formaldehyde was quenched by addition of 0.125 M glycine, and muscles were washed twice in PBS. The muscles were incubated on ice in lysis buffer (10 mM Tris-HCl pH 7.9, 85 mM KCl, 0.5% NP40, protease inhibitors (cOmplete, Roche)) for 10 minutes and homogenized with a mortar and, subsequently with a Dounce homogenizer. The nuclei were obtained by centrifugation, incubated in SDS lysis buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS, protease inhibitors) for 10 minutes, and sonicated in a Bioruptor apparatus (Diagenode). The debris

were removed by centrifugation. The sonicated DNA was incubated with 1 μg of *Six1* antibodies (HPA001893, Sigma) under agitation at +4 °C overnight. 20 μl of Dynabeads protein G (Invitrogen) were added to the samples and incubated under rotation at +4 °C for 1 hour. The beads were washed with low-salt buffer (2 mM EDTA, 20mM Tris-HCl pH 8, 150 mM NaCl, 1% TritonX-100, 0.1% SDS), high salt buffer (2 mM EDTA, 20mM Tris-HCl pH 8, 0.5M NaCl, 1% TritonX-100, 0.1% SDS), LiCl buffer (1 mM EDTA, 10mM Tris-HCl pH 8, 0.25M LiCl, 1% NP40, 1% deoxycholate) and TE buffer (1 mM EDTA, 10mM Tris-HCl pH 8). The DNA was eluted with elution buffer (1% SDS, 0.1 M $NaHCO_3$) containing 0.1mg/ml proteinase K (Invitrogen) at 62 °C for 2 hours, and, proteinase K was inactivated by incubation at 95°C for 10 minutes. The DNA was finally purified with MinElute PCR purification kit (Qiagen). The amount of specific amplified DNA is normalized by *beta-Actin* promoter amplification. The sequences of the oligonucleotides used in this study are as follows. Enh LCR 1F, ATCTCCACCTCCCTCCAACT; Enh LCR 1R, ACCCCCTAGCTTTGACAGGT; Enh LCR 2F, AATCTGACGACAGGGTGAGC; Enh LCR 2R, GGTCGCCTGACCTGATAGAG; AldoaF, CTCTCAAGGCAAACCAAAGC; AldoaF, CCAGTGTCCCAGACCTTCTC; ActbF, TGTTACCAACTGGGACGACA; ActbR, ACCTGGGTCATCTTTTCACG.

3C. 3C experiments were performed on adult EDL muscles as described [33]. Single myofibers were obtained from adult EDL muscles as previously described [26] and fixed. The sequences of the oligonucleotides used in this study are given in Table S3.

Mice. Animals were bred and handled as recommended by European Community guidelines. Experiments were performed in accordance with the guidelines of the French Veterinary Department. *cSix1*KO mice were obtained by breeding the *Six1-LoxP* mice [26] and

transgenic mice expressing a CRE recombinase under the control of the human skeletal actin promoter (HSA) [25].

RNA preparation. TA, back, soleus and GP muscles were collected from *cSix1 KO* and control mice. Total RNAs were extracted by Trizol Reagent (Invitrogen) according to manufacturer's instruction.

cDNA synthesis and QPCR. RNAs were treated with DNase I (Turbo DNA-free, Invitrogen) and were reverse transcribed with Superscript III kit (Invitrogen) according to manufacture's instruction. Reverse transcription was performed with 1 μg of total RNA. Quantitative real time PCR (Light Cycler 480, Roche) was performed using Light Cycler 480 SYBR Green I Master Kit (Roche) according to the manufacturer's protocols. PCR was performed for 40 cycles of 95 °C for 15 seconds, 60 °C for 15 seconds, and 72 °C for 15 seconds. Genes expression level was normalized by the expression level of the housekeeping gene *Actb*. The sequences of the oligonucleotides used in this study are given in Table S4.

Muscle contraction test. Skeletal muscle function was evaluated by measuring *in situ* muscle contraction, as described previously [34]. 12 week-old male mice were anesthetized (intraperitoneal injection of pentobarbital sodium, 50 mg/kg). Body temperature was maintained at 37°C using radiant heat. The distal tendon of the TA muscle was attached to an isometric transducer (Harvard Bioscience) using a silk ligature. The sciatic nerves were proximally crushed and distally stimulated by a bipolar silver electrode using supramaximal square wave pulses of 0.1 ms duration. Responses to tetanic stimulation (pulse frequency 50–143 Hz) were successively recorded. Maximal forces were determined at optimal length (length at which maximal force was obtained during the tetanus). Fatigue resistance was then determined after a 5-minutes rest period. The muscle was continuously stimulated at 50 Hz

for 2 minutes (sub-maximal continuous tetanus), and the duration corresponding to a 20% decrease in force was recorded.

RNA-FISH. Fluorescent-labeled antisense *linc-MYH* probes were synthesized according to manufacturer's instruction (FISH Tag RNA kit, Invitrogen). FISH experiments were performed on isolated EDL myofibres and images acquired on a Leica SP2 confocal microscope.

Generation of shRNA against mouse *linc-MYH*. Five distinct shRNAs targeting mouse *linc-MYH* were designed, called sh*lincMYH*, and inserted into the psiSTRIKE hMGFP system (Promega). The efficiency of each shRNA was established by determination of *linc-MYH* transcript levels in TA muscles transfected by each sh*lincMYH*. The shRNA against 5'-TTCTGCTCACCACCTACAATT-3' sequence was selected for the knockdown experiment.

Electroporation. Ten μg of shRNA-expressing vector were introduced into TA muscles of 8 week-old mice by electroporation as previously described [11]. Two weeks following electroporation, TA myofibers expressing GFP were dissected under a Nikon SMZ1500 stereo microscope and frozen in liquid nitrogen before processing.

Immunohistochemistry. TA, soleus and gastrocnemius muscles were embedded in cryomatrix and quickly frozen in isopentane cooled with liquid nitrogen. Cryostat sections (10 mm) were fixed in 4% PFA, washed in PBS, permeabilized with 0.1% Triton X-100 and left for 1 hour in blocking solution (1x PBS, 1.5% goat serum, 0.1% Triton X-100). Rabbit poly-clonal antibodies directed against Laminin (Z0097, Dako) (1/100 dilution), and monoclonal antibodies against *MYH*7 (NOQ7.5.4D, Sigma) (1/1000 dilution), *MYH*2 (SC-71, Santa Cruz biotechnology) (1/20 dilution) and against *MYH*4 (BF-F3, Developmental Studies Hybridoma

Bank) (1/20 dilution) were applied overnight at 4 °C to the treated sections. The next day, after three washes with 1× PBS containing 0.05% Tween-20, cryosections were incubated for 1 h with appropriate fluorescent secondary antibodies (Alexa Fluor 488 goat anti-rabbit IgG 1/1000 dilution, Alexa Fluor 594 goat anti-mouse IgG 1/1000 dilution, Invitrogen). After three washes with 1× PBS containing 0.05% Tween 20, samples were mounted in Vectashield mounting medium.

Microarray. After validation of RNA quality with the Bioanalyzer 2100 (using Agilent RNA6000 nano chip kit), 50 ng of total RNA were reverse transcribed following the Ovation PicoSL WTA System (Nugen). Briefly, the resulting double-strand cDNA was used for amplification based on SPIA technology. After purification according to Nugen protocol, 5 mg of single strand DNA was used for generation of Sens Target DNA using Ovation Exon Module kit (Nugen). 2.5 mg of Sens Target DNA were fragmented and labelled with biotin using Encore Biotin Module kit (Nugen). After control of fragmentation using Bioanalyzer 2100, the cDNA was then hybridized to GeneChip® Mouse Gene 1.0 ST (Affymetrix) at 45°C for 17 hours. After overnight hybridization, the ChIPs were washed using the fluidic station FS450 following specific protocols (Affymetrix) and scanned using the GCS3000 7G. The scanned images were then analyzed with Expression Console software (Affymetrix) to obtain raw data (cel files) and metrics for Quality Controls. The analysis of some of these metrics and the study of the distribution of raw data show no outlier experiment. RMA normalization was performed using R and normalized data was subjected to statistical tests.

EMSA. EMSA was carried out with *Six1* and Six4 full-length mouse cDNA cloned into the pCR3 vector (Clontech) as previously described [35]. Recombinant mouse *Six1* and Six4 proteins were obtained with a T7 transcription/translation kit (Promega). The oligonucleotide

containing double-stranded myogenin MEF3 site was incubated with recombinant proteins. Competition experiments were performed in the presence of a ten-fold and hundred-fold molar excess of unlabeled identified *Myh* enhancer MEF3 sites (Enh1 to Enh6), or Myogenin promoter NFI or MEF3 sites. The sequences of the oligonucleotides are as follows;

Enh1F CTCTTGGGTAACTGGAGCCCCTC

Enh1R GAGGGGCTCCAGTTACCCAAGAG

Enh2R GGTTGACTTAGATTTCCTTATGA

Enh2F TCATAAGGAAATCTAAGTCAACC

Enh3F TGTAAGAGAAACTGAAATAAAAT

Enh3R ATTTTATTTCAGTTTCTCTTACA

Enh4F GGGGTAAGAAATCTGACGACAGG

Enh4R CCTGTCGTCAGATTTCTTACCCC

Enh5F CTATCAGGTCAGGCGACCTCAGT

Enh5R ACTGAGGTCGCCTGACCTGATAG

Enh6F CGTCAAGGAAACCTTATTCCATC

Enh6R GATGGAATAAGGTTTCCTTGACG

MyogF TGGGGGGGCTCAGGTTTCTGTGGCGT

MyogR ACGCCACAGAAACCTGAGCCCCCCCA

NF1F TATCTCTGGGTTCATGCCAGCAGGG

NF1R CCCTGCTGGCATGAACCCAGAGATA

Western blot. Western blots were performed with protein extracts of GP and soleus muscles from c*Six1*KO mice and control mice as previously described [9]. 1:1000 dilutions of anti-Six1 antibodies (HPA001893, Sigma) or anti-b-tubulin antibodies (2128, Cell Signaling) were used.

Statistical analysis. All graphs represent mean values ± SEM.

References

1. Gundersen K (2011) Excitation-transcription coupling in skeletal muscle: the molecular pathways of exercise. Biological reviews of the Cambridge Philosophical Society 86: 564–600.
2. Schiaffino S, Reggiani C (2011) Fiber types in mammalian skeletal muscles. Physiological reviews 91: 1447–1531.
3. Braun T, Gautel M (2011) Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. Nature reviews Molecular cell biology 12: 349–361.
4. Greising SM, Gransee HM, Mantilla CB, Sieck GC (2012) Systems biology of skeletal muscle: fiber type as an organizing principle. Wiley interdisciplinary reviews Systems biology and medicine 4: 457–473.
5. Shrager JB, Desjardins PR, Burkman JM, Konig SK, Stewart SK, et al. (2000) Human skeletal myosin heavy chain genes are tightly linked in the order embryonic-IIa-IId/x-ILb-perinatal-extraocular. Journal of muscle research and cell motility 21: 345–355.
6. Palstra R-J, De Laat W, Grosveld F (2008) Beta-globin regulation and long-range interactions. Advances in genetics 61: 107–142.
7. Grifone R, Demignon J, Houbron C, Souil E, Niro C, et al. (2005) Six1 and Six4 homeoproteins are required for Pax3 and Mrf expression during myogenesis in the mouse embryo. Development (Cambridge, England) 132: 2235–2249.
8. Relaix F, Demignon J, Laclef C, Pujol J, Santolini M, et al. (2013) Six Homeoproteins Directly Activate Myod Expression in the Gene Regulatory Networks That Control Early Myogenesis. PLoS Genetics 9: e1003425.
9. Richard A-F, Demignon J, Sakakibara I, Pujol J, Favier M, et al. (2011) Genesis of muscle fiber-type diversity during mouse embryogenesis relies on Six1 and Six4 gene expression. Developmental biology 359: 303–320.
10. Niro C, Demignon J, Vincent S, Liu Y, Giordani J, et al. (2010) Six1 and Six4 gene expression is necessary to activate the fast-type muscle gene program in the mouse primary myotome. Developmental Biology 338: 168–182.
11. Grifone R, Laclef C, Lopez S, Demignon J, Guidotti J, et al. (2004) Six1 and Eya1 Expression Can Reprogram Adult Muscle from the Slow-Twitch Phenotype into the Fast-Twitch Phenotype. Molecular and cellular biology 24: 6253–6267.
12. Laclef C, Hamard G, Demignon J, Souil E, Houbron C, et al. (2003) Altered myogenesis in Six1-deficient mice. Development (Cambridge, England) 130: 2239–2252.
13. Mercer TR, Mattick JS (2013) Structure and function of long noncoding RNAs in epigenetic regulation. Nature Structural & Molecular Biology 20: 300–307.
14. Lee JT (2012) Epigenetic Regulation by Long Noncoding RNAs. Science 338: 1435–1439.
15. Guttman M, Rinn JL (2012) Modular regulatory principles of large non-coding RNAs. Nature 482: 339–346.
16. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, et al. (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell 147: 358–369.
17. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annual review of biochemistry 81: 145–166.

18. Cabianca DS, Casa V, Bodega B, Xynos A, Ginelli E, et al. (2012) A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. Cell 149: 819–831.
19. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, et al. (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472: 120–124.
20. Yang L, Lin C, Liu W, Zhang J, Ohgi K a, et al. (2011) ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell 147: 773–788.
21. Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, et al. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. Science (New York, NY) 329: 689–693.
22. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311–1323.
23. Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445: 666–670.
24. Schmitz K-M, Mayer C, Postepska A, Grummt I (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes & development 24: 2264–2269.
25. Miniou P, Tiziano D, Frugier T, Roblot N, Le Meur M, et al. (1999) Gene targeting restricted to mouse striated muscle lineage. Nucleic acids research 27: e27.
26. Le Grand F, Grifone R, Mourikis P, Houbron C, Gigaud C, et al. (2012) Six1 regulates stem cell repair potential and self-renewal during skeletal muscle regeneration. The Journal of cell biology 198: 815–832.
27. An C-I, Dong Y, Hagiwara N (2011) Genome-wide mapping of Sox6 binding sites in skeletal muscle reveals both direct and indirect regulation of muscle terminal differentiation by Sox6. BMC developmental biology 11: 59.
28. Potthoff MJ, Wu H, Arnold MA, Shelton JM, Backs J, et al. (2007) Histone deacetylase degradation and MEF2 activation promote the formation of slow-twitch myofibers. The Journal of clinical investigation 117: 2459–2467.
29. Quiat D, Voelker K a, Pei J, Grishin N V, Grange RW, et al. (2011) Concerted regulation of myofiber-specific gene expression and muscle performance by the transcriptional repressor Sox6. Proceedings of the National Academy of Sciences of the United States of America 108: 10196–10201.
30. Toyoshima C, Iwasawa S, Ogawa H, Hirata A, Tsueda J, et al. (2013) Crystal structures of the calcium pump and sarcolipin in the Mg2+-bound E1 state. Nature 495: 260–264.
31. Winther A-ML, Bublitz M, Karlsen JL, Møller J V., Hansen JB, et al. (2013) The sarcolipin-bound calcium pump stabilizes calcium sites exposed to the cytoplasm. Nature 495: 265–269.
32. Liu Y, Chu A, Chakroun I, Islam U, Blais A (2010) Cooperation between myogenic regulatory factors and SIX family transcription factors is important for myoblast differentiation. Nucleic acids research 38: 6857–6871.
33. Hagège H, Klous P, Braem C, Splinter E, Dekker J, et al. (2007) Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nature protocols 2: 1722–1733.
34. Joanne P, Hourdé C, Ochala J, Caudéran Y, Medja F, et al. (2012) Impaired adaptive response to mechanical overloading in dystrophic skeletal muscle. PloS one 7: e35346.

35. Giordani J, Bajard L, Demignon J, Daubas P, Buckingham M, et al. (2007) Six proteins regulate the activation of Myf5 expression in embryonic mouse limbs. Proceedings of the National Academy of Sciences of the United States of America 104: 11310–11315.
36. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL (2008) The Vienna RNA websuite. Nucleic acids research 36: W70–4.

Acknowledgements

We thank V. Moncollin at ENS Lyon for help with the adult muscle ChIP experiments, the imaging facility at Institute Cochin for technical assistance, the sequencing and genomic platform at Institute Cochin for microarray experiments and F. Dumont and S. Jacques for advice. We thank P. Billuart for the shRNA expression vector, F. Le Grand for teaching isolated myofibers isolation, E. Perret for his help with gene expression analysis, D. Blanchot for advice on the FISH experiment and S. Gautron, L. Dandolo, F. Le Grand and A. Sotiropoulos for reading the manuscript. I.S. is supported by ANR, The Uehara Memorial Foundation and JSPS Postdoctoral Fellowships for Research Abroad. Financial support was provided by the Institut National de la Santé et la Recherche Médicale (INSERM), the "Association Française contre les Myopathies" (AFM), the Centre National de la Recherche Scientifique (CNRS) and the Agence Nationale pour la Recherche (ANR RPV09108KKA). We also acknowledge a contribution to the Institut Cochin animal care facility, made by the Région Ile de France.

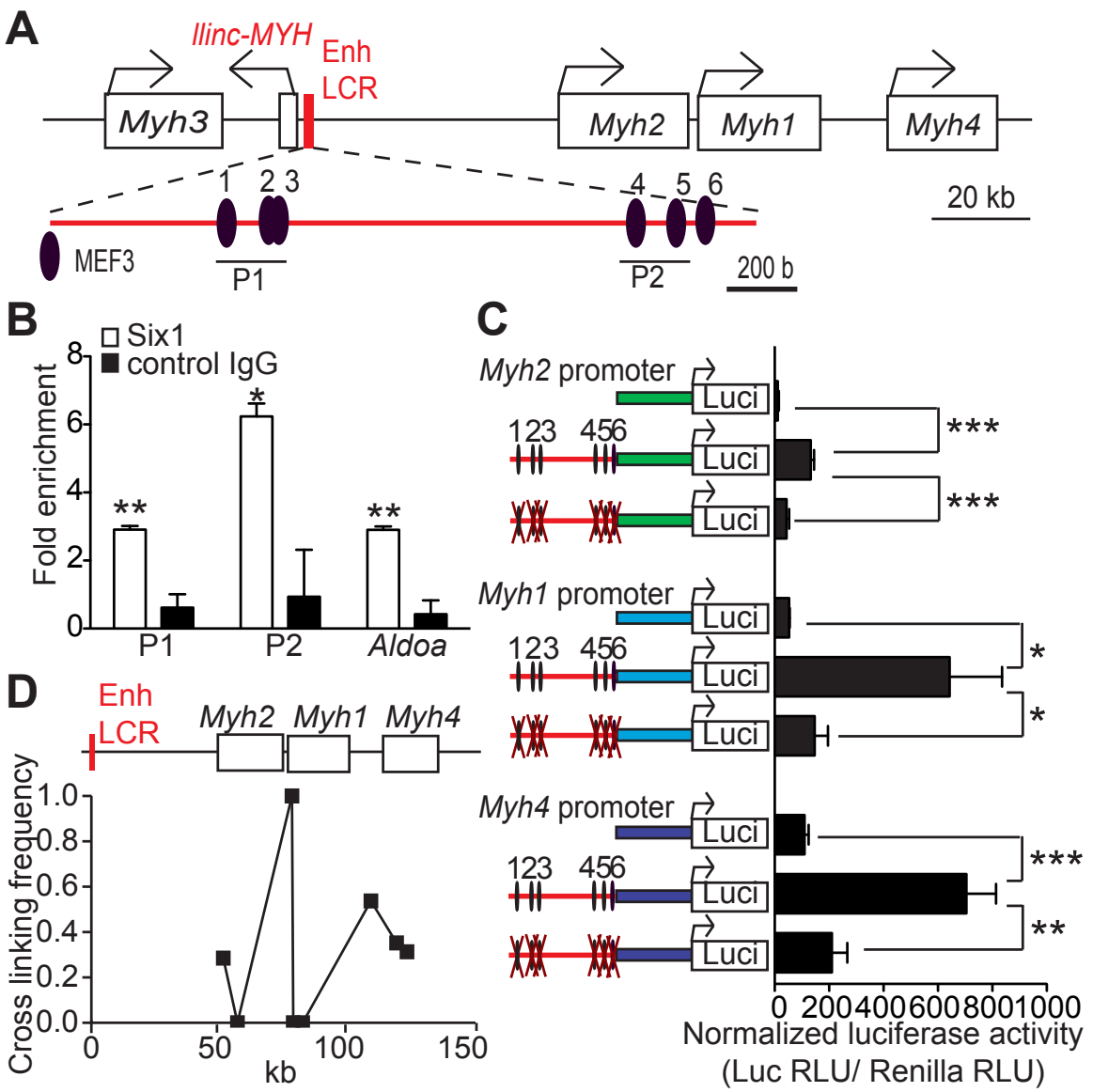


Figure 1. *Six1* binds directly the enhancer/LCR of the fast *Myh* gene cluster. (A) Schematic representation of the fast *Myh* gene cluster. (B) qPCR values of ChIP experiments performed with Six1 antibodies, or IgG on GP and TA chromatin, showing Six1 binding to P1 and P2 regions of the fast *Myh* enhancer and to the muscle promoter of *AldoA* (n=3). (C) Luciferase assays from adult TA muscles electroporated with luciferase vectors (indicated) and a TK-Rennila luciferase vector allowing normalization (n=4). (D) qPCR experiments from 3C assays of wild type EDL muscle, showing the direct interactions of *Myh2*, *Myh1* and *Myh4* promoters with the fast *Myh* LCR/enhancer. *P<0.05, **P<0.01, ***P<0.001.

