Découverte d'éléments cis-régulateurs impliqués dans l'activation transcriptionnelle du génome zygotique dans l'embryon précoce de Drosophila melanogaster

Elodie Darbo

Univeristé de la Méditerranée
Faculté des Sciences de Luminy
Ecole Doctorale des Sciences de la Vie et de la Santé

Thèse
présentée en vue de l'obtention du grade de

Docteur de l'Université de la Méditerranée
Spécialité Bioinformatique, Biochimie Structurale et Génomique
par
Élodie Darbo

DÉCOUVERTE D'ÉLÉMENTS CIS-RÉGULATEURS IMPLIQUÉS DANS L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE DANS L'EMBRYON PRÉCOCE DE DROSOPHILA MELANOGASTER

sous la co-direction de Denis THIEFFRY et Jacques VAN HELDEN
TAGC INSERM U928 case 928, 163 avenue de Luminy, 13288 Marseille

Date Prévue le vendredi 16 Décembre

Philipp BUCHER	Rapporteur
Nathalie DOSTATNI	Rapporteur
Pedro COUTINHO	Examinateur
Yacine GRABA	Examinateur
Julien ROYET	Examinateur
Denis THIEFFRY	Directeur de thèse
Jacques VAN HELDEN	Co-directeur de thèse

"Au som deu malh, que i a ua lutz,

Qu'i cau guardar los uelhs dessùs,

Que'ns cau traucar tot lo segàs,

Tà ns'arrapar, sonque las mans,

Lhèu veiram pas jamei la fin,

La libertat qu'ei lo camin,

Après lo malh, un aute malh,

Après la lutz, ua auta lutz ..."

D'après Los de Nadau

Résumé

Chez les métazoaires, la transcription est inactive durant les étapes précoces du développement embryonnaire. Chez Drosophila melanogaster, des études récentes de l'activation du génome zygotique (AGZ) ont mis en évidence l'implication de quelques acteurs moléculaires (Zelda, STAT92E), mais les mécanismes régulateurs généraux restent à découvrir. En appliquant des méthodes bioinformatiques à l'analyse de données à haut débit de différentes sources, j'ai recherché de nouveaux éléments cis-régulateurs impliqués dans l'AGZ. Tout d'abord, par l'analyse de données transcriptomiques, j'ai sélectionné un groupe de gènes activés pendant l'AGZ. L'analyse de leurs régions non codantes a mis en évidence neuf motifs, dont trois correspondent à des motifs connus (Zelda, Trl et TTK). La recherche systématique de ces motifs m'a permis de prédire des modules cis-régulateurs (CRMs) potentiels pour lesquels j'ai défini un environnement chromatinien spécifique en analysant des profils d'occupation de (co-) facteurs de transcription pertinents et d'histones modifiées (ChIP-seq) ainsi que des profils d'ouverture de la chromatine. L'ensemble de ces résultats m'a permis de définir un modèle de régulation de l'AGZ et de sélectionner des régions candidates pour une validation expérimentale.

Table des matières

Table des figures

Liste des tableaux

1

Introduction

Chez les métazoaires, le développement de l'embryon comprend plusieurs étapes clés dont les principales sont la segmentation de l'embryon menant à la formation du blastoderme, la gastrulation et l'organogenèse. Chacune de ces étapes repose sur des différenciations et mouvements cellulaires successifs. La différenciation des cellules requiert l'expression spécifique de gènes dont les produits vont déterminer un type cellulaire donné. Lors de cette thèse je me suis particulièrement intéressée à la période englobant la fécondation et la formation du blastoderme chez *Drosophila melanogaster*. Juste après la fécondation, durant quelques divisions mitotiques, alors que le génome zygotique n'est pas transcrit, le contrôle du développement dépend des ARN messagers et protéines déposés dans l'oocyte par la mère durant l'oogenèse. L'activation transcriptionnelle du génome zygotique (AGZ), concomitante à la dégradation des produits maternels, conduit à la synthèse des nouveaux transcrits qui seront nécessaires à la formation du blastoderme ainsi qu'aux étapes ultérieures du développement. La transition du contrôle maternel vers le contrôle embryonnaire du développement est appelée "*transition maternelle-zygotique*". Ce phénomène est intensément étudié mais les mécanismes de sa régulation restent encore mal compris. Les travaux menés durant cette thèse ont porté particulièrement sur la compréhension des mécanismes de régulation de l'activation du génome zygotique. Dans ce chapitre, je vais tout d'abord introduire les concepts inhérents à la régulation transcriptionnelle ainsi que les outils bioinformatiques nécessaires à son étude. Je vais, par la suite, détailler les mécanismes connus comme étant impliqués dans la transition maternelle-zygotique chez la drosophile, et plus particulièrement dans l'activation du génome zygotique.