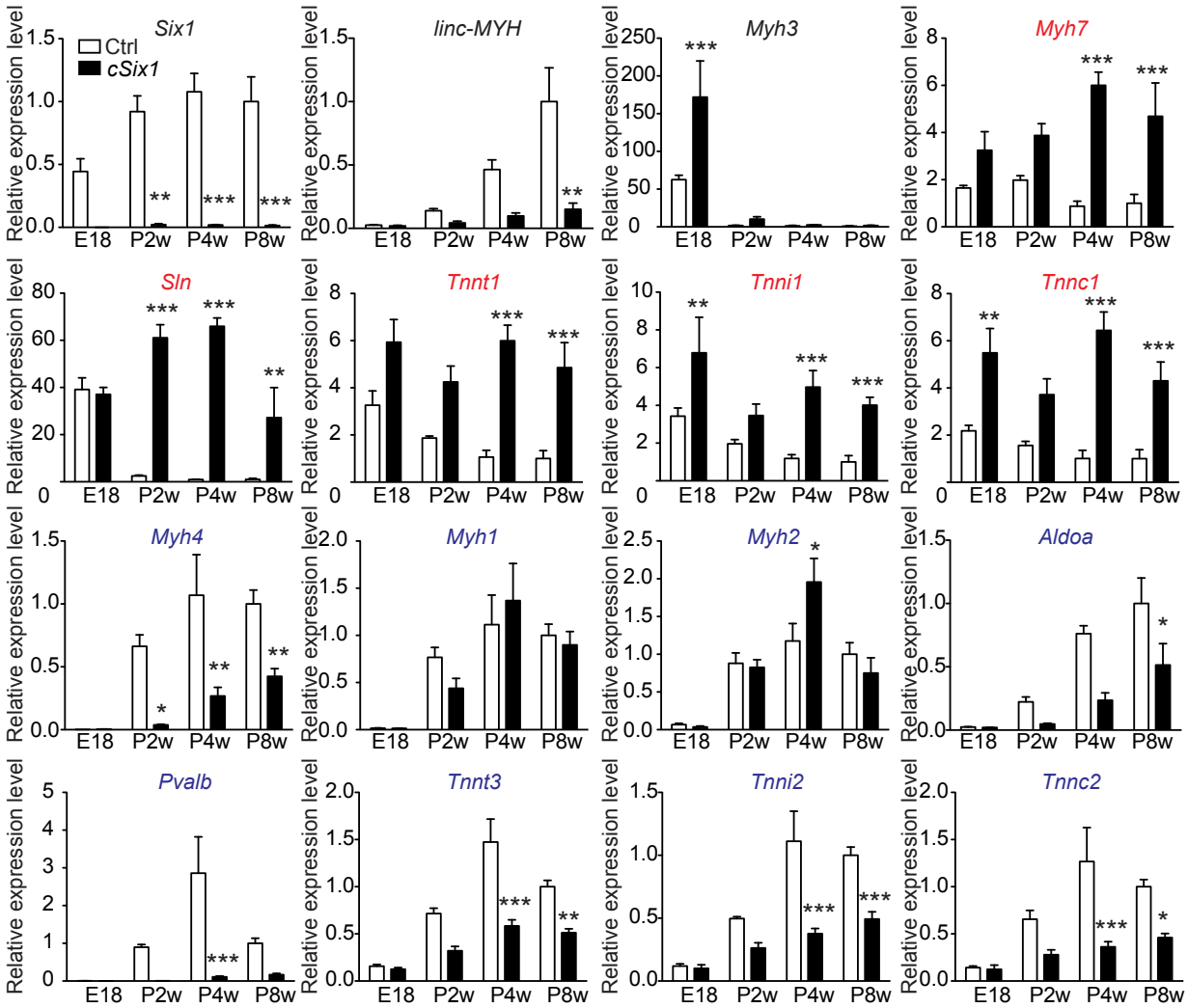


Figure 2. The expression of fast-type genes and *linc-MYH* is impaired in *cSix1 KO* mice during postnatal development. mRNA expression level of *Six1*, *linc-MYH*, *Myh3*, slow-type genes (red) and fast-type genes (blue) in back muscles of *cSix1 KO* mice at E18.5, P2W, P4W and P8W, as determined by qPCR experiments, (n=3 to 6 for each point). *P<0.05, **P<0.01, ***P<0.001.

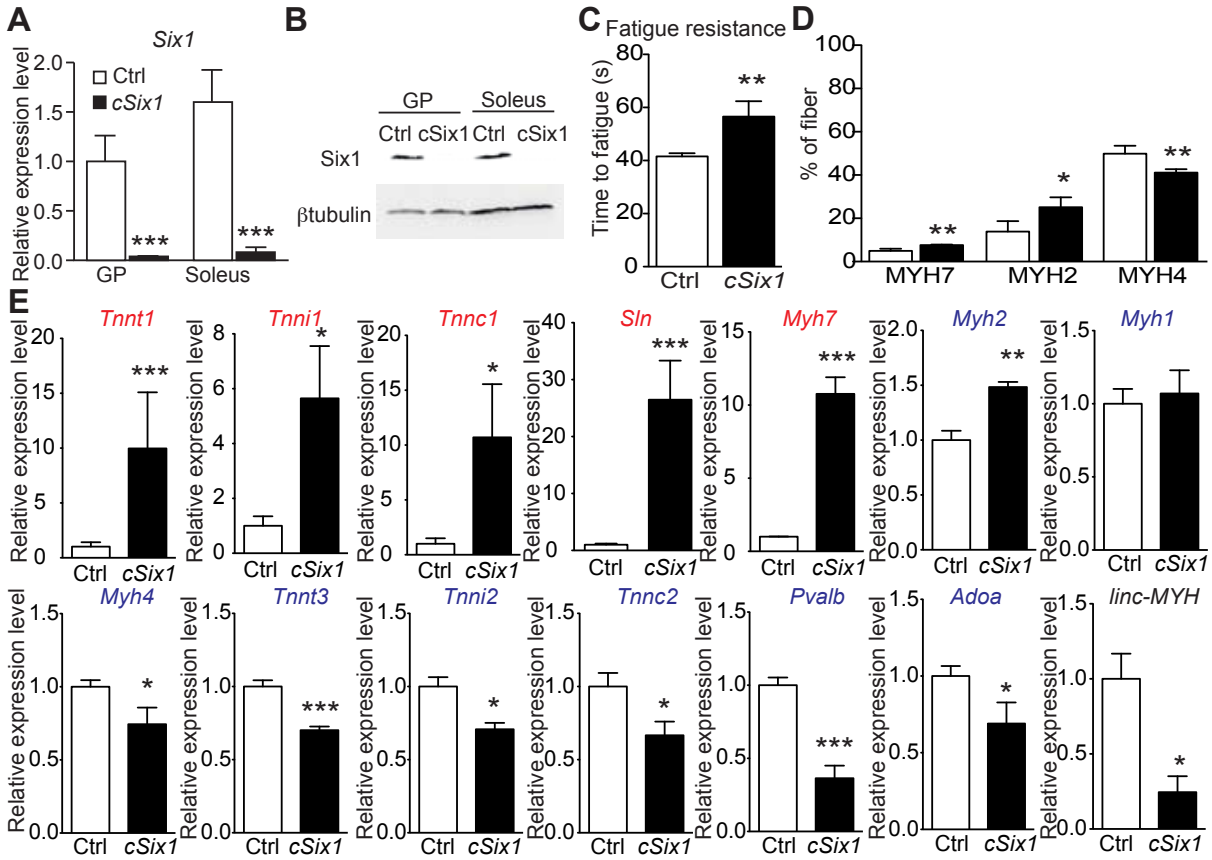


Figure 3. *Six1* deficiency impairs the adult phenotype of fast muscle. (A) *Six1* mRNA expression levels in GP and Sol muscles of three month-old control (Ctrl, n=4) and *cSix1 KO* (n=3) mice. (B) Western blot analysis of Six1 and βtubulin expression in Sol and GP of Ctrl and *cSix1 KO* mice. (C) Time to fatigue ratio of TA muscles of Ctrl (n=4) and *cSix1 KO* (n=4) mice. (D) Percentage of myofibers expressing MYH7, MYH2 and MYH4 in TA muscles of three month-old Ctrl (n=4) and *cSix1 KO* (n=4) mice. (E) mRNA expression levels of slow-type genes (red), fast-type genes (blue) and *linc-MYH* in TA muscles of three month-old Ctrl (n=4) and *cSix1 KO* (n=4) mice. *P<0.05, **P<0.01, ***P<0.001.

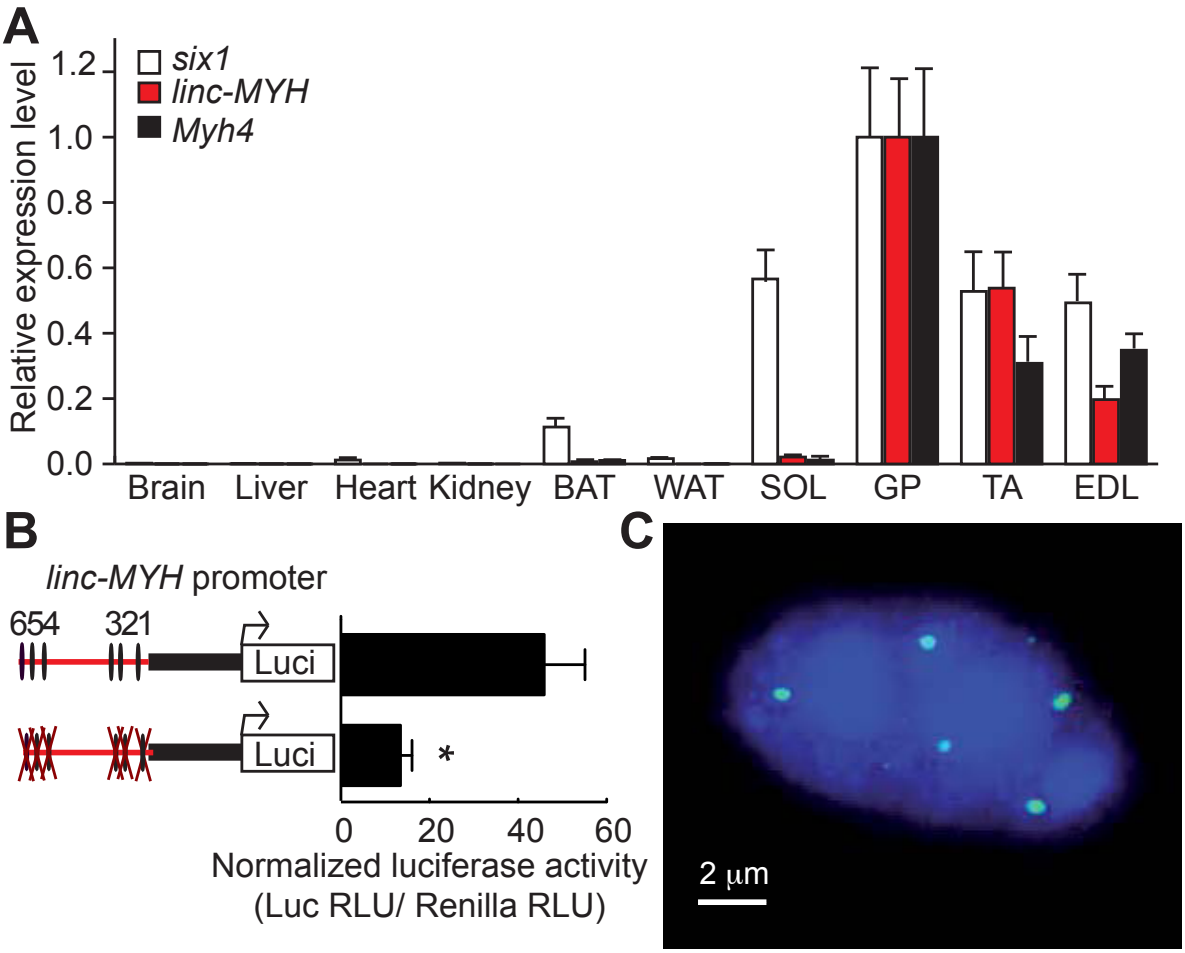


Figure 4. *linc-MYH* is expressed in adult fast-type skeletal muscles and accumulates in their nuclei. (A) Tissue distribution of *Six1, Myh4* and *linc-MYH* RNAs. BAT, brown adipose tissue; WAT, white adipose tissue. (n=4). (B) Luciferase assays of adult TA electroporated with *linc-MYH* promoter luciferase vectors (indicated) and a TK-Renilla luciferase vector (allowing normalization). *P<0.05. (C) FISH of isolated EDL myofiber with a *linc-MYH* antisense RNA fluorescent-labeled probe (green) and Dapi staining (blue).

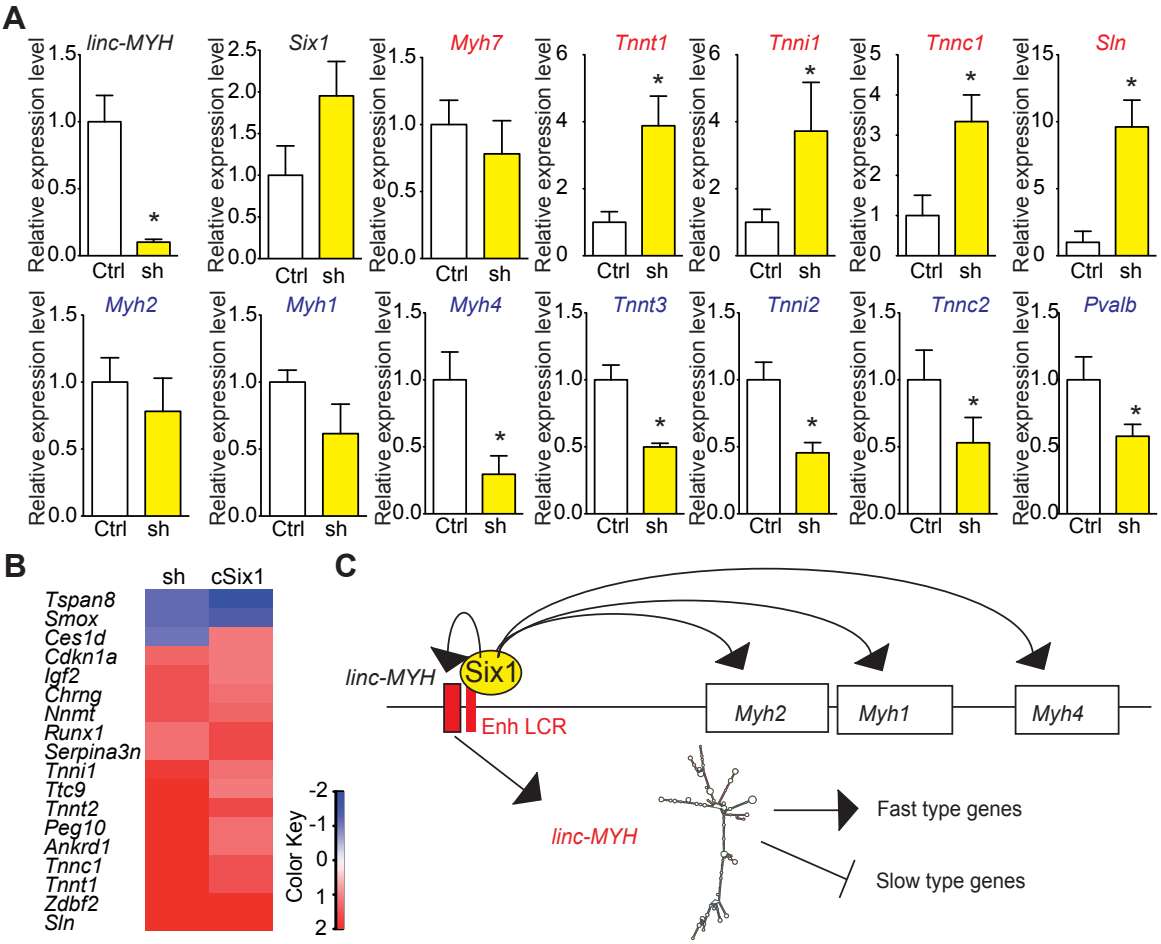


Figure 5. Slow-type gene expression is suppressed by *linc-MYH*. (A) qPCR experiments revealing mRNA expression levels of *linc-MYH*, *Six1*, slow-type genes (red) and fast-type genes (blue) in TA muscles expressing a shRNA directed against *linc-MYH* (n=3), or a sh*CtrlRNA* (n=3). (B) Microarray analysis of TA muscles transfected by shRNA against *linc-MYH*: a heat map of genes (red) upregulated to more than double the levels observed in c*Six1*KO. (C) A model of *Six1* controlling the expression of the different *MYH* and of *linc-MYH* at the fast *Myh* locus. Below, the hypothesis explaining the *linc-MYH* mode of action, as supported by the transcriptomic-wide analysis performed after *linc-MYH* knockdown in fast TA. *P<0.05.

Supporting Information

```
>P1 MusMus chr11 66933250-66933440
                                 MEF3 1          EBOX
MusMus      GTCACTCACCTCTTGGGTAACTGGAGCCCCTCAGCTGTTTGGGTCCTGCTTTGGAAGGACATTCCAGTGT
RatNor      GTCACTCACATCTTGGGTAACTGGAGCTCCTCAGCTGTTTGGGTCTTGGTTTGGAAGGACATTCCAGTGT
HomSap      GTCACTCCCCTCTAGGGTAACTGGAGCCCTTCAGCTGTTTGGGTGCCTGTTTGAAAGGACATCTCAGCAT
BosTau      GTCACTCACGTCTTGGGTAACTGGAGCCCTTCAGCTGTTTGGGTCCCTGTTTGGAAGGACGTCTCAGCAT
EquCab      GTCACTCACTTCTTGGGTAACTGGAGCCCTTCAGCTGTTTGGGTCCCTGTTTGGAAGGACATCTCAGCAC
                                                      EBOX                   MEF3 2
MusMus      TGTCCTG-GCTCCTGGAACAGCCAGCCTCCCAGAGGGCTTCACCTGTCAAAGCTAGGGGGTTGACTTAGA
RatNor      TGTCCTG-GCTCCTGGAACAGCTAGCCTCCCAGAGGACTTCACCTGTCAAAGCTG-GGGGTTGACTTAGA
HomSap      CACCCAGGGCTCCCGGGACAGCCAACCTCCCAGAAGGCTTCATCTGTCAGAGCAG-GAG-CTTACTTAGA
BosTau      TGCCCAGGGCTCTGGGGGCAGCCAACTTCTCAGAAGGTTTCATCTGTCAGCCCAA-GAGGCTGACTTAGC
EquCab      TTTCTGGGGCTCCTGGGGCGGCCAACTGCCCACAAGGCTTCATGTGTCAAAGCAG-GAGGCTGACGCAGA
                                        MEF3 3
MusMus      TTTCCTTATGAGCACTCTGTAAGAGAAACTGAAATAAAATAAATCAATAAAT
RatNor      TTTCCTTATGAGAACTCTGTAAGAGAAATCGAAATAAAATAAATCACCA---
HomSap      ATTCCTTACAAGAACACTGCAAGAGCAATTAAAATAAGTTA-----TCT---
BosTau      TTTCCTCAGAAGAACTTTCCAAGAGCAATTAAAATAAGCTA-----TCT---
EquCab      TCTCCTTACAAGAACTTTGCGCGAGCAATGAAA----GTTA-----CCT---

>P2 MusMus chr11 66934405-66934621
                                 MEF3 4
MusMus      CTTGATCCCTCTGGGGTAAGAAATCTGACGACAGGGTGAGCCTGCCAGGCGTGGCCTCTTGACTCTGA--
RatNor      CTTGATCCCTCTGGGGTAAGAAATCCGAGGACAGGATGAGCCTGCCAGGCATGGCCTCTTGACTCTGA--
HomSap      CTTGATCCCCATGGGGTAAGAAATCTGAGGACAGAATGAGCCTGCCAGCCATGTCCTCCTGACATTGCAA
BosTau      CTTGATCCCCACATGGAAAGAAATCTGAGGACAGAATGAGCTTCTCC---GTGTCCTCCCCACACTGCAA
EquCab      CTTGATCCCTGTGGGGAAGGAAATCTGAGGACA--------------------------------TTGCAA
                                                                          MEF3 5
MusMus      ATCTGGCTGAGAGATGGGCCCGGAGTTACAGCTCCTGCTGGGAATGTTCTCAGAAACTCT---ATCAGGT
RatNor      ATCTGGCTGAGAGCTGGGGCCAGAGCTACAGCGCCTGCTGGGAATGTCCTCAGAAACTCT---ATCAGGT
HomSap      ATCTGGCCAAGGGATGGGGCTGAACTTGCAAATCCCTCCGGGGCTATTCTCAGAAACCAAGTGATGGGGA
BosTau      ACTTGGCCAAGGGATGGGGCCGAAGCTGTAAATCCCATCGGGGCTGTTCTCAGAAATCAAGTGGTGGGGT
EquCab      ATCTGGCCAAGAAATGGGGCCAAAGCTGTAAATCCCAACAGGGCTGTTCTCAGAAACCAAGCCATGGGAT
                      EBOX                                                MEF3 6
MusMus      CAGGCGACCTCAGTTGATCTGCCCGACC--CTGGGTTTCTCGGTCGACCCTCGTCAAGGAAACCTT----
RatNor      CAAGCGACCTCAGTTGATCTGCCCGACA--CTGGGTTTCTCTGTCGACCCTCGTCAAGGAAACCGA----
HomSap      AACCTGACCTCAGCTGATCTACCTGATATGCTGGGTTTCTCTACTGACCCTGACCAAGGTAATGTT----
BosTau      TACCTGACCTCAACTGATCTACCCAACT--CCGGACTTCTGCACTGACC-TGACCAAGGAGACCTGGCCT
EquCab      CATCTGGCCTCAATTGATCTACCTGACC--CTGGGCTTCTCTAGCAACCCCGACCAAGGAAATGTT----

MusMus      --A-----TT--CCATCATG--TTATTCT
RatNor      --ACTCAATT--CCGTCATG--CTATACT
HomSap      --A-----TTATCAATCATACGTTATTCA
BosTau      ATA-----TCATCAATCATAGGTTTCCCA
EquCab      ----------ATCAATCATATGTTATTCA
```

Figure S1. Sequences of P1 and P2 boxes of the *Myh* enhancer. Sequences of P1 and P2 boxes of the *Myh* enhancer in mouse, rat, human, bovine and equinides species, and showing the sequence conservation of the six MEF3 sites and E boxes.

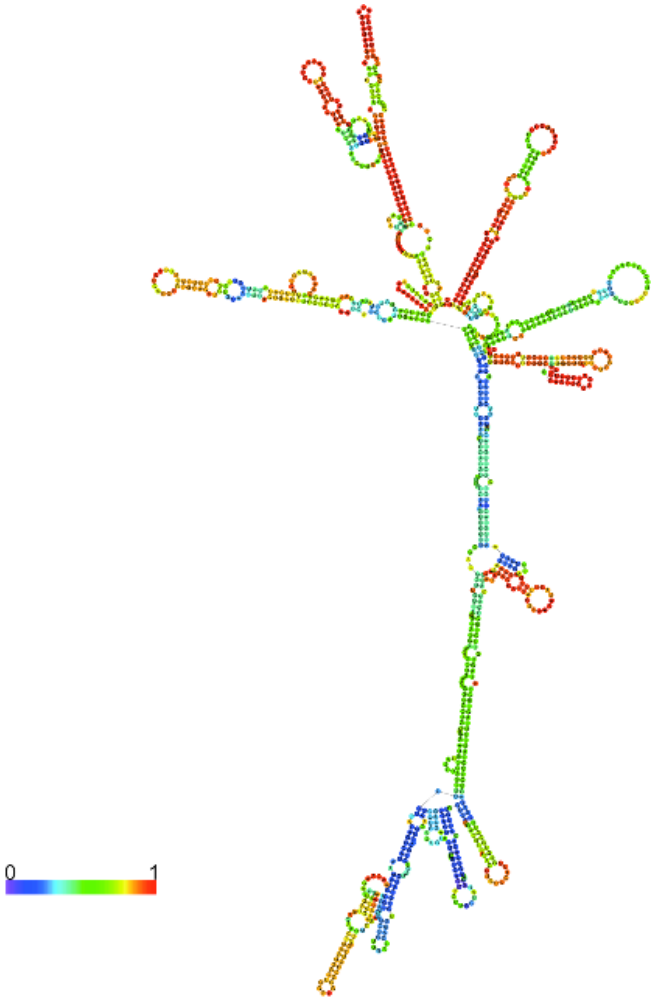


Figure S2. Predicted linc-MYH structure. Predicted minimum free energy (MFE) structure of the 1050nt long linc-MYH, as determined by RNAfold [36]. The color encodes base-pair probabilities.

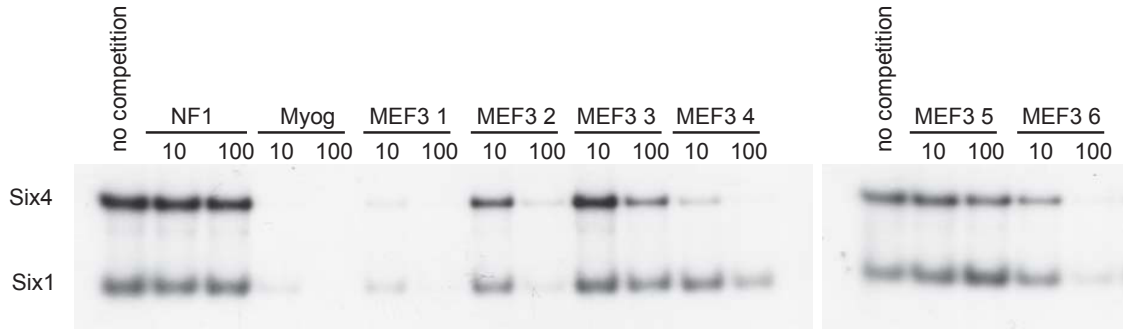


Figure S3. Competitive Electromobility shift assays. Competitive Electromobility shift assays performed with recombinant Six1 and Six4 proteins and labeled Myogenin MEF3 oligonucleotide and 10 or 100 fold molar excess of unlabelled oligonucleotides containing Myogenin MEF3 or NF1 site, or with MYH MEF3 sites (1, 2, 3, 4, 5, 6) whose sequence is presented on Figure S1.

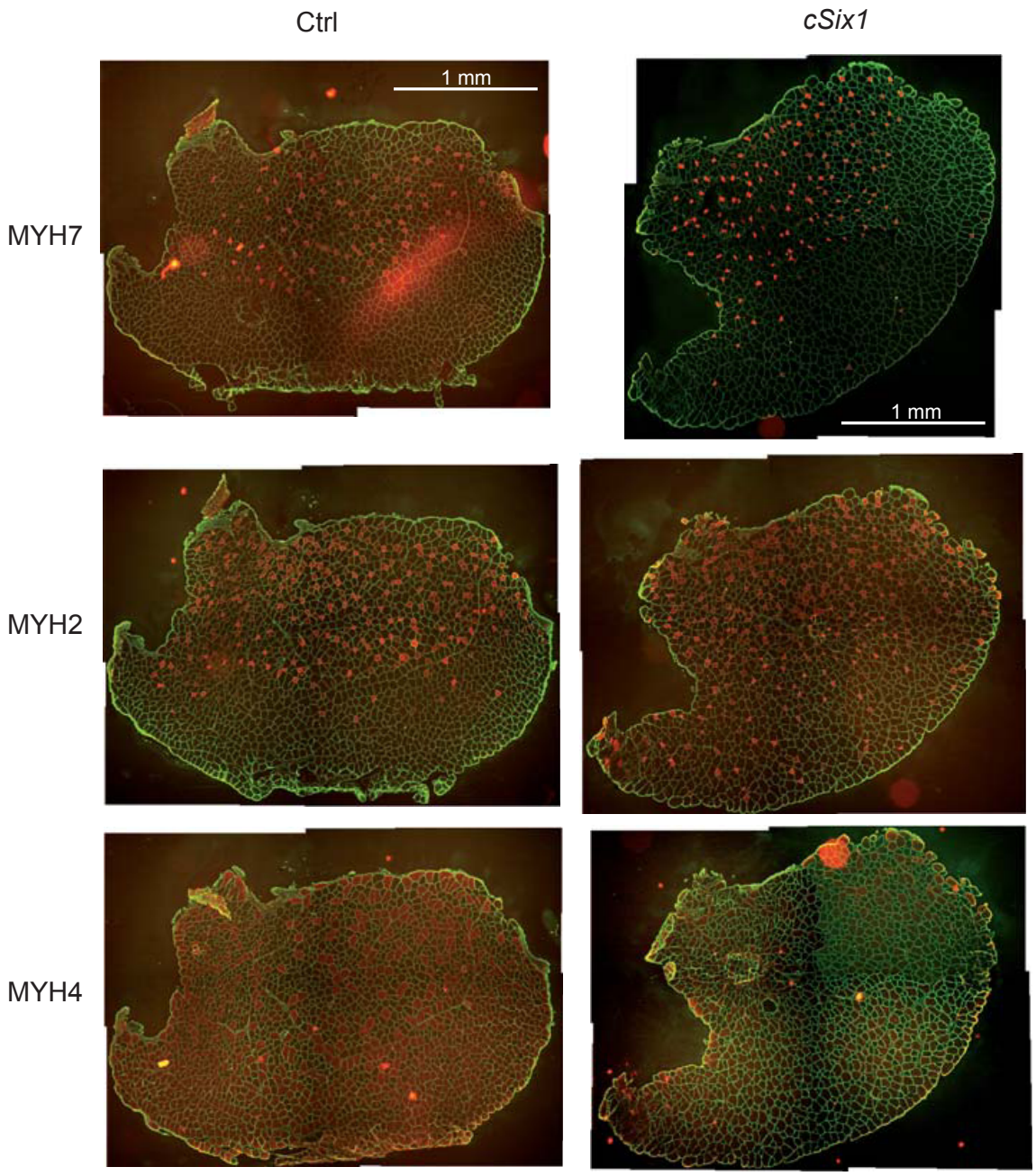


Figure S4. Immunostaining of MYH proteins. Immunostaining of MYH7 (red), MYH2 (red), MYH4 (red) and laminin (green) in TA of 12 weeks old control and *cSix1KO* male mice.

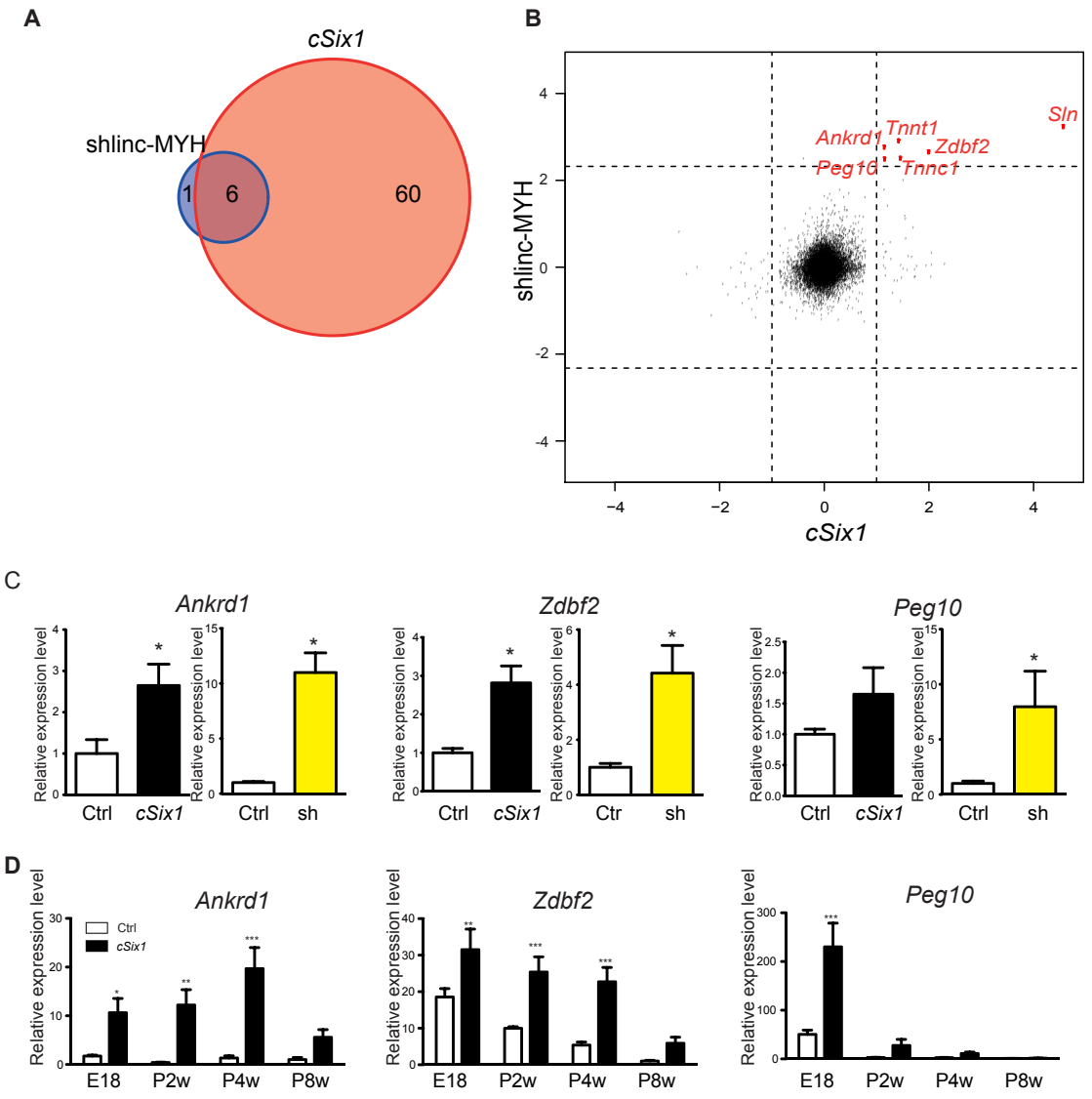


Figure S5. Comparison between *cSix1* mice and shlinc-MYH treated mice. (*A*) Venn diagram showing the overlap between genes that are up-regulated more than five fold in shlinc-MYH and two fold in *cSix1* muscles. ($p=10^{-17}$ as given by a hypergeometric test). (*B*) Scatter plot of mRNA expression fold change (log2) as determined by microarray analysis of shlinc-MYH and *cSix1*. Genes that are up-regulated more than five fold in shlinc-MYH and two fold in cSix1 muscles are indicated in red. **(*C*)** mRNA expression levels in shlinc-MYH knock-down and cSix1 TA muscles, as measured by qPCR experiments. Ctrl (means ± SEM; n=4), *cSix1* (means ± SEM; n=4), shCtrl (means ± SEM; n=3), shlinc-MYH (means ± SEM; n=3). (*D*) mRNA expression level of *Ankrd1*, *Zdbf2* and *Peg10* and in back muscles of *cSix1* KO mice

at E18.5, P2W, P4W and P8W, as determined by qPCR experiments, (means ± SEM; n=3 to 6 for each point).

Table S1. Microarrays analysis of TA muscles electroporated by shRNA

Table S2. Microarrays analysis of GP muscles of cSix1

Table S3. Sequence of the oligonucleotides used for 3C.

name of oligonucleotides	Sequence (5'- 3')
probe	FAM- TCAGCTGCCCAGGGTGACCA-Tamra
Enh_3CF	CCAGCCTGTTCTGGGTACAT
MYH1_3CR	ACCCCTTGGAATGAGAGTGA
MYH2_3CR	TGAAGCAGTGTGGAACAAGC
MYH4_3CR	CCAAATTGGTTGATGCTCATT
78038_3CR	CCTGACGCACCATGTCTAAA
104838_3CR	CAGGATTTGGTAGGGGATGA
106285_3CR	CGCACAGCCTAATGAAGACA
134682_3CR	CCCTCTCATATGGTGCCAGT
148019_3CR	GGTCACTGGAGGGATCTGAA

Table S4. Sequence of the oligonucleotides for qPCR.

gene name	Forward (5'- 3')	Reverse (5'- 3')
Six1	CTTTAAGGAGAAGTCTCGGG	TTCCAGAGGAGAGAGTTGAT
MYH7	AGGGCGACCTCAACGAGAT	CAGCAGACTCTGGAGGCTCTT
MYH2	CCAAGAAAGGTGCCAAGAAG	CGGGAGTCTTGGTTTCATTG
MYH1	CGGTGGTGGAAAGAAAGG	CAGGAGTCTTGGTTTCATT
MYH4	GCTTGAAAACGAGGTGGAAA	CCTCCTCAGCCTGTCTCTTG
MYH3	GCAAAGACCCGTGACTTCACCT CTAG	GCATGTGGAAAAGTGATACGTG G
Tnnt1	CCCCCGAAGATTCCAGAAGG	TGCGGTCTTTTAGTGCAATGAG
Tnnt3	GGAACGCCAGAACAGATTGG	TGGAGGACAGAGCCTTTTTCTT
Tnni1	ATGCCGGAAGTTGAGAGGAAA	TCCGAGAGGTAACGCACCTT
Tnni2	AGAGTGTGATGCTCCAGATAGC	AGCAACGTCGATCTTCGCA
Tnnc1	GCGGTAGAACAGTTGACAGAG	CCAGCTCCTTGGTGCTGAT

Tnnc2	ATGGCAGCGGTACTATCGACT	CCTTCGCATCCTCTTTCATCTG
Sln	GGTCCTTGGTAGCCTGAGTG	CGGTGATGAGGACAACTGTG
Pvalb	ATCAAGAAGGCGATAGGAGCC	GGCCAGAAGCGTCTTTGTT
linc-MYH	GTGCAGCCAGAACAAGACAG	CAAGATGGGAGGCTCTCAAA
Aldoa	ACTCTCTGCTGACCGGGCTCT	AATGCTTCCGGTGGACTCAT
Zdbf2	TAGCGGCTCTTTCGAGAGAC	CCCTGATCTGGGGAGTCAA
Peg10	TGCACAACTACACTGCCTTTATG	CTGGGCAATCATCTGGAATGC
Ankrd1	TGCGATGAGTATAAACGGACG	GTGGATTCAAGCATATCTCGGA A
Actb	GGCTGTATTCCCCTCCATCG	CCAGTTGGTAACAATGCCATGT

5.4 Synergie entre Six et MyoD au cours de la myogenèse

Dans cette dernière partie, nous présentons un autre projet que nous avons réalisé en collaboration avec Iori Sakakibara dans l'équipe de P Maire. La question sous-tendant ce projet est la suivante : étant donné que l'expérience ChIP-seq de MyoD prédit plus de 30,000 régions liées (Cao et al., 2010), mais que le nombre de gènes cibles est bien inférieur – peut-être quelques centaines (Bergstrom et al., 2002) –, peut-on prédire les pics effectivement fonctionnels, c'est-à-dire ceux qui induisent l'expression d'un gène cible ? Pour ce faire, nous pouvons utiliser le fait que les CRMs fonctionnels ont tendance à concentrer des sites de fixation pour différents TFs et à être conservés chez des espèces proches (voir section 1.6). Dans notre cas, les protéines Six constituent un cofacteur important de la myogenèse et notamment de MyoD : nous avons donc voulu voir si la présence de sites MEF3 conservés au sein des pics ChIP-seq de MyoD permettait de filtrer les données et de récupérer les pics fonctionnels.

5.4.1 État de l'art sur la coopération Six/MyoD

L'idée d'une coopération entre Six et MyoD au cours de la myogenèse n'est pas nouvelle. Ainsi, le promoteur de Myog contient un site MEF3 fixant Six1,4 et une boîte E fixant MyoD. La mutation de MEF3 abolit totalement l'expression d'un transgène *in vivo* (Spitz et al., 1998), et la mutation de la boîte E plus partiellement (Cheng et al., 1993). De manière similaire, dans le cas du CE de *Myod1*, la mutation de la boîte E abolit l'expression dans les somites (Kucharczuk et al., 1999) et celle des sites MEF3 l'abolit plus généralement (voir section 5.3.1). On observe dans les deux cas un phénomène du type « tout ou rien » dans lequel l'expression du gène cible dépend de la présence conjointe de Six1,4 et MyoD sur l'élément régulateur : on parle de synergie. Celle-ci a été récemment étudiée *in vitro* de manière quantitative. Ainsi, Liu et al. (2010) ont montré dans des cellules 293T (modèles de cellules du rein) qu'un promoteur synthétique composé de 4 boîtes E et de 6 sites MEF3 multimérisés est activé 5 fois plus lorsque Six4 et MyoD sont présents dans le milieu que lorsque seulement l'une ou l'autre des protéines est présente. Dans les mêmes conditions, le promoteur de Myog montre une synergie [21] d'un facteur 1.6, abolie (1.04) lors de la mutation du site MEF3. Par ailleurs, les auteurs ont réalisé une expérience de ChIP-on-chip pour Six1 dans des cellules C2C12 couvrant 17% du génome (voir 5.2.2), et ont noté que 40% des sites de fixation de Six1 détectés empiètent sur un pic ChIP-seq de MyoD obtenu précédemment par Cao et al. (2010). Ces données suggèrent un rôle plus étendu de la synergie Six/MyoD sur le génome.

Le mécanisme sous-jacent à cette synergie *in vivo* n'est pas complètement clair. Il peut y avoir l'activation par Six et MyoD d'étapes distinctes du processus de transcription, ou encore la mise en place d'une structure de chromatine permissive au niveau des éléments régulateurs ou du promoteur (Fuda et al., 2009). Par exemple, MyoD peut recruter des co-activateurs transcriptionnels comme l'acétyltransférase *P300/CBP-associated factor* ou PCAF (Puri et al., 1997), ou encore la protéine *TATA-binding protein Associated Factor* 3 ou TAF3 (Deato and Tjian, 2007) au niveau des séquences régulatrices. Les protéines Six, quant à elles, semblent pouvoir recruter la protéine CBP (*CREB-binding Protein*) liée à l'acétyltransférase p300 *via* leurs cofacteurs Eya (Ikeda et al., 2002). De plus, il a été montré que le recrutement par Six4 de l'histone déméthylase UTX au niveau des promoteurs de *Myog* et de *Ckm* (*Muscle Creatine Kinase*) est requis pour permettre l'induction de ces gènes au début de la différenciation des myoblastes (Aziz

21. La synergie est calculée comme étant le rapport entre la valeur de l'expression du rapporteur dans le cas Six4+MyoD et la valeur de la somme des expressions dans les cas Six4 et MyoD seuls.

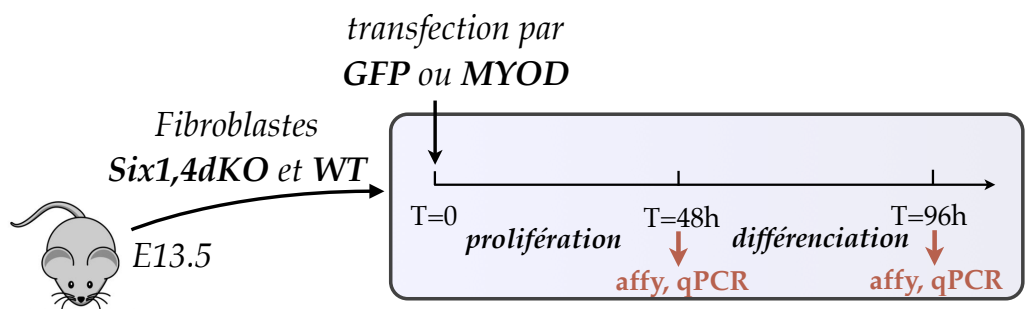


FIGURE 5.10 – **Description de l'expérience permettant d'étudier la synergie Six+MyoD.** Des fibroblastes primaires sont prélevés à partir d'un embryon de souris *wild-type* (WT) ou Six1,4dKO à E13.5 et sont mis en culture. À T= 0 ils sont électroporés par un vecteur d'expression codant pour MyoD ou pour la GFP puis cultivés en milieu de prolifération pendant 48h et en milieu de différenciation pendant 48h. Trois répliquas sont prélevés à 48h et à 96h pour étudier l'expression génétique dans les différentes conditions.

et al., 2010). Une autre étude a montré que le locus de Myog est méthylé au niveau de l'ADN dans des myoblastes non différenciés, et que Six1 permet la perte de la méthylation CpG et l'activation de Myog (Palacios et al., 2010). Ces études suggèrent la possibilité que les facteurs Six recrutent des enzymes rendant la chromatine accessible aux MRFs et permettant ainsi la transcription, ce qui peut expliquer la synergie.

Ainsi, ces différentes études suggèrent une action concertée entre les protéines Six et MyoD pour activer des gènes cibles à l'échelle génomique. Plusieurs questions restent néanmoins ouvertes : est-ce que la combinaison de sites de fixation pour Six et MyoD suffit à prédire les enhancers liés à la régulation du destin myogénique ? Quels sont les autres corégulateurs, et comment modulent-ils l'activité ?

5.4.2 Obtention de données d'expression Six+MyoD

Afin d'étudier la coopération entre Six1,4 et MyoD à l'échelle du génome, nous avons d'abord cherché à mettre en place un modèle cellulaire au sein duquel nous pourrions contrôler la présence ou l'absence de Six1,4 et MyoD. Nous avons pour cela choisi les fibroblastes (fig. 5.10). En effet, ces cellules n'expriment pas MyoD, mais lorsque l'expression de MyoD y est forcée, elles se transdifférencient en cellules musculaires (Davis et al., 1987). Par ailleurs, l'équipe de P. Maire possède des souris hétérozygotes en Six1 et Six4 qui permettent de générer des embryons Six1,4dKO (Grifone et al., 2005). Ainsi, il est possible d'obtenir des fibroblastes provenant d'embryons de souris à E13.5 qui sont soit *wild-type* (WT, Six+/MyoD-) soit Six1,4dKO (Six-/MyoD-). Enfin, ces fibroblastes peuvent être électroporés par un vecteur d'expression codant pour MyoD, générant les cas restants Six+/MyoD+ et Six-/MyoD+. Les fibroblastes qui ne sont pas transfectés par MyoD sont transfectés par un vecteur exprimant la GFP afin de produire des conditions similaires. Les fibroblastes sont ensuite mis en culture pendant 48h en milieu de prolifération (« prolif »), simulant ainsi les stades précoces de prolifération des myoblastes au niveau des somites, puis à nouveau pendant 48h en milieu de différenciation (« diff »), modélisant la formation de myotubes. Les fibroblastes sont prélevés

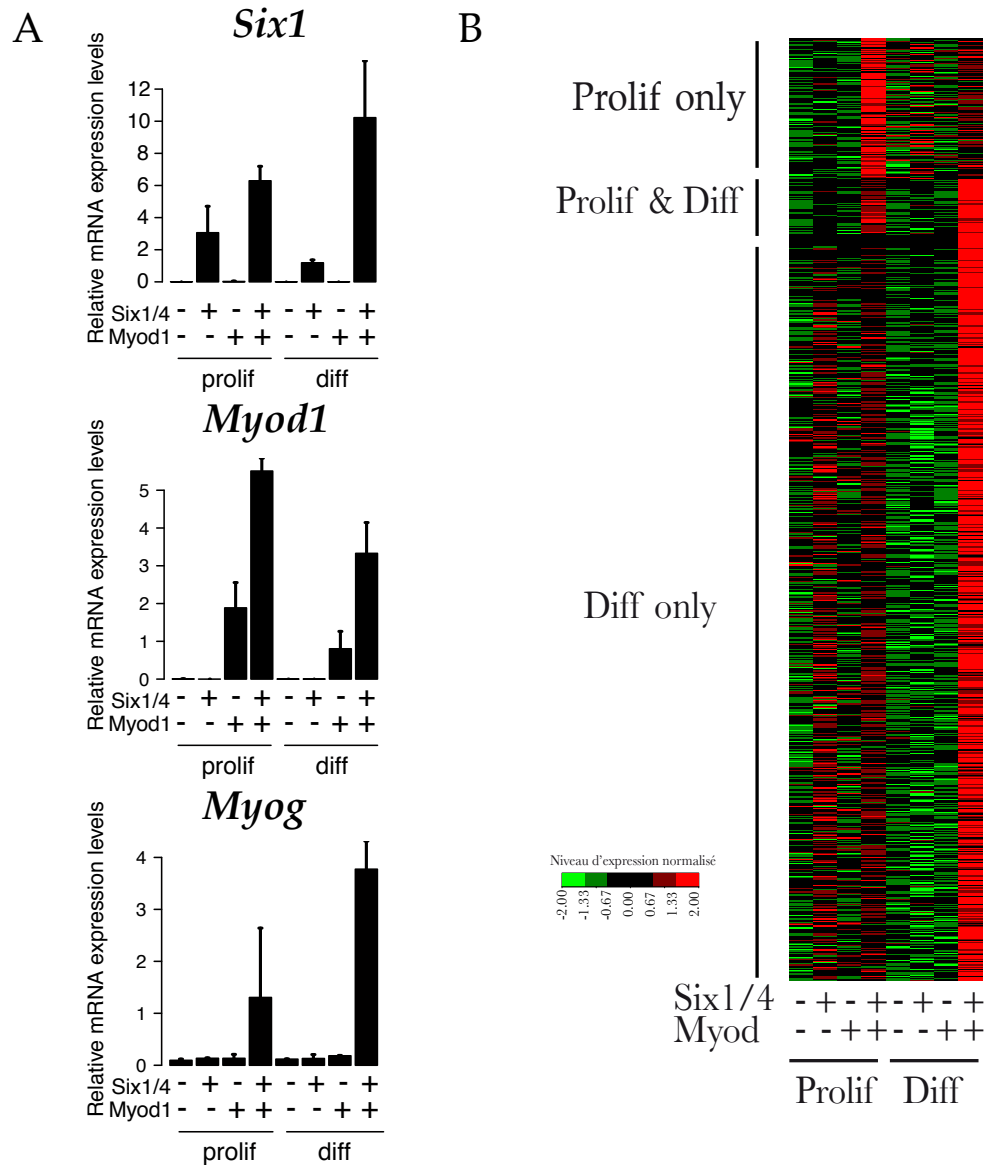


FIGURE 5.11 – **Données d'expression des gènes activés par la synergie Six+MyoD.**
(A) Quantification de l'expression des ARNs pour *Six*1, *Myod*1 et *Myog* par RT-PCR quantitative dans les différentes conditions considérées : respectivement milieu de prolifération ou milieu de différenciation, et pour chaque cas fibroblastes Six1,4dKO et WT transfectés par GFP ou par MyoD. On note l'absence d'expression de MyoD dans les fibroblastes qui ne sont pas transfectés par le vecteur d'expression codant pour MyoD, et de Six1 dans les fibroblastes Six1,4dKO. On observe pour *Myog* une expression synergique, c'est-à-dire qu'elle dépend de la présence conjointe de Six1 et MyoD. Les barres d'erreur correspondent à l'erreur type de la moyenne.
(B) Données affymetrix pour les 761 gènes synergiques. Par souci de clarté, les données sont normalisées par gène et par milieu de culture (prolif ou diff). Les gènes sont classés en trois catégories selon leur profil d'expression : prolif only, prolif & diff, et diff only.

à T= 48h et à T= 96h afin de recueillir des données d'expression sur l'ensemble du génome par hybridation sur puce à ADN (technologie Affymetrix) ainsi que de manière plus précise sur certains gènes cibles par RT-PCR quantitative [22] (abrégée dans la suite par qPCR). Les données sont scindées en trois répliquas afin d'évaluer le bruit expérimental. Les barres d'erreur présentées dans la suite réfèrent à l'erreur type de la moyenne ou SEM (*Standard Error of the Mean*) de ces répliquas, soit dans ce cas $\sigma/\sqrt{3}$, où σ est l'écart type mesuré sur les 3 points.

Nous montrons en figure 5.11 les résultats des données d'expression obtenues. D'abord, les données d'expression de Six1, MyoD et Myog par qPCR sont montrées en fig.5.11A. On observe que Six1 n'est exprimé que dans les fibroblastes WT, comme attendu. Les autres gènes de la famille Six ne sont pas exprimé dans ce contexte (données affymetrix, non montré). De manière similaire, MyoD n'est exprimé que dans les fibroblastes transfectés par le vecteur d'expression codant pour MyoD. Enfin, nous retrouvons bien que Myog n'est exprimé que lorsque Six et MyoD sont tous deux exprimés, corroborant les résultats *in vitro* de Liu et al. (2010). En figure 5.11B, nous présentons les résultats des puces affymetrix pour 761 gènes dits « synergiques ». Ces gènes ont été définis par le fait que l'expression obtenue dans le cas Six+/MyoD+ est au moins 1.5 fois plus grande que l'expression maximale des cas Six-/MyoD-, Six-/MyoD+ et Six+/MyoD-, que ce soit en prolif ou en diff (nous justifions ce critère après). Ces gènes se partagent en 125 gènes synergiques seulement en milieu de prolifération (prolif only), 49 en milieux de prolifération et de différenciation (prolif & diff), et 665 seulement en milieu de différenciation (diff only). Nous avons aussi noté la présence de 375 gènes montrant une synergie négative au même seuil de 1.5. Néanmoins, ceux-ci ne présentent pas de gènes ayant des ontologies liées au développement musculaire, et ne sont pas corrélés à la présence d'enhancers Six+MyoD (voir après), laissant à penser qu'ils sortent du cadre du modèle de différenciation musculaire que nous cherchons à étudier.

5.4.3 Obtention de régions de régulation Six+MyoD

Afin de comprendre la régulation des gènes synergiques mis en évidence, nous avons cherché à identifier des éléments régulateurs pouvant exhiber une coopération Six+MyoD. En utilisant le motif MEF3 présenté en figure 5.5, nous avons scanné le génome à la recherche d'instances conservées chez les mammifères avec un seuil de 10.5 bits (cf note de bas de page 19). La conservation est définie selon les critères d'Imogene (voir article en section 3.2). Ainsi, nous trouvons 45,628 instances MEF3 conservées dans le génome. Nous avons ensuite comparé ces données aux 32,272 pics ChIP-seq de Cao et al. (2010) obtenus après fusion des conditions MB50, MB95 et MT. Pour chaque pic ChIP-seq de MyoD, nous avons cherché les instances MEF3 conservées dans une région de 20kb et avons mesuré la distance entre ces instances et le centre du pic. Les données agrégées sont montrées dans l'histogramme de la figure 5.12A. On observe une concentration de sites MEF3 autour du centre des pics MyoD qui décroît rapidement après $\sim$ 500bp vers une répartition uniforme. Ce signal clair nous permet de définir des pics MyoD+MEF3 comme étant les régions de 1kb (signal fort) centrées sur un pic MyoD et contenant au moins un site MEF3 conservé. De cette manière, nous obtenons 1,230 « enhancers putatifs » MyoD+MEF3 (fig. 5.12B).

22. *Reverse Transcription Polymerase Chain Reaction*. Les ARNs sont d'abord transcrits en ADN complémentaire par la transcriptase inverse (RT), puis ils sont amplifiés par réaction en chaîne par polymérase (PCR) et sont quantifiés grâce à une sonde fluorescente se fixant sur l'ADN (technologie Taqman).

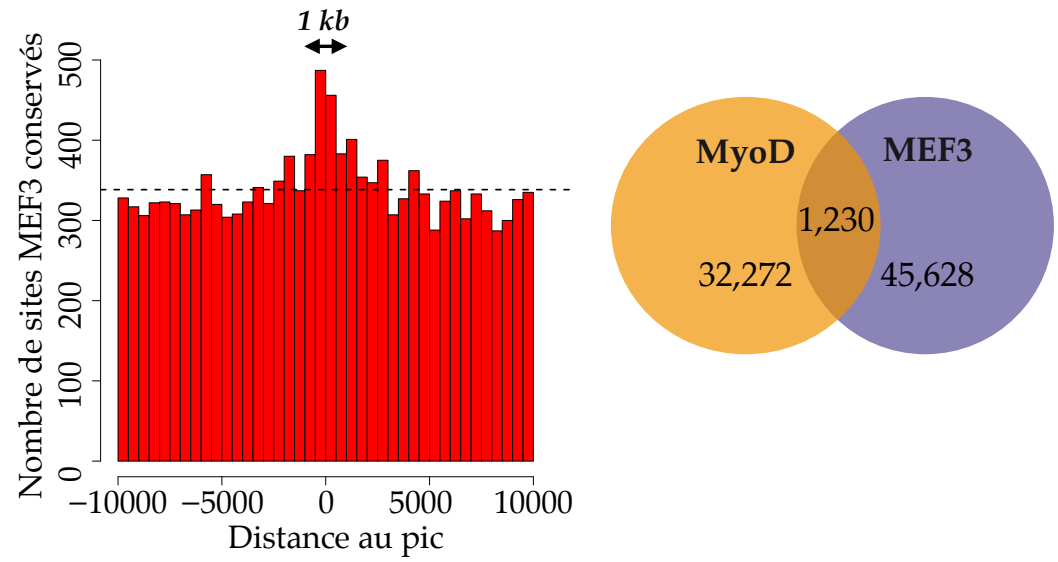


FIGURE 5.12 – **Définition d'enhancers putatifs MyoD+MEF3.**
(A) Corrélation croisée entre les sites MEF3 conservés et les pics ChIP-seq MyoD. Le nombre de sites MEF3 conservés est indiqué en fonction de la distance à un pic MyoD. Une barre représente 500bp. On observe un signal clair dans une région de 1kb autour des pics MyoD. La ligne pointillée indique ce qu'on observerait si la distribution des sites MEF3 était uniforme.
(B) Diagramme de Venn indiquant le nombre de pics ChIP-seq de MyoD contenant au moins un site MEF3.

5.4.4 Croisement des données d'expression et des enhancers putatifs

Nous souhaitons maintenant comparer les données d'expression avec les données bioinformatiques pour voir si la synergie peut s'expliquer par une régulation transcriptionnelle directe. Tout d'abord, nous avons voulu savoir si les enhancers putatifs MyoD+MEF3 avaient un effet sur l'expression des gènes proximaux. Pour chaque région MyoD+MEF3, nous avons regardé le changement d'expression entre la condition Six-/MyoD- et les conditions Six+/MyoD+, Six+/MyoD- et Six-/MyoD+ pour les 10 gènes dont les TSSs sont les plus proches en terme de distance absolue au centre de l'enhancer putatif. Le résultat est montré en figure 5.13A. Alors que dans les cas Six-/MyoD+ et Six+/MyoD- les gènes ne montrent en moyenne pas de changement d'expression, dans le cas synergique Six+/MyoD+ on observe un accroissement d'expression moyen d'un facteur 1.5 pour le gène le plus proche, diminuant rapidement sur les gènes suivants vers la moyenne génomique (pointillés). Nous avons donc décidé d'associer les enhancers putatifs MyoD+MEF3 au gène dont le TSS est le plus proche, un gène pouvant donc avoir plusieurs enhancers putatifs associés. Nous montrons en figure 5.13B l'enrichissement des puces affymetrix en gènes associés à un enhancer putatif MyoD+MEF3 lorsque le taux d'accroissement (Fold-Change ou FC) est supérieur ou égal à une valeur donnée. Cet enrichissement est défini par le rapport entre la proportion de gènes associés à un enhancer putatif au-delà d'un seuil de FC et la proportion moyenne dans la puce affymetrix (seuil FC= 0). Au-delà de FC= 1.5, l'enrichissement passe rapidement de 1 (pro-

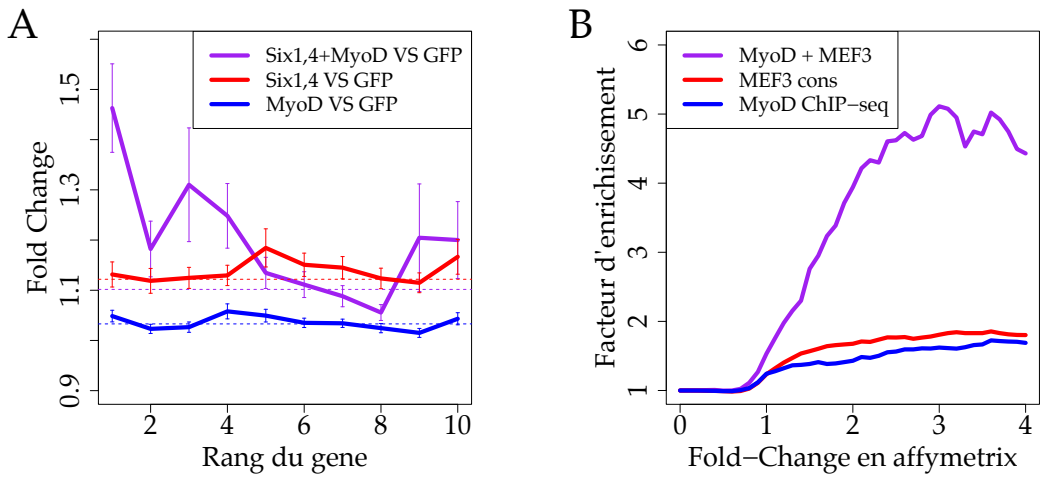


FIGURE 5.13 – **Rôle des enhancers putatifs MyoD+MEF3 dans la synergie Six+MyoD.** (A) Changement d'expression moyen des gènes entourant les enhancers putatifs Six+MyoD en condition de différenciation. Les gènes sont classés en fonction de leur distance absolue au centre de l'enhancer. On observe une expression synergique du gène le plus proche (accroissement d'expression dans le cas Six1,4+MyoD mais pas Six1,4 ou MyoD seuls) diminuant rapidement pour les gènes suivants. Les barres d'erreur indiquent le SEM.
(B) Proportion de gènes associés à un ChIP-seq MyoD, un site MEF3 conservé, ou un enhancer putatif MyoD+MEF3 au sein des gènes ayant un Fold-Change en puce affymetrix supérieur ou égal à une valeur donnée en condition de différenciation. Cette proportion est normalisée par rapport à sa valeur moyenne sur la puce entière (point FC= 0, proportions moyennes de 51% pour MyoD, 41% pour MEF3, et 4.5% pour MyoD+MEF3). Ceci définit l'enrichissement et permet de comparer les courbes. Les gènes ayant un Fold-Change supérieur à 1.5 sont clairement enrichis en enhancers MyoD+MEF3.

portion moyenne) à 5. En comparaison, nous montrons les enrichissements obtenus lorsque l'on utilise seulement les sites MEF3 ou les ChIP-seq MyoD lors de l'association au gène le plus proche. Les enrichissements obtenus sont dans les deux cas inférieurs à 2. Ainsi, ces deux résultats montrent qu'il y a une association claire entre la présence d'une région MyoD+MEF3 et l'expression synergique du gène le plus proche.

Au final, nous trouvons 82 gènes synergiques associés à 96 enhancers putatifs MyoD+MEF3, partagés en 9 gènes prolif only, 8 prolif & diff, et 65 diff only (fig. 5.14). Nous en avons sélectionné plusieurs afin de vérifier les données de puce par qPCR. Les gènes indiqués en rouge dans la figure sont ceux dont les données qPCR ont reproduit les données de puce. Nous avons par ailleurs testé un certain nombre de gènes dont les données qPCR étaient soit négatives (pas d'expression) soit non synergiques (Six-only ou MyoD-only) : *Bhlhe*41, *Chd*7, *Cxcr*4, *E2f*8, *Grem*1, *Mef2a*, *Meis*1, *Nfib*, *Sestd*1, *Srgap*3, *Trps*1, et *Zfp*238. Au final, le taux de validation est de 54% . Il est à noter que l'enhancer MyoD+MEF3 de Myog, situé au niveau de son promoteur, n'est pas détecté par notre approche car il n'y a pas assez d'espèces présentes dans l'alignement pour considérer le site MEF3 comme conservé selon les critères d'Imogene (bien que celui-ci soit strictement conservé entre la souris et les primates). Il a donc été rajouté aux données *a posteriori*. Il est aussi à noter que certains gènes ne figurent pas dans cette liste bien

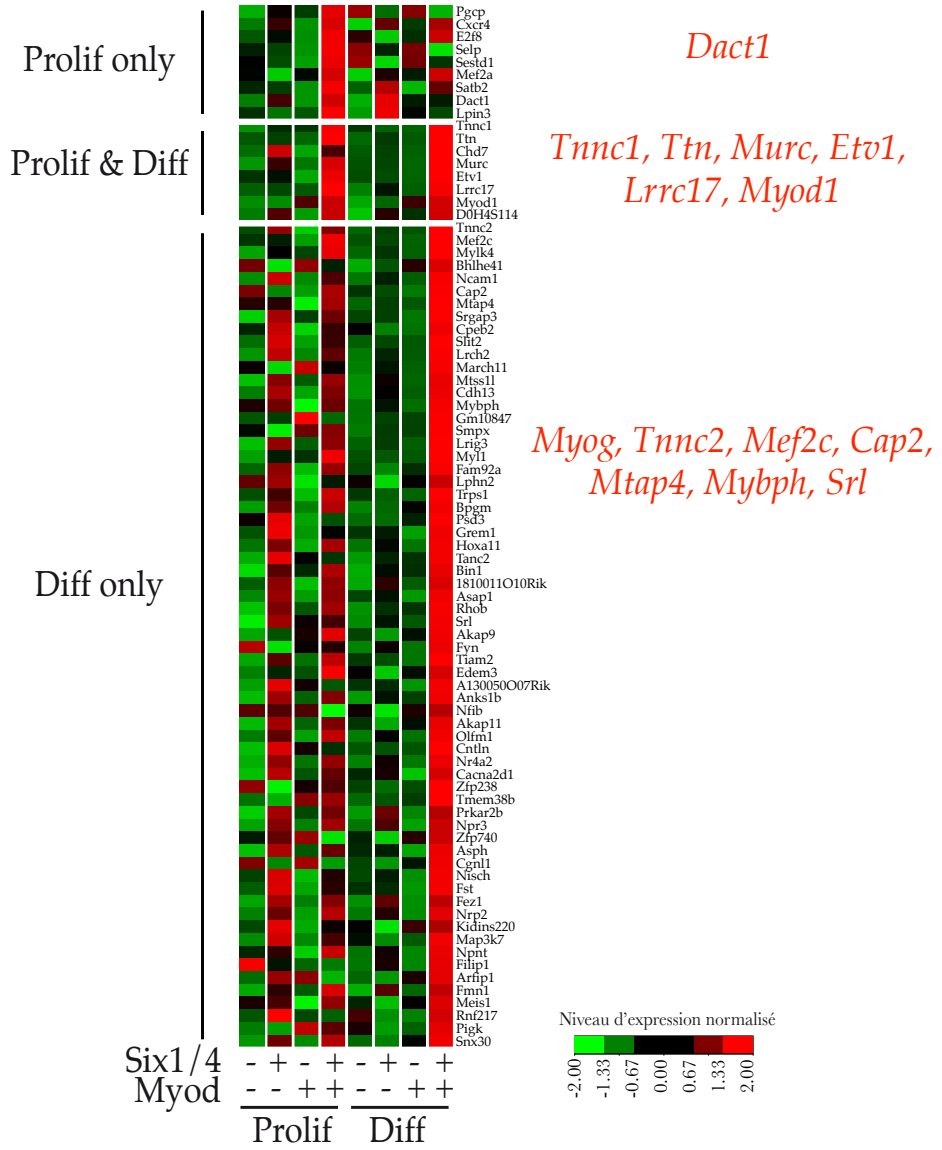


FIGURE 5.14 – **Liste finale de gènes synergiques associés à un enhancer MyoD+MEF3.** Liste des gènes ayant une expression synergique (fig. 5.11) et qui sont associés à un ou plusieurs enhancers MyoD+MEF3 (ceux-ci étant attribués au gène dont le TSS est le plus proche). Nous indiquons en rouge les gènes dont nous avons testé le ou les élément(s) régulateur(s) MyoD+MEF3.

qu'ils aient un comportement synergique, tel que *Myh*3, *Actc*1, *Tnnt*3, *Tnnt*1, *Casq*2, ou encore *Acta*1. Les raisons sont variées, par exemple le site MEF3 peut être situé juste au-delà du seuil de 500bp du centre du pic MyoD (*Actn*3), l'enhancer peut être associé à un autre gène (LCR du locus myosines), ou encore comme pour Myog le site MEF3 n'est pas considéré comme conservé (*Tnnt*3 ou *Acta*1).

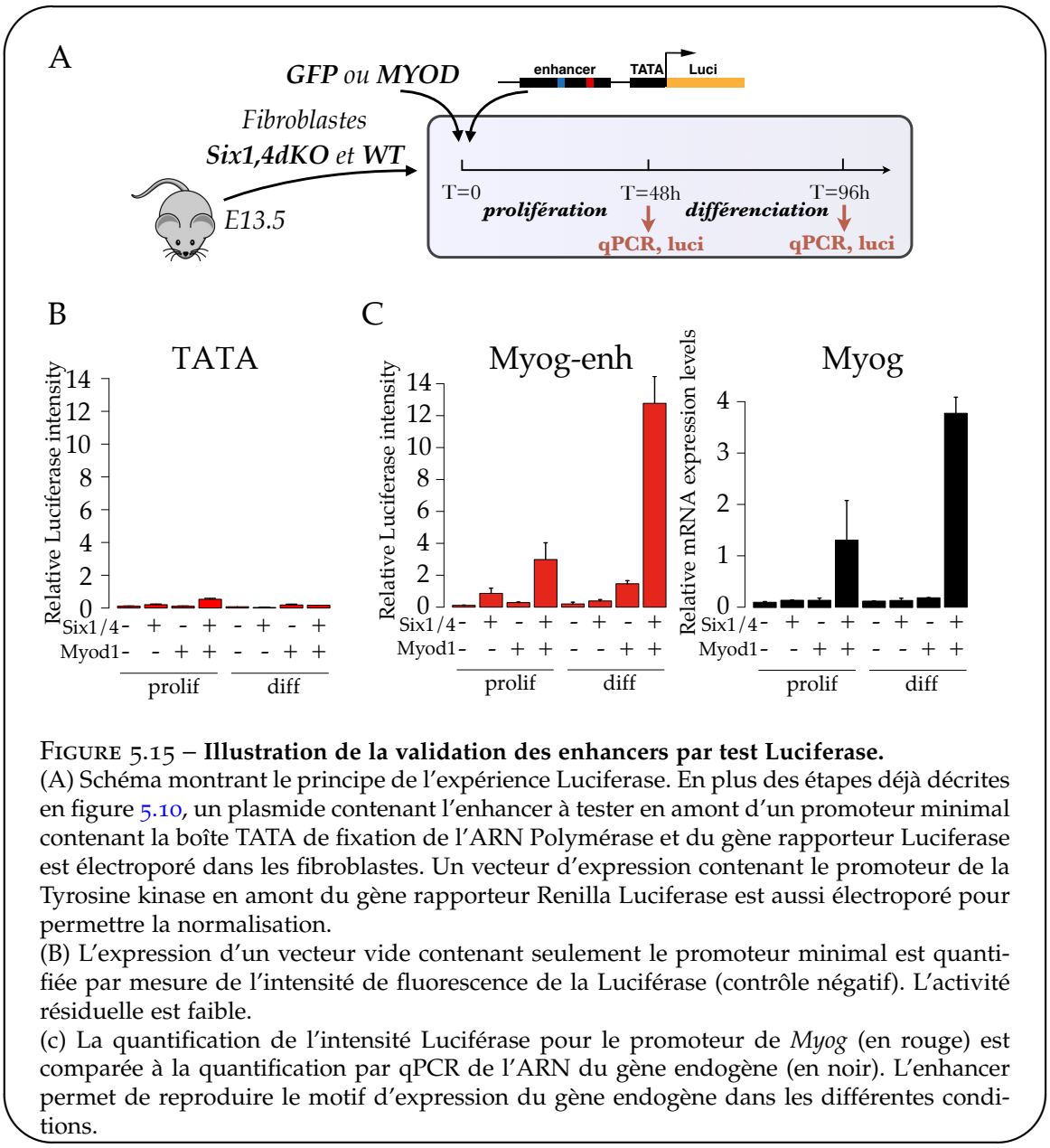


FIGURE 5.15 – **Illustration de la validation des enhancers par test Luciferase.**
(A) Schéma montrant le principe de l'expérience Luciferase. En plus des étapes déjà décrites en figure 5.10, un plasmide contenant l'enhancer à tester en amont d'un promoteur minimal contenant la boîte TATA de fixation de l'ARN Polymérase et du gène rapporteur Luciferase est électroporé dans les fibroblastes. Un vecteur d'expression contenant le promoteur de la Tyrosine kinase en amont du gène rapporteur Renilla Luciferase est aussi électroporé pour permettre la normalisation.
(B) L'expression d'un vecteur vide contenant seulement le promoteur minimal est quantifiée par mesure de l'intensité de fluorescence de la Luciférase (contrôle négatif). L'activité résiduelle est faible.
(c) La quantification de l'intensité Luciférase pour le promoteur de *Myog* (en rouge) est comparée à la quantification par qPCR de l'ARN du gène endogène (en noir). L'enhancer permet de reproduire le motif d'expression du gène endogène dans les différentes conditions.

5.4.5 Validation des enhancers MyoD+MEF3

Nous souhaitons maintenant tester expérimentalement l'hypothèse que les enhancers putatifs associés aux gènes synergiques permettent de récapituler le motif d'expression du gène. Nous avons pour cela 14 gènes synergiques dont l'expression a été confirmée par qPCR et qui sont associés à 21 enhancers putatifs. Afin de tester l'activité de ces enhancers, nous avons mis en place une expérience de mesure de l'intensité d'un gène rapporteur Luciferase mis sous le contrôle d'un enhancer. Le protocole est décrit en figure 5.15A.

Lors de la transfection des fibroblastes par le vecteur exprimant MyoD ou GFP, un autre vecteur est électroporé, contenant l'enhancer à tester en amont d'un promoteur minimal contenant la boîte TATA de fixation de l'ARN polymerase et contrôlant l'expression du gène rapporteur Luciferase. Pour des raisons de normalisation, un autre plasmide est aussi électroporé,

contenant le promoteur de la Tyrosine kinase – qui n'est pas lié au développement musculaire – en amont du rapporteur Renilla Luciferase. D'abord, l'expression du vecteur vide (TATA seule) a été mesurée afin d'estimer le biais introduit par ce promoteur minimal (fig. 5.15B). L'intensité Luciferase relative mesurée est au plus de 0.5 (prolif Six+/MyoD+). Cette expression est à comparer avec celle du vecteur contenant le promoteur de Myog affichant une intensité maximale plus de 20 fois plus grande, et qui nous sert de contrôle positif (fig. 5.15C). Dans ce cas, l'expression du rapporteur reproduit le motif d'expression du gène endogène obtenu par qPCR (barres noires).

Nous présentons les résultats « positifs » de l'expérience Luciferase en figure 5.16 et les résultats « négatifs » en figure 5.17. Par positif et négatif, nous entendons le fait que le gène rapporteur reproduit le motif d'expression du gène endogène. Ainsi, les enhancers putatifs MyoD+MEF3 « négatifs » peuvent avoir une expression forte (voir par exemple les deux enhancers « Myod-only ») mais ne reproduisant pas la synergie observée sur le gène associé. Au total, 70% des prédictions sont validées (hors Myog, utilisé comme contrôle positif). Nous notons dans certains cas que plusieurs enhancers associés à un même gène récapitulent un motif d'expression similaire, comme par exemple les 3 enhancers positifs de *Mef2c*. Ce résultat est reminiscent des enhancers fantômes ou *shadow enhancers* introduits par M. Levine pour décrire les enhancers redondants chez la Drosophile (voir introduction, section 1.5.4).

5.4.6 Recherche de motifs et mutagenèses

- **Recherche de motifs**

Maintenant que nous avons obtenu 15 régions MyoD+MEF3 validées, nous pouvons appliquer l'algorithme Imogene pour détecter les motifs enrichis. Pour cela, nous avons créé une version modifiée de l'algorithme utilisant les bases de données de TFs Transfac et Jaspar (voir section 1.7.2) pour guider la recherche. L'algorithme est d'abord utilisé pour générer des motifs *de novo* à partir des séquences. Ensuite, les PWMs de Transfac et Jaspar sont raffinées sur les séquences d'apprentissage. Pour cela, nous utilisons simplement ces PWMs connues comme *Prior* lors de l'initialisation du motif dans Imogene. Une fois tous les motifs générés et raffinés, ceux-ci sont classés en fonction de leur enrichissement dans l'ensemble d'apprentissage par rapport à des séquences *background* de tailles similaires issues des régions intergéniques. L'enrichissement est défini comme étant la proportion de séquences de l'ensemble d'apprentissage prédites, c'est-à-dire possédant au moins un site pour le motif considéré, pour un taux de séquences *background* prédites fixé à 1% (FPR = 0.01). Ce critère permet de définir un seuil stringent adapté à chaque motif évalué.

Les résultats sont montrés en figure 5.18. Les motifs sont classés par enrichissement décroissant. Les deux premiers motifs correspondent à la boîte E sur laquelle se fixe MyoD et au site MEF3 sur lequel se fixent les protéines Six, ce qui est attendu. À noter que le seuil utilisé ici est plus stringent que celui utilisé lors de la détection des sites MEF3 sur le génome : on ne retrouve donc pas de sites MEF3 sur tous les enhancers. Parmi les autres motifs détectés, on retrouve Meis1, Ap1 et Mef2, motifs qui avaient déjà été trouvés par Cao et al. (2010) comme étant surreprésentés au sein des pics ChIP-seq de MyoD en MT (Meis et Ap1) et dans les pics augmentant d'intensité entre MB et MT (Mef2). On trouve aussi le motif associé aux protéines EBF, qui ont récemment été trouvées comme étant impliquées dans la régulation du développement musculaire chez l'amphibien *Xenopus laevis* (Green and Vetter, 2011). Par ailleurs, plusieurs motifs *de novo* figurent dans la liste. Il est à noter qu'après avoir réalisé un premier classement de l'ensemble des motifs générés, nous fusionnons les motifs du bas du

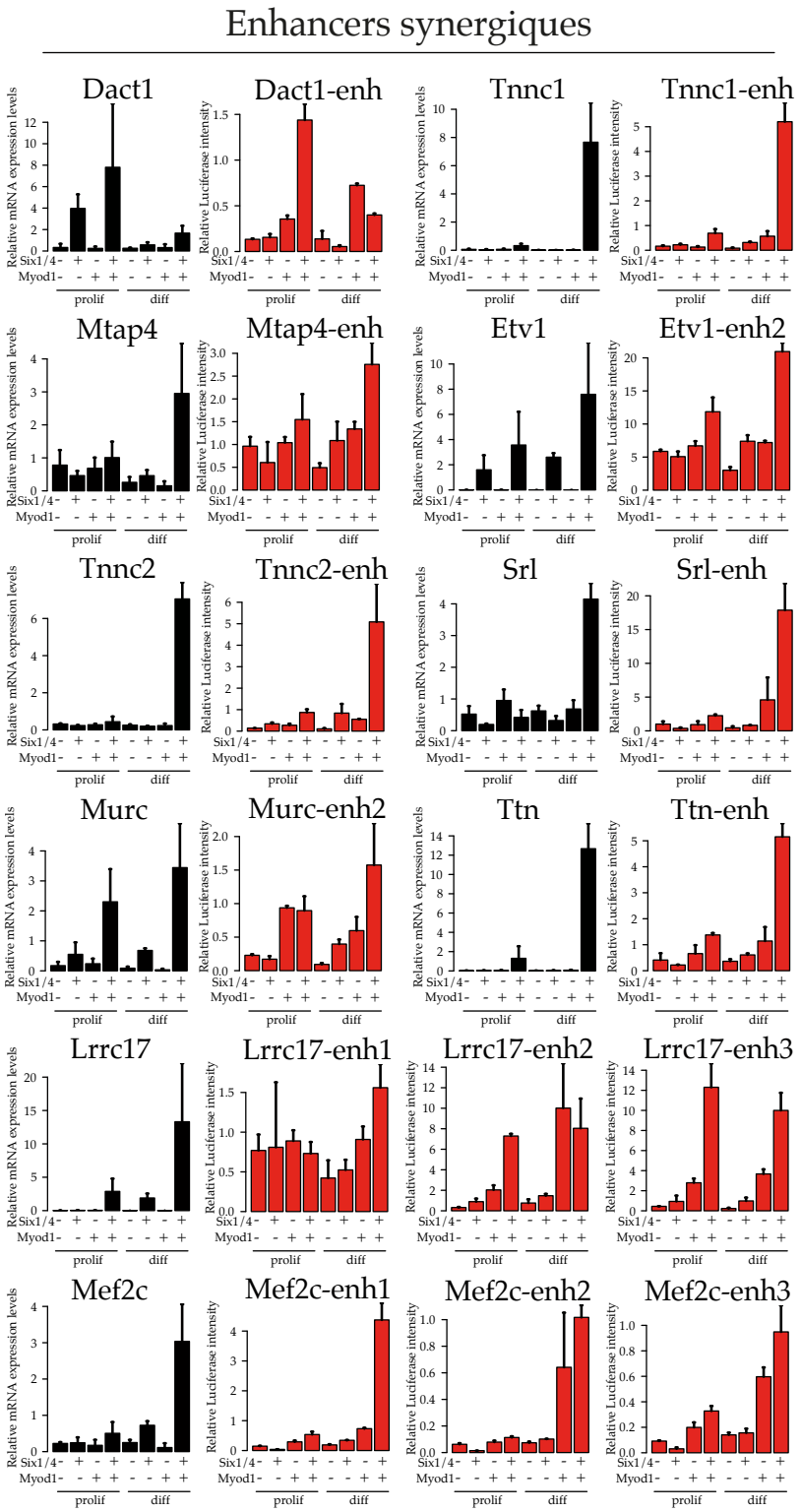


FIGURE 5.16 – **Enhancers MyoD+MEF3 « positifs ».**
Les données qPCR pour différents gènes testés (barres noires) sont comparées aux résultats Luciferase des enhancers prédits correspondants (barres rouges). Ces enhancers « positifs » reproduisent le motif d'expression du gène endogène.

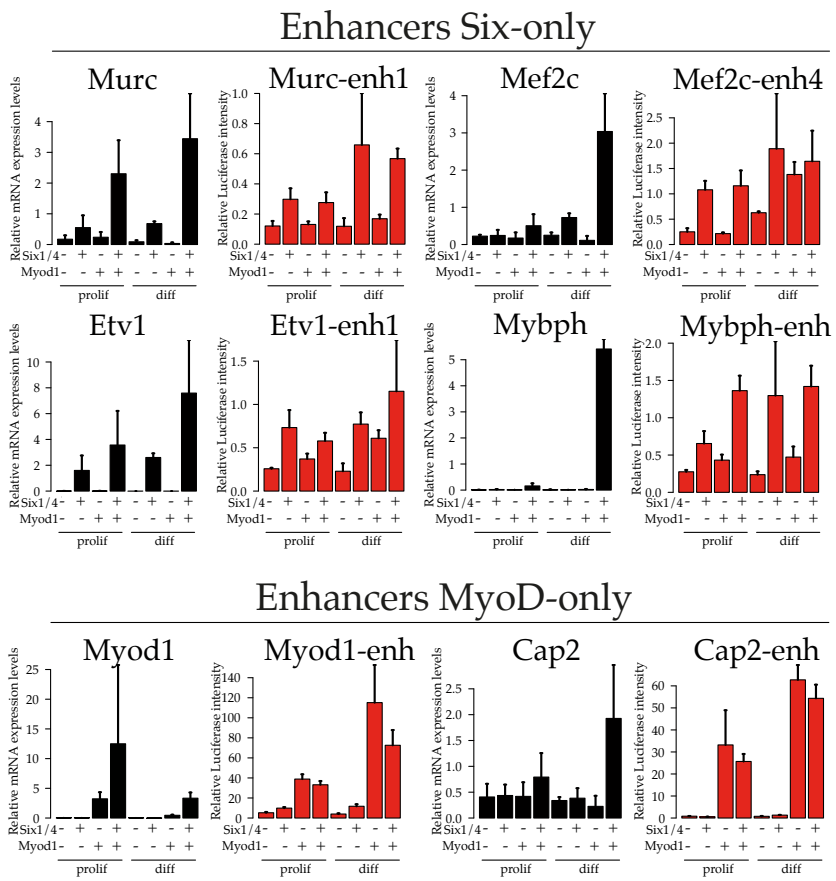


FIGURE 5.17 – **Enhancers MyoD+MEF3 « négatifs ».**
Les enhancers ne récapitulant pas la synergie observée au niveau du gène endogène montrent un comportement lié à Six seul (Six-only) ou à MyoD seul (Myod-only). Contrairement à la figure 5.11 montrant la quantité totale (endogène + transfecté) d'ARN de *Myod*1 mesurée par qPCR, nous montrons ici seulement la quantité relative au gène endogène, obtenue en utilisant un primer ciblant la région 5′ UTR de l'ARN endogène qui n'est pas présente dans l'ARN exprimé par le *Myod*1 transfecté.

classement avec les motifs mieux classés lorsque leur distance est inférieure à un seuil, ou de manière équivalente lorsqu'il partagent suffisamment de sites en commun au-delà d'un seuil de détection donné, dans ce cas 10 bits (voir article décrivant Imogene en 3.2). Ainsi, un motif *de novo* peut être une meilleure description (en ce qui concerne le cas d'étude) d'un motif Transfac existant. Ce seuil de fusion étant arbitraire, il peut y avoir certains doublons, par exemple nous retrouvons une deuxième E-Box en 11ème position.

- **Mutagenèse systématique des sites de fixation**

Afin de tester l'importance des motifs générés, nous avons entrepris des expériences de mutagenèse systématique sur 3 enhancers validés : Srl-enh, Tnnc2-enh et Lrrc17-enh3. Ces travaux sont en cours et nous ne sommes pour le moment qu'en mesure de présenter les résultats pour Srl-enh. La séquence de Srl-enh est montrée en 5.19A, ainsi que les sites prédits pour les différents motifs (lorsqu'il y en a). L'expérience de mutagenèse systématique consiste à mu-

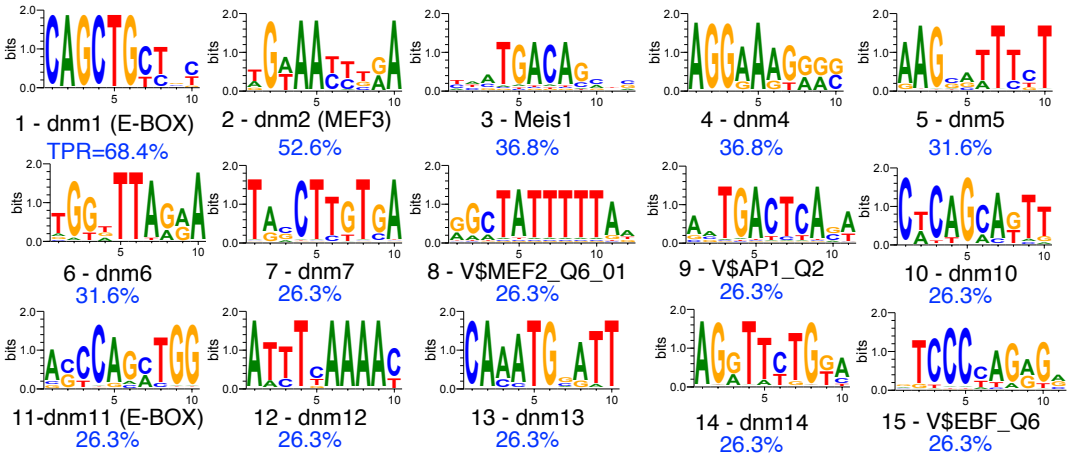


FIGURE 5.18 – **Motifs sur-représentés sur les enhancers synergiques.**
Les logos des 15 premiers motifs sont montrés, ainsi que la proportion des séquences d'apprentissage détectées pour FPR= 1%. Les motifs générés *de novo* par Imogene sont signalés par le préfixe *dnm* (*de novo* motif).

ter l'ensemble des sites d'un motif donné puis de tester l'expression de l'enhancer muté par Luciférase. Pour chaque site, nous remplaçons les 6 nucléotides consécutifs les plus conservés dans la PWM par le site de restriction GAATTC reconnu par l'enzyme EcoRI, sauf dans le cas où le (les) site(s) est (sont) en extrémité de séquence, auquel cas nous procédons par délétion, comme par exemple pour le site MEF3 en début de séquence et les trois sites EBF en fin de séquence de Srl-enh. La substitution par un site de restriction permet de rapidement tester que les mutations ont été introduites au bon endroit : il suffit de regarder par retard sur gel la taille des fragments d'ADN obtenus après digestion par EcoRI. Les résultats Luciferase des mutagenèses de Srl-enh sont montrés en figure 5.19. Ces résultats sont inattendus : on observe peu d'effet dans tous les cas, voire un effet d'activation dans certains cas (Srl:ebox, Srl:mot5, Srl:mot14). De tels résultats pointent vers une robustesse *in vitro* de l'élément régulateur de Srl aux mutations. Comment l'interpréter ?

- **Interprétation des résultats**

L'hypothèse d'une activité résiduelle du vecteur muté dû au promoteur minimal TATA a déjà été écartée en figure 5.15B. La robustesse de l'expression pourrait par ailleurs être due à la surabondance de MyoD dans le milieu, ce qui pourrait activer le rapporteur même avec une fixation faible. Pour tester cette hypothèse, nous avons réalisé un étalonnage de la quantité de MyoD injectée lors de l'électroporation : 500ng (valeur par défaut), 50ng et 5ng (fig. 5.20). Nous espérions ainsi voir si une quantité moindre de MyoD, par exemple 50ng, permettait d'observer une synergie pour l'enhancer WT mais pas pour l'enhancer muté. Le résultat de l'expérience ne conforte pas cette hypothèse : même en diminuant la dose de MyoD, l'enhancer muté garde un comportement similaire à l'enhancer non muté. Par ailleurs, la valeur de 500ng utilisée est adaptée à l'étude, comme on peut le voir sur le contrôle positif du promoteur de Myog.

Une autre hypothèse est qu'il y a des mécanismes de fixation protéine-protéine compensant l'impossibilité de se fixer à l'ADN. Nous avons donc cherché à savoir quelle était la

A

SRL: MEF3, EBOX, EBF, MEIS1, MOT5, MOT14, AP1, MEF2

GGGGAAATTCAACCACTAGCAAAGACTGCTGTTCTAGGGGTATGGGGAAAGCAT
CAGGTGTCTTGTAACCTCCCACTGAGTTGAAGCAGCAGCCACCTCGGGGTACACA
ATCTACAACTGGGCAGAGTAGTGGTAACTGAAGTAGAACAAGCAGGCAGGGGGA
TTTGAGACAGGCAAGAGCAAGCTCACACTCCTGGGTACAAAAAGAAACAGCCTA
GAAAAACTTTGATATGAATCACATGCCTCTTTCCCATAGCAAGAGCTCTGAACAT
AGCCTTCAGAATAGTCTAGAAATCCTGTCCCCTTTATACAAAGAAAGAAAAAAA
ATTTAAATGTCCCTAAAAGGACAGAAATTCCTGGGGCTTTGCTGTGTGGTAGCTT
CAAGGTGGCAGCTGGGGCTGGAGCAGCAGAGGAGGGGCTGGGAAGAGGAGAGG
GTGGCAGTGGCTTCGGGCTCAGGATCCAATTACAGGACTGTTTCCATACTTAGCC
ACCTGACAGCCGCTTGGAATACTCACCACTCTAGCCACCACAGCCCCATGGAGAC
CAAATTAACTCAAGGGCAGACAAGGGCAGGGCCCAACAGCAAGCCCAATGGCCC
AGAGCACCCAGCCTTACCTCTCACACCCCCAAAGAATGGCTTTTGTTCCTCATAT
GCCTATCCTCTCACTGGACTGCACCTCTGTCACTCCAGTCCCTGAGACAGAACCTT
AAAGGTCTAAAACAAGGACCCAGAAATCTCACAAAGATGAAGTACCAGCATAAA
AGACAGACCAGGGTGTCCTAGATCCCTAGGGGGAATCTTAGAGCACCTCCCCTAA
GAATCCCCAAGGGGAAGCCAGTCTTGGGAGTC

B

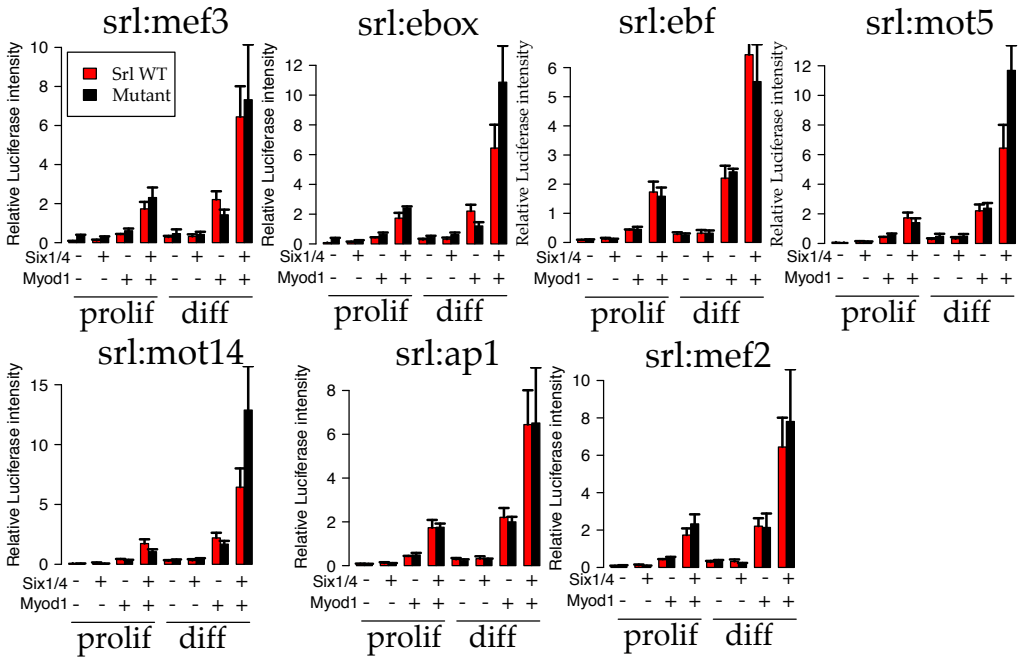


FIGURE 5.19 – **Mutagenèse systématique des sites prédits sur l'enhancer de *Srl*.**
(A) Les différents sites prédits par les motifs de la figure 5.18 sont indiqués en couleur sur la séquence d'ADN de l'enhancer de *Srl*.
(B) Les différents graphiques à barres comparent l'intensité Luciférase de l'enhancer Srl non muté (rouge) à celle de l'enhancer muté dans les sites du motif indiqué (noir).

réponse attendue pour les sites de fixation de Six1, MyoD et Mef2c seuls : peuvent-ils exhiber une synergie ? Tout d'abord, dans le cas du polyMEF3, l'activité de Six1 est perceptible mais faible (fig. 5.21A). Ainsi, les protéines Six1 seules, dans ce contexte, ne sont pas de bonnes activatrices (ce qui est attendu dans le cas d'une synergie). On note par ailleurs que l'activité du polyMEF3 est semblable à celle observée pour les enhancers Six-only de la figure 5.17, suggérant que la régulation de ces enhancers passe bien par les sites MEF3 prédits. Dans le

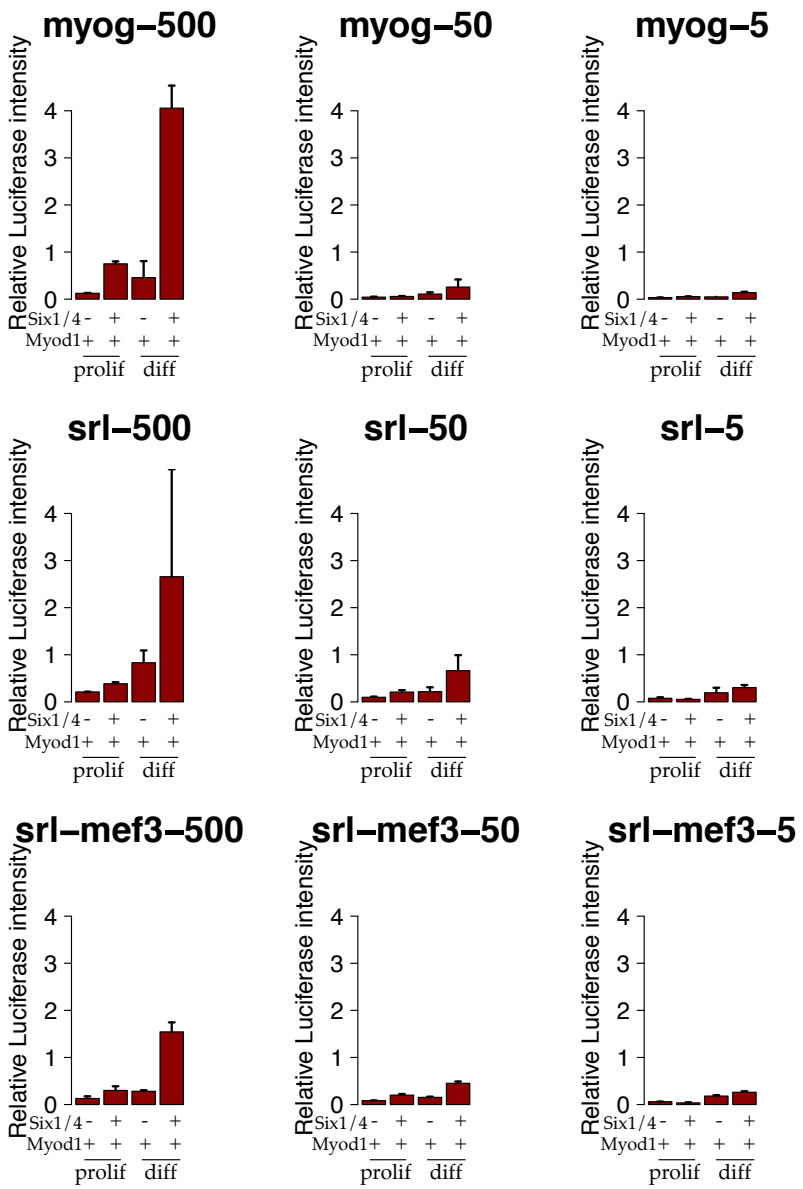


FIGURE 5.20 – **Variation de la quantité de vecteur MyoD électroporé.**
Expériences Luciférase pour Myog, TATA, Srl, et Srl:MEF3 pour plusieurs valeurs de la quantité de MyoD électroporé : 500 ng, 50 ng ou 5 ng. Sont montrées les conditions Six-/MyoD+ et Six+/MyoD+ en prolif et en diff. La valeur utilisée dans les expériences est 500ng.

cas du polyE, on n'observe pas d'expression en prolif alors que l'ARN de *Myod*1 est présent. Par contre, on observe une forte expression en diff lorsque Six1 et MyoD sont présents. Cet effet peut-être dû à la présence de Myog qui peut aussi se fixer sur les boîtes E, ou encore à la levée d'un processus inhibant la fixation de MyoD sur ses cibles (cf par exemple le cas des protéines ID, Yokoyama et al. (2009)). De la même manière, dans le cas du polyMEF2, on observe une expression synergique seulement en diff qui semble récapituler celle du gène

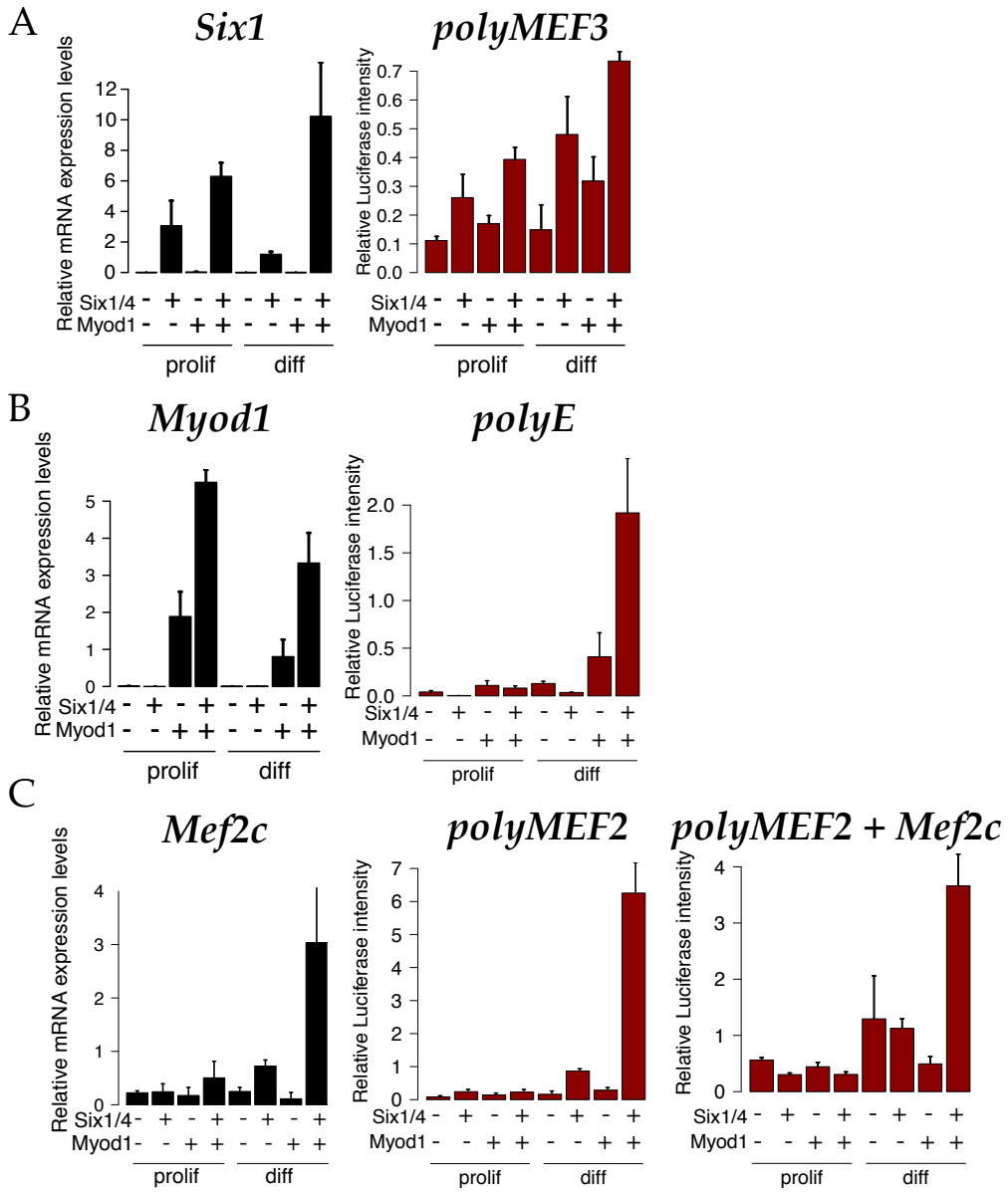


FIGURE 5.21 – **Expression Luciférase de multimères synthétiques de sites de fixation.** Diverses expériences de Luciférase sont montrées pour des enhancers synthétiques contenant des multimères de sites de fixation, ainsi que l'expression du TF correspondant. (A) Séquence polyMEF3 contenant 6 sites MEF3. (B) Séquence polyE contenant 4 E-Box. (C) Séquence polyMEF2. Dans ce dernier cas, l'un des deux multimères a été co-transfecté avec un vecteur d'expression *Mef2c*

Mef2c endogène. Néanmoins, lors de la co-transfection dans le milieu d'un vecteur exprimant Mef2c, l'expression dans les autres conditions n'est pas significativement augmentée. La protéine MEF2 semble donc n'avoir d'activité que dans le cas Six+/MyoD+ en milieu de différenciation, ce qui peut être lié à son activation par phosphorylation (Ferrari et al., 1997). Ces données semblent pointer vers une forte activité de Mef2 et MyoD dans le cas Six+/MyoD+ en

différenciation, qui pourrait effectivement créer une compensation lors de la mutation d'un des deux sites. Pour tester cette hypothèse, il conviendrait dont maintenant de réaliser des doubles mutations de type E-Box/MEF2, E-Box/MEF3, ou MEF3/MEF2.

Enfin, dans le cas de l'enhancer testé par mutagenèse, il est possible que sa longueur (1kb) soit trop importante (densité de sites trop grande). Une possibilité serait de le couper en deux pour réduire sa taille et produire une région minimale montrant une perte d'expression lors de la mutation des sites correspondant à un motif.

5.5 Conclusion et perspectives du chapitre 5

Dans ce chapitre, nous avons présenté diverses analyses bioinformatiques appliquées au cas de la différenciation musculaire. Nous avons montré comment, en combinant des données expérimentales publiées et des outils bioinformatiques simples (PWM), il était possible de réaliser des prédictions validées *in vivo*, que ce soit dans le cas de la régulation de *Myod*1 ou celle d'un lincRNA et des myosines de type rapide. Par ailleurs, nous avons présenté un système modèle de différenciation musculaire montrant que la présence des homéoprotéines Six1 est nécessaire à la transdifférenciation par MyoD de fibroblastes en cellules musculaires : on parle de synergie. En combinant la présence de sites MEF3 conservés avec celle d'un pic ChIP-seq pour MyoD, nous avons exhibé des régions de régulation putatives de cette synergie et avons montré que la majorité (70%) reproduit le motif d'expression du gène cible. Enfin, nous avons cherché des motifs surreprésentés sur ces régions de régulation synergiques et avons trouvé plusieurs motifs, dont notamment le co-régulateur connu Mef2, dont nous montrons qu'il est lui-même activé de manière synergique par Six1 et MyoD. La mutation de plusieurs sites prédits a été réalisée, sans effet sur l'expression du gène rapporteur, montrant une robustesse *in vitro* de la région régulatrice aux mutations.

Cette observation semble contredire les résultats obtenus par transgenèse *in vivo* où la simple mutation de sites MEF3 abolit l'expression du rapporteur. Néanmoins, la comparaison est biaisée, puisque dans le cas de la transgenèse, le contexte chromatinien de l'enhancer introduit dans le génome est important, et l'ouverture de la chromatine induite par la fixation sur l'ADN des protéines Six pourrait expliquer la différence de comportement observée entre les cas *in vivo* et *in vitro*. En effet, il a été montré que Six1 interagit directement avec la sous-unité Brg1 du complexe SWI/SNF de remodelage de la chromatine, et que cette interaction permet la différenciation de cellules épithéliales de l'oreille interne en cellules neuronales (Ahmed et al., 2012). Par ailleurs, il a plus récemment été montré que Six1 contrôle l'accessibilité du *Core Enhancer* CE de *Myod1* en C2C12 : en l'absence de Six1, les nucléosomes y sont plus abondants et MyoD ne peut plus s'y fixer (Liu et al., 2013). Ces résultats laissent à penser que le rôle de Six1 aux différents enhancers putatifs que nous avons étudiés serait d'ouvrir la chromatine, rendant l'accession à l'élément de régulation possible. L'élément serait par la suite fixé par un complexe de TFs dont certains (MyoD/Myog, Mef2c) sont eux mêmes activés de manière synergique. Enfin, un tel complexe montre une capacité de compensation rendant les enhancers robustes aux mutations.

Plusieurs hypothèses restent finalement à être testées pour mieux cerner la synergie observée entre Six1 et MyoD au cours de la myogenèse. D'abord, l'idée que Six1 pourrait agir comme un élément crucial de l'ouverture de chromatine au niveau des enhancers validés peut être testée par quantification de l'accessibilité de l'ADN par DNAse1, ainsi qu'en quantifiant les niveaux de méthylation H3K4me1 des histones par ChIP. Par ailleurs, la fixation des TFs sur les sites prédits reste aussi à confirmer, par exemple par retard sur gel avec des extraits cel-

lulaires des fibroblastes utilisés. Enfin, des doubles et triples mutations des sites MEF3, E-box et MEF2 permettraient de clarifier les effets de compensation au sein des enhancers testés.

Annexe A

Statistiques génomiques

Nous nous intéressons ici aux distributions de tailles intergéniques et introniques dans les génomes de différentes espèces, et aux questions qu'elles soulèvent.

Nous l'avons vu en introduction (section 1.7.1), l'interface Galaxy permet d'obtenir un certain nombre d'annotations génomiques à partir de différentes bases de données, comme UCSC. Les annotations génétiques consistent généralement en des coordonées sur le génome de TSSs et de leurs exons et introns. Ceux-ci sont associés à un gène, qui peut avoir plusieurs TSSs différents et dont le transcrit peut avoir plusieurs épissages alternatifs.

Nous avons utilisé Galaxy pour récupérer les annotations génomiques de différentes espèces, allant de l'unicellulaire à l'homme : la bactérie *Escherichia coli*, la levure *Saccharomyces cerevisiae*, le ver *Caenorhabditis elegans*, la mouche *Drosophila melanogaster* , la souris *Mus Musculus*, le poulet *Gallus gallus* et l'homme *Homo Sapiens*.

Les données intergéniques ont été obtenues en prenant les coordonnées complémentaires à l'ensemble fusionné des transcrits annotés. Dit autrement, chaque gène a été défini par les coordonnées les plus extrêmes de ses transcrits alternatifs, et les régions entre deux gènes définissent les régions intergéniques. La distribution de la taille de l'intergénique pour ces différentes espèces est montrée en figure A.1A. On observe une loi proche d'une log-normale et qui semble, à une remise à l'échelle près, être relativement conservée chez les différentes espèces. Les distributions sont essentiellement unimodales, mais on note l'apparition d'un deuxième pic à $\sim$ 100bp chez la souris et l'humain, peut-être dû à l'annotation récente de nombreuses régions transcrites non codantes proches des gènes. On note par ailleurs l'inflation importante de la taille des régions intergéniques entre la bactérie (180 bp en moyenne, médiane à 100bp) et l'homme (80kb de moyenne, médiane à 15kb).

Les données introniques montrées en figure A.1B exhibent quant à elles un comportement très stéréotypé. Toutes les distributions présentent un pic fort autour de 80bp, que l'on peut interpréter comme une taille minimale pour que l'épissage puisse avoir lieu, et l'apparition d'un deuxième pic à 1kb chez les vertébrés. On observe par ailleurs que la présence d'introns longs semble corrélée à la taille des génomes.

La présence de distributions des longueurs d'intergénique et d'intronique stéréotypées au sein d'organismes aussi divers que la bactérie, le ver ou l'homme laisse à penser qu'il existe des mécanismes universels régissant la croissance des régions non codantes du génome. Il serait donc intéressant de comprendre quels mécanismes d'insertions-délétions (indels) permettent de modéliser ce phénomène. De manière générale, l'équation maîtresse caractérisant l'évolution de la distribution $P(L,t)$ des longueurs intergéniques L au cours du temps t s'écrit

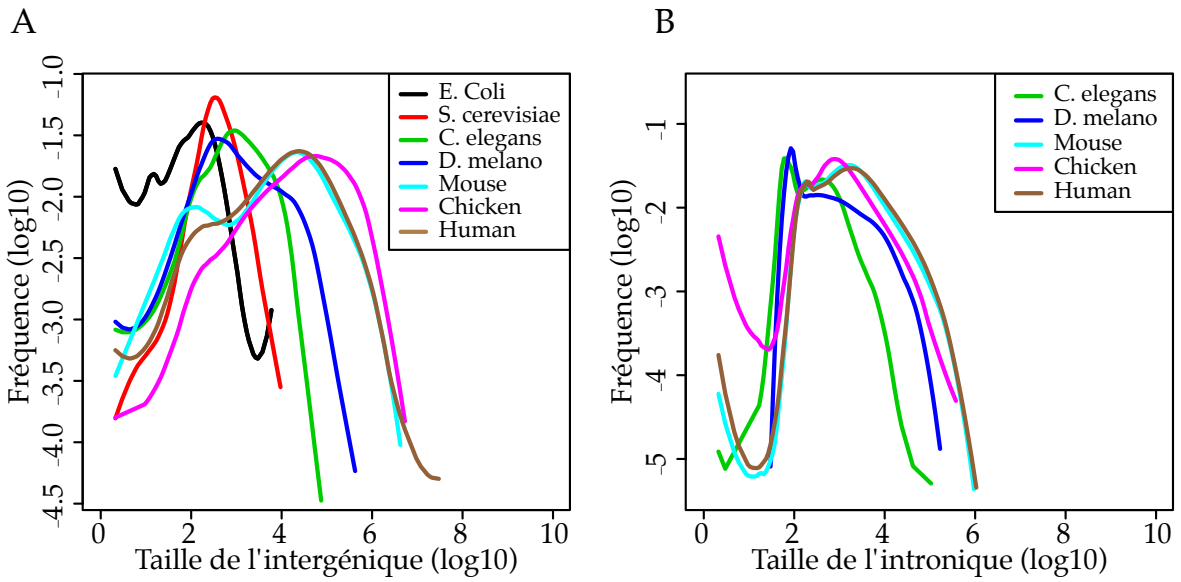


FIGURE A.1 – **Distribution des tailles intergéniques et introniques chez différentes espèces.** Distributions log-log de la taille des régions intergéniques (A) et introniques (B) chez différentes espèces. Les histogrammes sont réalisés avec un intervalle de 0.05 (en log 10) puis lissés avec l'estimateur local LOESS de paramètre $span = 0.3$ (logiciel R). (A) Les régions intergéniques sont définies comme les régions complémentaires aux régions transcrites (données UCSC), celles-ci étant préalablement fusionnées pour éviter les redondances liées aux multiples transcrits d'un même gène. De la bactérie à l'homme, on observe une inflation de la quantité de génome non codant. (B) Les régions introniques sont définies par le fait qu'elles sont entourées par deux exons d'un même gène. Pour pouvoir être épissés lors de la maturation des preARNm, les introns doivent posséder des sites d'épissage, imposant une borne inférieure à leur taille pour que l'ARNm final soit fonctionnel.

$$P(L, t + \mathrm{d}t) = P(L, t) + \int_0^\infty \mathrm{d}L' \Big[P(L', t)\tau(L' \to L) - P(L, t)\tau(L \to L') \Big] \mathrm{d}t \tag{A.1}$$

où $\tau(L \to L')$ est le taux de transition de la taille L' vers la taille L, qui dépend des insertions-délétions. On imagine un mécanisme similaire pour les introns, avec une longueur minimale $L_{\min}$. De nombreuses données sur les statistiques des indels ont récemment été rendues accessibles chez l'homme dans le cadre du projet 1000 genomes (Mills et al., 2011) ainsi que chez la souris (Yalcin et al., 2011). Ces données pourraient servir à décrire ces transitions et voir si l'on peut reproduire les lois observées.

Bibliographie

Aerts, S. (2012). Chapter 5 - Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets. *Current Topics in Developmental Biology : Transcriptional Switches During Development*, 98 :121–145. (Pages 31 et 96.)

Ahmed, M., Xu, J., and Xu, P.-X. (2012). EYA1 and SIX1 drive the neuronal developmental program in cooperation with the SWI/SNF chromatin-remodeling complex and SOX2 in the mammalian inner ear. *Development*, pages 1–13. (Page 280.)

Alon, U. (2007a). An Introduction to Systems Biology : Design Principles of Biological Circuits (Mathematical and Computational Biology Series vol 10). (Page 13.)

Alon, U. (2007b). Network motifs : theory and experimental approaches. *Nat Rev Genet*, 8(6) :450–461. (Page 13.)

An, C.-I., Dong, Y., and Hagiwara, N. (2011). Genome-wide mapping of Sox6 binding sites in skeletal muscle reveals both direct and indirect regulation of muscle terminal differentiation by Sox6. *BMC Dev Biol*, 11(1) :59. (Page 228.)

Asakura, A., Lyons, G. E., and Tapscott, S. J. (1995). The regulation of MyoD gene expression : conserved elements mediate expression in embryonic axial muscle. *Dev Biol*, 171(2) :386–98. (Pages 223 et 229.)

Asp, P., Blum, R., Vethantham, V., Parisi, F., Micsinai, M., Cheng, J., Bowman, C., Kluger, Y., and Dynlacht, B. D. (2011). PNAS Plus : Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proceedings of the National Academy of Sciences*, pages 1–11. (Pages 46 et 226.)

Attanasio, C., Reymond, A., Humbert, R., Lyle, R., Kuehn, M. S., Neph, S., Sabo, P. J., Goldy, J., Weaver, M., Haydock, A., Lee, K., Dorschner, M., Dermitzakis, E. T., Antonarakis, S. E., and Stamatoyannopoulos, J. A. (2008). Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. *Genome Biol*, 9(12) :R168. (Page 40.)

Aurell, E., d'Hérouël, A., Malmnäs, C., and Vergassola, M. (2007). Transcription factor concentrations versus binding site affinities in the yeast S. cerevisiae. *Physical biology*, 4 :134. (Page 20.)

Aziz, A., Liu, Q.-C., and Dilworth, F. J. (2010). Regulating a master regulator : Establishing tissue-specific gene expression in skeletal muscle. *epigenetics*, 5(8) :691–695. (Page 265.)

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 324(5935) :1720–1723. (Pages 50, 53 et 54.)

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2 :28–36. (Page 96.)

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependencies in protein-DNA binding sites. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, RECOMB '03, pages 28–37. ACM, New York, NY, USA. ISBN 1-58113-635-8. (Pages 51, 53 et 54.)

Bartel, D. P. (2009). MicroRNAs : target recognition and regulatory functions. *Cell*, 136(2) :215–33. (Page 11.)

Baxter, R. (2007). *Exactly Solved Models in Statistical Mechanics*. Dover Books on Physics Series. DOVER PUBN Incorporated. ISBN 9780486462714. (Page 58.)

Baylies, M. K., Bate, M., and Ruiz Gomez, M. (1998). Myogenesis : a view from Drosophila. *Cell*, 93(6) :921–7. (Page 13.)

Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3) :387–96. (Page 31.)

Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002). Additivity in protein-DNA interactions : how good an approximation is it ? *Nucleic Acids Res*, 30(20) :4442–51. (Page 52.)

Bentzinger, C. F., Wang, Y. X., and Rudnicki, M. A. (2012). Building muscle : molecular regulation of myogenesis. *Cold Spring Harb Perspect Biol*, 4(2). (Page 222.)

Berg, O. and von Hippel, P. (1987). Selection of DNA binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*, 193(4) :723–743. (Page 16.)

Berg, O. G., Winter, R. B., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20(24) :6929–48. (Page 15.)

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., 3rd, and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11) :1429–35. (Page 22.)

Bergstrom, D., Penn, B., Strand, A., Perry, R., Rudnicki, M., and Tapscott, S. (2002). Promoter-specific regulation of MyoD binding and signal transduction cooperate to pattern gene expression. *Molecular cell*, 9(3) :587–600. (Page 265.)

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A*, 99(2) :757–62. (Page 38.)

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1) :6–21. (Page 11.)

Blackwell, T. K. and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science*, 250(4984) :1104–10. (Page 22.)

Blais, A., Tsikitis, M., Acosta-Alvear, D., Sharan, R., Kluger, Y., and Dynlacht, B. (2005). An initial blueprint for myogenic differentiation. *Genes & development*, 19(5) :553. (Page 219.)

Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganière, J., Lefèbvre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., Coulombe, B., and Robert, F. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*, 16(5) :656–68. (Page 40.)

Blau, H. M., Pavlath, G. K., Hardeman, E. C., Chiu, C. P., Silberstein, L., Webster, S. G., Miller, S. C., and Webster, C. (1985). Plasticity of the differentiated state. *Science*, 230(4727) :758–66. (Page 6.)

Bolouri, H. and Davidson, E. H. (2002). Modeling DNA sequence-based cis-regulatory gene networks. *Dev Biol*, 246(1) :2–13. (Page 31.)

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., and Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*, 18(11) :1752–1762. (Page 34.)

Brazma, A., Parkinson, H., Schlitt, T., and Shojatalab, M. (2001). A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. `http://www.ebi.ac.uk/microarray/biology_intro.html`. (Page 4.)

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9) :5136–41. (Page 31.)

Buckingham, M. and Relaix, F. (2007). The role of Pax genes in the development of tissues and organs : Pax3 and Pax7 regulate muscle progenitor cell functions. *Annual review of cell and developmental biology*, 23(1) :645. (Page 220.)

Bulyk, M. L., Johnson, P. L. F., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5) :1255–61. (Page 50.)

Bussemaker, H. J., Li, H., and Siggia, E. D. (2000). Building a dictionary for genomes : identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A*, 97(18) :10096–100. (Page 99.)

Campbell, C. T. and Kim, G. (2007). SPR microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics. *Biomaterials*, 28(15) :2380–92. (Page 22.)

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G. J., Parker, M. H., Macquarrie, K. L., Davison, J., Morgan, M. T., Ruzzo, W. L., Gentleman, R. C., and Tapscott, S. J. (2010). Genome-wide MyoD binding in skeletal muscle cells : a potential for broad cellular reprogramming. *Developmental Cell*, 18(4) :662–74. (Pages 39, 46, 104, 219, 226, 265, 268 et 273.)

Carlson, C. D., Warren, C. L., Hauschild, K. E., Ozers, M. S., Qadir, N., Bhimsaria, D., Lee, Y., Cerrina, F., and Ansari, A. Z. (2010). Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci U S A*, 107(10) :4544–9. (Page 22.)

Carninci, P., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6) :626–35. (Page 28.)

Carvajal, J. J., Keith, A., and Rigby, P. W. J. (2008). Global transcriptional regulation of the locus encoding the skeletal muscle determination genes Mrf4 and Myf5. *Genes & development*, 22(2) :265–76. (Page 30.)

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3) :167–174. (Page 52.)

Chan, B. Y. and Kibler, D. (2005). Using hexamers to predict cis-regulatory motifs in Drosophila. *BMC Bioinformatics*, 6 :262. (Page 104.)

Cheng, T. C., Wallace, M. C., Merlie, J. P., and Olson, E. N. (1993). Separable regulatory elements governing myogenin transcription in mouse embryogenesis. *Science*, 261(5118) :215–8. (Page 265.)

Cheng, Y., King, D. C., Dore, L. C., Zhang, X., Zhou, Y., Zhang, Y., Dorman, C., Abebe, D., Kumar, S. A., Chiaromonte, F., Miller, W., Green, R. D., Weiss, M. J., and Hardison, R. C. (2008). Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res*, 18(12) :1896–905. (Page 41.)

Chung, J. H., Whiteley, M., and Felsenfeld, G. (1993). A 5′ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, 74(3) :505–14. (Page 31.)

Ciglar, L. and Furlong, E. E. M. (2009). Conservation and divergence in developmental networks : a view from Drosophila myogenesis. *Current opinion in cell biology*, 21(6) :754–60. (Page 14.)

Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10) :691–703. (Page 35.)

Davis, R. L., Weintraub, H., and Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, 51(6) :987–1000. (Pages 7 et 266.)

Deato, M. D. E. and Tjian, R. (2007). Switching of the core transcription machinery during myogenesis. *Genes Dev*, 21(17) :2137–49. (Page 265.)

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38. (Page 98.)

Dermitzakis, E. T. and Clark, A. G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions : conservation and turnover. *Mol Biol Evol*, 19(7) :1114–21. (Page 33.)

Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet*, 6(2) :151–7. (Page 40.)

D'haeseleer, P. (2006). How does DNA sequence motif discovery work ? *Nat Biotechnol*, 24(8) :959–61. (Page 97.)

Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13(11) :2381–90. (Page 18.)

Donaldson, I. J., Chapman, M., Kinston, S., Landry, J. R., Knezevic, K., Piltz, S., Buckley, N., Green, A. R., and Göttgens, B. (2005). Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Hum Mol Genet*, 14(5) :595–601. (Page 40.)

Elemento, O. and Tavazoie, S. (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol*, 6(2) :R18. (Page 143.)

ENCODE Project Consortium, et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74. (Page 47.)

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345) :43–9. (Page 41.)

Felsenstein, J. (1981). Evolutionary trees from DNA sequences : a maximum likelihood approach. *J Mol Evol*, 17(6) :368–76. (Pages 100 et 101.)

Ferrari, S., Molinari, S., Melchionna, R., Cusella-De Angelis, M. G., Battini, R., De Angelis, L., Kelly, R., and Cossu, G. (1997). Absence of MEF2 binding to the A/T-rich element in the muscle creatine kinase (MCK) enhancer correlates with lack of early expression of the MCK gene in embryonic mammalian muscle. *Cell Growth Differ*, 8(1) :23–34. (Page 279.)

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*, 9(5) :397–405. (Page 35.)

Fields, D. S., He, Y., Al-Uzri, A. Y., and Stormo, G. D. (1997). Quantitative specificity of the Mnt repressor. *J Mol Biol*, 271(2) :178–94. (Page 22.)

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, pages 1–5. (Page 36.)

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3) :131–163. (Page 53.)

Fuda, N. J., Ardehali, M. B., and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261) :186–92. (Page 265.)

Furusawa, C. and Kaneko, K. (2012). A Dynamical-Systems View of Stem Cell Biology. *Science*, 338(6104) :215–217. (Page 5.)

Gerland, U., Moroz, J., and Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor–DNA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 99(19) :12015. (Pages 16, 18, 19 et 20.)

Giocomo, L. M., Moser, M.-B., and Moser, E. I. (2011). Computational models of grid cells. *Neuron*, 71(4) :589–603. (Page 17.)

Giordani, J., Bajard, L., Demignon, J., Daubas, P., Buckingham, M., and Maire, P. (2007). Six proteins regulate the activation of Myf5 expression in embryonic mouse limbs. *Proceedings of the National Academy of Sciences*, 104(27) :11310. (Page 223.)

Grad, Y. H., Roth, F. P., Halfon, M. S., and Church, G. M. (2004). Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura. *Bioinformatics*, 20(16) :2738–50. (Page 103.)

Graf, T. and Enver, T. (2009). Forcing cells to change lineages. *Nature*, 462(7273) :587–94. (Page 7.)

Green, Y. S. and Vetter, M. L. (2011). EBF proteins participate in transcriptional regulation of Xenopus muscle development. *Developmental Biology*, 358(1) :240–250. (Page 273.)

Greer, E. L. and Shi, Y. (2012). Histone methylation : a dynamic mark in health, disease and inheritance. *Nat Rev Genet*, 13(5) :343–57. (Page 11.)

Grendar Jr, M. and Grendar, M. (2001). MiniMax Entropy and Maximum Likelihood : complementarity of tasks, identity of solutions. In *AIP Conference Proceedings*, volume 568, page 49. (Page 57.)

Grifone, R., Demignon, J., Giordani, J., Niro, C., and Souil, E. (2007). Eya1 and Eya2 proteins are required for hypaxial somitic myogenesis in the mouse embryo. *Developmental Biology*, 302(2) :602–616. (Pages 221 et 223.)

Grifone, R., Demignon, J., Houbron, C., Souil, E., Niro, C., Seller, M. J., Hamard, G., and Maire, P. (2005). Six1 and Six4 homeoproteins are required for Pax3 and Mrf expression during myogenesis in the mouse embryo. *Development*, 132(9) :2235–49. (Pages 221 et 266.)

Gurdon, J. B. and Melton, D. A. (2008). Nuclear reprogramming in cells. *Science*, 322(5909) :1811–5. (Page 7.)

Halpern, A. and Bruno, W. (1998). Evolutionary distances for protein-coding sequences : modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7) :910. (Page 102.)

Hammond, S. M., Caudy, A. A., and Hannon, G. J. (2001). Post-transcriptional gene silencing by double-stranded RNA. *Nat Rev Genet*, 2(2) :110–9. (Page 11.)

Hannon, G. J. (2002). RNA interference. *Nature*, 418(6894) :244–51. (Page 11.)

Hardison, R. C. and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13(7) :469–483. (Pages 29 et 41.)

Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature*, 402(6761) :47. (Pages 38 et 42.)

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2) :160–74. (Page 101.)

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109. (Pages 132 et 133.)

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3) :311–8. (Page 41.)

Heintzman, N. D., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243) :108–12. (Pages 31 et 41.)

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4) :576–89. (Page 104.)

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4) :283–9. (Page 27.)

Hong, J.-W., Hendrix, D. A., and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894) :1314. (Page 36.)

Hu, M., Yu, J., Taylor, J. M., Chinnaiyan, A. M., and Qin, Z. S. (2010). On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic acids research*, 38(7) :2154–2167. (Page 52.)

Ikeda, K., Watanabe, Y., Ohto, H., and Kawakami, K. (2002). Molecular interaction and synergistic activation of a promoter by Six, Eya, and Dach proteins mediated through CREB binding protein. *Molecular and Cellular Biology*, 22(19) :6759–66. (Page 265.)

Ivan, A., Halfon, M. S., and Sinha, S. (2008). Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol*, 9(1) :R22. (Page 103.)

Jaynes, E. T. (1978). Where do we stand on maximum entropy ? pages 1–105. (Page 54.)

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9) :939–952. (Page 57.)

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, 20(6) :861–73. (Page 23.)

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(1-2) :327–39. (Pages 23, 50, 53 et 93.)

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314) :430–435. (Page 37.)

Kantorovitz, M. R., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G. E., Göttgens, B., Halfon, M. S., and Sinha, S. (2009). Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. *Developmental Cell*, 17(4) :568–79. (Pages 39, 41, 103 et 104.)

Kantorovitz, M. R., Robinson, G. E., and Sinha, S. (2007). A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13) :i249–55. (Page 103.)

Kassar-Duchossoy, L., Gayraud-Morel, B., Gomès, D., Rocancourt, D., Buckingham, M., Shinin, V., and Tajbakhsh, S. (2004). Mrf4 determines skeletal muscle identity in Myf5 :Myod double-mutant mice. *Nature*, 431(7007) :466–71. (Page 219.)

Kaufmann, S. (1993). The origins of order. (Page 6.)

Kawakami, K., Sato, S., Ozaki, H., and Ikeda, K. (2000). Six family genes–structure and function as transcription factors and their roles in development. *Bioessays*, 22(7) :616–26. (Page 221.)

Kazemian, M., Zhu, Q., Halfon, M. S., and Sinha, S. (2011). Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. *Nucleic Acids Res*, 39(22) :9463–72. (Page 103.)

Keim, C. N., Martins, J. L., Abreu, F., Rosado, A. S., de Barros, H. L., Borojevic, R., Lins, U., and Farina, M. (2004). Multicellular life cycle of magnetotactic prokaryotes. *FEMS Microbiol Lett*, 240(2) :203–8. (Page 4.)

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res*, 17(12) :1919–31. (Page 40.)

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47(6) :713. (Page 102.)

Kinney, J. B., Tkacik, G., and Callan, C. G. (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA*, 104(2) :501–6. (Page 22.)

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science*, 329(5989) :336–9. (Page 11.)

Krauth, W. (2006). *Statistical mechanics : algorithms and computations*, volume 13. Oxford University Press. (Pages 52, 132 et 134.)

Kucharczuk, K. L., Love, C. M., Dougherty, N. M., and Goldhamer, D. J. (1999). Fine-scale transgenic mapping of the MyoD core enhancer : MyoD is regulated by distinct but overlapping mechanisms in myotomal and non-myotomal muscle lineages. *Development*, 126(9) :1957–65. (Pages 223, 228, 229, 230 et 265.)

Kulessa, H., Frampton, J., and Graf, T. (1995). GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev*, 9(10) :1250–62. (Page 7.)

Kulkarni, M. M. and Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development*, 130(26) :6569–75. (Pages 31 et 32.)

Lander, E. S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921. (Page 43.)

Lässig, M. (2007). From biophysics to evolutionary genetics : statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6) :S7. (Pages 16 et 19.)

Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1) :41–51. (Pages 96 et 97.)

Le Grand, F., Grifone, R., Mourikis, P., Houbron, C., Gigaud, C., Pujol, J., Maillet, M., Pagès, G., Rudnicki, M., Tajbakhsh, S., and Maire, P. (2012). Six1 regulates stem cell repair potential and self-renewal during skeletal muscle regeneration. *J Cell Biol*, 198(5) :815–32. (Page 223.)

Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., and Simon, I. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594) :799. (Pages 12 et 13.)

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14) :1725–35. (Pages 30 et 43.)

L'honoré, A., Ouimette, J.-F., Lavertu-Jolin, M., and Drouin, J. (2010). Pitx2 defines alternate pathways acting through MyoD during limb and somitic myogenesis. *Development*, 137(22) :3847–56. (Page 229.)

Liberman, L. M. and Stathopoulos, A. (2009). Design flexibility in cis-regulatory control of gene expression : Synthetic and comparative evidence. *Developmental Biology*, 327(2) :578–589. (Pages 33 et 34.)

Liu, Y., Chakroun, I., Yang, D., Horner, E., Liang, J., Aziz, A., Chu, A., De Repentigny, Y., Dilworth, F. J., Kothary, R., and Blais, A. (2013). Six1 Regulates MyoD Expression in Adult Muscle Progenitor Cells. *PLoS One*, 8(6) :e67762. (Page 280.)

Liu, Y., Chu, A., Chakroun, I., Islam, U., and Blais, A. (2010). Cooperation between myogenic regulatory factors and SIX family transcription factors is important for myoblast differentiation. *Nucleic acids research*. (Pages 46, 228, 265 et 268.)

Loots, G. G. and Ovcharenko, I. (2004). rVISTA 2.0 : evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue) :W217–21. (Page 46.)

Loots, G. G. and Ovcharenko, I. (2005). Dcode.org anthology of comparative genomic tools. *Nucleic Acids Res*, 33(Web Server issue) :W56–64. (Page 46.)

Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769) :564–7. (Page 33.)

Maerkl, S. and Quake, S. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809) :233. (Page 21.)

Majoros, W. H. and Ohler, U. (2010). Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. *PLoS Comput Biol*, 6(12) :e1001037. (Page 40.)

Man, T. K. and Stormo, G. D. (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res*, 29(12) :2471–8. (Page 50.)

Maniatis, T., Goodbourn, S., and Fischer, J. A. (1987). Regulation of inducible and tissue-specific gene expression. *Science*, 236(4806) :1237–45. (Page 30.)

Masuya, H., Sezutsu, H., Sakuraba, Y., Sagai, T., Hosoya, M., Kaneda, H., Miura, I., Kobayashi, K., Sumiyama, K., Shimizu, A., Nagano, J., Yokoyama, H., Kaneko, S., Sakurai, N., Okagaki, Y., Noda, T., Wakana, S., Gondo, Y., and Shiroishi, T. (2007). A series of ENU-induced single-base substitutions in a long-range cis-element altering Sonic hedgehog expression in the developing mouse limb bud. *Genomics*, 89(2) :207–14. (Page 43.)

McGregor, A., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D., Payre, F., and Stern, D. (2007). Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, 448(7153) :587–590. (Page 36.)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21 :1087. (Page 132.)

Mills, R. E., et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332) :59–65. (Page 284.)

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–7. (Page 13.)

Molkentin, J. and Olson, E. (1996). Combinatorial control of muscle development by basic helix-loop-helix and MADS-box transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 93(18) :9366. (Page 219.)

Moses, A., Chiang, D., Pollard, D., Iyer, V., and Eisen, M. (2004). MONKEY : identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, 5(12) :R98. (Page 100.)

Moses, A. M., Chiang, D. Y., Kellis, M., Lander, E. S., and Eisen, M. B. (2003). Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol*, 3 :19. (Page 101.)

Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D., and Eisen, M. B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*, 2(10) :e130. (Page 33.)

Mullen, A. C., Orlando, D. A., Newman, J. J., Lovén, J., Kumar, R. M., Bilodeau, S., Reddy, J., Guenther, M. G., Dekoter, R. P., and Young, R. A. (2011). Master Transcription Factors Determine Cell-Type-Specific Responses to TGF-&beta ; Signaling. *Cell*, 147(3) :565–576. (Page 226.)

Nagaraj, V. H., O'flanagan, R. A., and Sengupta, A. M. (2008). Better estimation of protein-DNA interaction parameters improve prediction of functional sites. *BMC Biotechnol*, 8(1) :94. (Page 23.)

Naidu, P. S., Ludolph, D. C., To, R. Q., Hinterberger, T. J., and Konieczny, S. F. (1995). Myogenin and MEF2 function synergistically to activate the MRF4 promoter during myogenesis. *Mol Cell Biol*, 15(5) :2707–18. (Page 219.)

Niro, C., Demignon, J., Vincent, S., Liu, Y., Giordani, J., Sgarioto, N., Favier, M., Guillet-Deniau, I., Blais, A., and Maire, P. (2010). Six1 and Six4 gene expression is necessary to activate the fast-type muscle gene program in the mouse primary myotome. *Developmental Biology*, 338(2) :168–182. (Pages 221 et 231.)

Nurse, P. and Hayles, J. (2011). The Cell in an Era of Systems Biology. *Cell*. (Page 8.)

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6) :730–2. (Page 33.)

Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., Fraenkel, E., Bell, G. I., and Young, R. A. (2004). Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303(5662) :1378–81. (Pages 12 et 13.)

Oliphant, A. R., Brandl, C. J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides : analysis of yeast GCN4 protein. *Mol Cell Biol*, 9(7) :2944–9. (Page 22.)

Ondek, B., Gloss, L., and Herr, W. (1988). The SV40 enhancer contains two distinct levels of organization. *Nature*, 333(6168) :40–5. (Page 30.)

Palacios, D., Summerbell, D., Rigby, P. W. J., and Boyes, J. (2010). Interplay between DNA Methylation and Transcription Factor Availability : Implications for Developmental Activation of the Mouse Myogenin Gene. *Molecular and Cellular Biology*, 30(15) :3805–3815. (Page 266.)

Palstra, R.-J., de Laat, W., and Grosveld, F. (2008). Beta-globin regulation and long-range interactions. *Adv Genet*, 61 :107–42. (Page 233.)

Panne, D. (2008). The enhanceosome. *Curr Opin Struct Biol*, 18(2) :236–42. (Pages 31 et 33.)

Park, P. J. (2009). ChIP-seq : advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10) :669–80. (Page 26.)

Parker, M., Seale, P., and Rudnicki, M. (2003). Looking back to the embryo : defining transcriptional networks in adult myogenesis. *Nature Reviews Genetics*, 4(7) :497–507. (Page 219.)

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118) :499–502. (Pages 33 et 40.)

Perry, M. W., Boettiger, A. N., and Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proc Natl Acad Sci U S A*, 108(33) :13570–5. (Page 36.)

Phillips, J. E. and Corces, V. G. (2009). CTCF : master weaver of the genome. *Cell*, 137(7) :1194–211. (Page 31.)

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E. M., Couronne, O., and Pennacchio, L. A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res*, 16(7) :855–63. (Page 33.)

Puri, P. L., Sartorelli, V., Yang, X. J., Hamamori, Y., Ogryzko, V. V., Howard, B. H., Kedes, L., Wang, J. Y., Graessmann, A., Nakatani, Y., and Levrero, M. (1997). Differential roles of p300 and PCAF acetyltransferases in muscle differentiation. *Mol Cell*, 1(1) :35–45. (Page 265.)

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics*, 3 :30. (Pages 39 et 99.)

Recillas-Targa, F., Pikaart, M. J., Burgess-Beusse, B., Bell, A. C., Litt, M. D., West, A. G., Gaszner, M., and Felsenfeld, G. (2002). Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A*, 99(10) :6883–8. (Page 31.)

Relaix, F., Demignon, J., Laclef, C., Pujol, J., Santolini, M., Niro, C., Lagha, M., Rocancourt, D., Buckingham, M., and Maire, P. (2013). Six homeoproteins directly activate Myod expression in the gene regulatory networks that control early myogenesis. *PLoS Genet*, 9(4) :e1003425. (Pages 221, 223, 229 et 231.)

Richard, A.-F., Demignon, J., Sakakibara, I., Pujol, J., Favier, M., Strochlic, L., Le Grand, F., Sgarioto, N., Guernec, A., Schmitt, A., Cagnard, N., Huang, R., Legay, C., Guillet-Deniau, I., and Maire, P. (2011). Genesis of muscle fiber-type diversity during mouse embryogenesis relies on Six1 and Six4 gene expression. *Dev Biol*, 359(2) :303–20. (Page 221.)

Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, 81 :145–66. (Page 233.)

Roh, T.-Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev*, 19(5) :542–52. (Page 41.)

Rouault, H., Mazouni, K., Couturier, L., Hakim, V., and Schweisguth, F. (2010). Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proceedings of the National Academy of Sciences*, 107(33) :14615. (Pages 96, 104, 105 et 139.)

Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J. G., Mermod, N., and Bucher, P. (2002). High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol*, 20(8) :831–5. (Page 23.)

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development*, 132(4) :797–803. (Page 43.)

Sato, S., Ikeda, K., Shioi, G., Nakao, K., Yajima, H., and Kawakami, K. (2012). Regulation of Six1 expression by evolutionarily conserved enhancers in tetrapods. *Dev Biol*, 368(1) :95–108. (Page 46.)

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing : higher than you think ! *Genome Biol*, 12(8) :125. (Page 44.)

Schirm, S., Jiricny, J., and Schaffner, W. (1987). The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. *Genes Dev*, 1(1) :65–74. (Page 30.)

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell*, 148(1-2) :335–348. (Page 35.)

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding SI. *Science*, 328(5981) :1036–40. (Page 33.)

Schoborg, T. A. and Labrador, M. (2010). The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. *J Mol Evol*, 70(1) :74–84. (Page 31.)

Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet*, 9(3) :179–91. (Page 10.)

Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol*, 2(9) :E271. (Page 39.)

Sethna, J. (2006). *Statistical Mechanics : Entropy, Order Parameters and Complexity*. Oxford Master Series in Physics. OUP Oxford. ISBN 9780198566779. (Page 56.)

Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1) :64–68. (Page 13.)

Shumaker-Parry, J. S., Aebersold, R., and Campbell, C. T. (2004). Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy. *Anal Chem*, 76(7) :2071–82. (Page 21.)

Sinha, S. and He, X. (2007). MORPH : probabilistic alignment combined with hidden Markov models of cis-regulatory modules. *PLoS Comput Biol*, 3(11) :e216. (Page 40.)

Sinha, S., Nimwegen, E. V., and Siggia, E. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(Suppl 1) :i292. (Pages 99 et 100.)

Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U., and Siggia, E. D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics*, 5 :129. (Page 40.)

Sinha, S. and Tompa, M. (2000). A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol*, 8 :344–54. (Page 104.)

Slutsky, M. and Mirny, L. A. (2004). Kinetics of protein-DNA interaction : facilitated target location in sequence-dependent potential. *Biophys J*, 87(6) :4021–35. (Page 16.)

Smith, A. D., Sumazin, P., Xuan, Z., and Zhang, M. Q. (2006). DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A*, 103(16) :6275–80. (Page 39.)

Smith, A. D., Sumazin, P., and Zhang, M. Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A*, 102(5) :1560–5. (Page 39.)

Soleimani, V. D., Punch, V. G., ichi Kawabe, Y., Jones, A. E., Palidwor, G. A., Porter, C. J., Cross, J. W., Carvajal, J. J., Kockx, C. E. M., Ijcken, W. F. J. V., Perkins, T. J., Rigby, P. W. J., Grosveld, F., and Rudnicki, M. A. (2012). Transcriptional Dominance of Pax7 in Adult Myogenesis Is Due to High-Affinity Recognition of Homeodomain Motifs. *Developmental Cell*, pages 1–13. (Page 226.)

Spitz, F., Demignon, J., Porteu, A., Kahn, A., Concordet, J., Daegelen, D., and Maire, P. (1998). Expression of myogenin during embryogenesis is controlled by Six/sine oculis homeoproteins through a conserved MEF3 binding site. *Proceedings of the National Academy of Sciences*, 95(24) :14220. (Pages 223 et 265.)

Spitz, F., Salminen, M., Demignon, J., Kahn, A., Daegelen, D., and Maire, P. (1997). A combination of MEF3 and NFI proteins activates transcription in a subset of fast-twitch muscles. *Molecular and cellular biology*, 17(2) :656. (Page 223.)

Stormo, G. and Fields, D. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences*, 23(3) :109–113. (Page 18.)

Stormo, G. D. and Zhao, Y. (2007). Putting numbers on the network connections. *Bioessays*, 29(8) :717–21. (Page 21.)

Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nature Reviews Genetics*, 11(11) :751–60. (Pages 21 et 24.)

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4) :663–76. (Page 7.)

Taylor, J., Tyekucheva, S., King, D. C., Hardison, R. C., Miller, W., and Chiaromonte, F. (2006). ESPERR : learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res*, 16(12) :1596–604. (Page 40.)

Thayer, M. J., Tapscott, S. J., Davis, R. L., Wright, W. E., Lassar, A. B., and Weintraub, H. (1989). Positive autoregulation of the myogenic determination gene MyoD1. *Cell*, 58(2) :241–8. (Pages 219 et 228.)

Thurman, R. E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414) :75–82. (Page 27.)

Tijssen, M. R., Cvejic, A., Joshi, A., Hannah, R. L., Ferreira, R., Forrai, A., Bellissimo, D. C., Oram, S. H., Smethurst, P. A., Wilson, N. K., Wang, X., Ottersbach, K., Stemple, D. L., Green, A. R., Ouwehand, W. H., and Göttgens, B. (2011). Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell*, 20(5) :597–609. (Page 41.)

Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., and Barkai, N. (2008). On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol*, 4 :159. (Page 33.)

Trinklein, N. D., Aldred, S. J. F., Saldanha, A. J., and Myers, R. M. (2003). Identification and functional analysis of human transcriptional promoters. *Genome Res*, 13(2) :308–12. (Page 41.)

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment : RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968) :505–10. (Page 22.)

U.S. Department of Energy (2001). Genomes to life : accelerating biological discovery (Office of Biological and Environmental Research and Office of Advanced Scientific Computing Research of the U.S. Department of Energy). `http://genomicscience.energy.gov/roadmap/GTLcomplete_web.pdf`. (Pages 8 et 9.)

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors : function, expression and evolution. *Nat Rev Genet*, 10(4) :252–63. (Page 9.)

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009a). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231) :854–8. (Page 41.)

Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009b). Genomic views of distant-acting enhancers. *Nature*, 461(7261) :199–205. (Pages 25, 42 et 43.)

Vokes, S., Ji, H., Wong, W., and McMahon, A. (2008). A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb SI. *Genes & development*, 22(19) :2651. (Page 228.)

Waddington, C. H. et al. (1957). The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*, pages ix+–262. (Page 5.)

Wallace, J. A. and Felsenfeld, G. (2007). We gather together : insulators and genome organization. *Curr Opin Genet Dev*, 17(5) :400–7. (Page 31.)

Wang, J., Kumar, R. M., Biggs, V. J., Lee, H., Chen, Y., Kagey, M. H., Young, R. A., and Abate-Shen, C. (2011). The Msx1 Homeoprotein Recruits Polycomb to the Nuclear Periphery during Development. *Dev Cell*, pages 1–14. (Page 228.)

Wang, Q., Carroll, J. S., and Brown, M. (2005). Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol Cell*, 19(5) :631–42. (Page 31.)

Warren, C. L., Kratochvil, N. C. S., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., Phillips, G. N., Jr, and Ansari, A. Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci U S A*, 103(4) :867–72. (Page 22.)

Wasserman, W. and Fickett, J. (1998). Identification of regulatory regions which confer muscle-specific gene expression1. *Journal of molecular biology*, 278(1) :167–181. (Page 38.)

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4) :276–87. (Pages 17, 44 et 96.)

Weintraub, H., Tapscott, S. J., Davis, R. L., Thayer, M. J., Adam, M. A., Lassar, A. B., and Miller, A. D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci U S A*, 86(14) :5434–8. (Page 13.)

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2) :307–19. (Pages 36, 37, 227 et 229.)

Wilczynski, B. and Furlong, E. E. M. (2010). Challenges for modeling global gene regulatory networks during development : Insights from Drosophila. *Developmental Biology*, 340(2) :161–169. (Page 30.)

Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., Fisher, E. M. C., Tavaré, S., and Odom, D. T. (2008). Species-specific transcription in mice carrying human chromosome 21. *Science*, 322(5900) :434–8. (Page 34.)

Wilson, M. D. and Odom, D. T. (2009). Evolution of transcriptional control in mammals. *Curr Opin Genet Dev*, 19(6) :579–85. (Pages 30, 33 et 35.)

Winter, R. B., Berg, O. G., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor–operator interaction : kinetic measurements and conclusions. *Biochemistry*, 20(24) :6961–77. (Page 15.)

Winter, R. B. and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli repressor–operator interaction : equilibrium measurements. *Biochemistry*, 20(24) :6948–60. (Page 15.)

Wright, W. E., Binder, M., and Funk, W. (1991). Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol Cell Biol*, 11(8) :4104–10. (Page 22.)

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, 434(7031) :338–45. (Page 40.)

Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T. M., Gan, X., Nellåker, C., Goodstadt, L., Nicod, J., Bhomra, A., Hernandez-Pliego, P., Whitley, H., Cleak, J., Dutton, R., Janowitz, D., Mott, R., Adams, D. J., and Flint, J. (2011). Sequence-based characterization of structural variation in the mouse genome. *Nature*, 477(7364) :326–9. (Page 284.)

Yokoyama, S., et al. (2009). A Systems Approach Reveals that the Myogenesis Genome Network Is Regulated by the Transcriptional Repressor RP58. *Developmental Cell*, 17(6) :836–848. (Page 278.)

Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev*, 21(4) :385–90. (Page 36.)

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9) :R137. (Page 26.)

Zhao, Y., Granas, D., and Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12) :e1000590. (Page 18.)

Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6) :909–916. (Page 52.)

Zhou, Q. and Wong, W. H. (2004). CisModule : de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33) :12114–9. (Page 39.)

Zinzen, R., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269) :65–70. (Pages 33, 139 et 215.)

Zykovich, A., Korf, I., and Segal, D. J. (2009). Bind-n-Seq : high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res*, 37(22) :e151. (Page 23.)

Abstract

Cellular differentiation and tissue specification depend in part on the establishment of specific transcriptional programs of gene expression. These programs result from the interpretation of genomic regulatory information by sequence-specific transcription factors (TFs). Decoding this information in sequenced genomes is a key issue.

In a first part, we study the interaction between the TFs and the DNA sequences they bind to, called Transcription Factor Binding Sites (TFBSs). Using a Potts model inspired from spin glass physics along with high-throughput binding data for a variety of Drosophilae and mammalian TFs, we show that TFBSs exhibit correlations among nucleotides and that the account of their contribution in the binding energy greatly improves the predictability of genomic TFBSs.

Then, we present Imogene, an extension to mammalian genomes of a Bayesian, phylogeny-based algorithm designed to computationally identify the Cis-Regulatory Modules (CRMs) that control gene expression in a set of co-regulated genes, and that was previously applied to Drosophila regulation. Starting with a small number of CRMs in a reference species as a training set, but with no a priori knowledge of the factors acting in trans, the algorithm uses the over-representation and conservation of TFBSs among related species to predict putative regulatory elements along with genomic CRMs underlying co-regulation. We present several applications of this algorithm both in Drosophila and vertebrates. We also present an extension of the algorithm to the case of pattern recognition, showing that CRMs with different patterns of expression can be distinguished on the sole basis of their DNA motifs content.

Finally, we present applications of these modeling tools to real biological cases : the trichomes differentiation in Drosophila, and the skeletal muscle differentiation in the mouse. In both cases, predictions were experimentally validated in a joint work with biological teams, and point towards a great flexibility of the cis-regulatory processes.

Keywords: Gene regulation, Transcription Factor, Potts Model, Phylogeny, Bayesian algorithm, muscle differentiation, trichomes.

Résumé

La différenciation cellulaire et la spécification des tissus biologiques dépendent en partie de l'établissement de programmes d'expression génétique caractéristiques. Ces programmes sont le résultat de l'interprétation de l'information génomique par des Facteurs de Transcription (TFs) se fixant à des séquences d'ADN spécifiques. Décoder cette information dans les génomes séquencés est donc un enjeu majeur.

Dans une première partie, nous étudions l'interaction entre les TFs et leurs sites de fixation sur l'ADN. L'utilisation d'un modèle de Potts inspiré de la physique des verres de spin et de données de fixation à grande échelle pour plusieurs TFs de la drosophile et des mammifères permet de montrer que les sites de fixation exhibent des corrélations entre nucléotides. Leur prise en compte permet d'améliorer significativement la prédiction des sites de fixations sur le génome.

Nous présentons ensuite Imogene, l'extension au cas des mammifères d'un algorithme bayésien utilisant la phylogénie afin d'identifier les motifs et modules de cis-régulation (CRMs) contrôlant l'expression d'un ensemble de gènes co-régulés, qui a précédemment été appliqué au cas de la régulation chez les drosophiles. Partant d'un ensemble d'apprentissage constitué d'un petit nombre de CRMs chez une espèce de référence, et sans connaissance *a priori* des TFs s'y fixant, l'algorithme utilise la sur-représentation et la conservation des sites de fixation chez des espèces proches pour prédire des régulateurs putatifs ainsi que les CRMs génomiques sous-tendant la co-régulation. Nous montrons en particulier qu'Imogene peut distinguer des modules de régulation conduisant à différents motifs d'expression génétique sur la seule base de leur séquence ADN.

Enfin, nous présentons des applications de ces outils de modélisation à des cas biologiques réels : la différenciation des trichomes chez la drosophile, et la différenciation musculaire chez la souris. Dans les deux cas, les prédictions ont été validées expérimentalement en collaboration avec des équipes de biologistes, et pointent vers une grande flexibilité des processus de cis-régulation.

Mots-clés: Régulation génétique, Facteur de transcription, Modèle de Potts, Phylogénétique, Algorithme bayésien, différenciation musculaire, trichomes.