1.1 La régulation transcriptionnelle

1.1.1 Mise en place de la machinerie basale de transcription

La transcription des gènes codant pour des protéines est assurée par l'ARN polymérase II (ARNpolII) et conduit à la synthèse d'ARN messagers (ARNm). l'ARNpolII ne reconnaît pas à elle seule les sites d'initiation de la transcription (TSS pour "*Transcription Start Site*") présents en amont des gènes. Son recrutement dépend de protéines, appelées facteurs généraux de transcription, appartenant à la famille TFII (de A à H) (lire la référénce (1) pour une revue). Un complexe de pré-initiation se met en place sur une région spécifique appelée promoteur basal, qui s'étend sur environ 100 pb autour du TSS. La formation du complexe se déroule de façon séquentielle. En effet, TFIID et TFIIB vont d'abord se lier à des sites spécifiques sur l'ADN (2), l'ARNpolII et TFIIF viennent se positionner et recrutent TFIIE, qui à son tour va recruter TFIIH. TFIIH possède deux activités enzymatiques : une activité hélicase qui permet d'ouvrir la double hélice d'ADN, et une activité kinase qui permet la phosphorylation du domaine C-terminal (CTD) de l'ARNpolII. La phosphorylation du CTD permet à l'ARNpolII de se détacher du complexe de pré-initiation (en anglais "*Pre-Initiation Complex*", PIC) et d'initier la transcription. La formation du PIC est une étape essentielle, mais son recrutement spécifique aux gènes qui doivent être transcrits requiert l'action d'autres facteurs protéiques.

1.1.2 Les facteurs de transcription

1.1.2.1 Définition

Les facteurs de transcription (FT) sont des protéines qui reconnaissent de courtes séquences d'ADN (entre 5 et 30 pb), appelées sites de liaison, grâce aux domaines de liaison à l'ADN qu'elles comportent. Les FT peuvent être classés en super-familles selon la structure de leur domaine de liaison à l'ADN. En effet, certaines protéines contiennent des domaines présentant des structures tridimensionnelles similaires : hélice-tour-hélice, doigts de zinc, architecture à feuillets béta etc. Les séquences reconnues par les FT sont situées dans des régions en cis (liées génétiquement au gène-cible) pouvant se trouver dans une région de quelques centaines de bases en amont du site d'initiation de la transcription (promoteurs proximaux), dans les introns, ou dans des régions distales en amont ou en aval, parfois à plusieurs milliers de paires de bases du promoteur proximal. Quand ces régions permettent une augmentation de la transcription, on les nomme "enhancers" ; quand au contraire, elles diminuent le taux de transcription,

elles se nomment "silencers". On appelle module cis-régulateur (CRM) une région de quelques centaines de paires de bases caractérisée par la présence d'une série de sites de liaison pour un (CRM homotypique) ou plusieurs (CRM hétérotypique) facteurs transcriptionnels. L'effet d'activation ou de répression d'un CRM résulte des interactions particulières entre ces sites et les FT actifs à un moment donné et dans un type cellulaire particulier.

Les FT sont donc capables de réguler positivement et/ou négativement la transcription via des interactions directes ou indirectes (en collaborant avec des co-facteurs qui seront décrits plus tard) avec la machinerie basale de transcription. Un FT régule généralement l'activité de plusieurs gènes qui présentent des niveaux d'expression différents dans des tissus et à des stades de développement spécifiques. La modulation du niveau d'expression peut être due à la liaison du FT à différentes séquences avec différentes affinités (3) ou à l'effet de combinaison de facteurs qui interagissent au sein de CRM.

1.1.2.2 Caractérisation des sites de liaison protéine-ADN

Afin d'identifier les séquences reconnues par les FT, plusieurs méthodes ont été mises au point. Les sites collectés sont généralement alignés pour modéliser la spécificité des séquences reconnues par chaque facteur. La figure 1.1 présente les principales techniques utilisées pour détecter les sites de liaision des FT à l'ADN.

La technique de recherche d'empreinte de liaison de protéines à l'ADN grâce à la DNAse1 (4) ("*DNAse1 footprinting*" en anglais) requiert la définition préalable d'une région régulatrice relativement courte (généralement une partie de promoteur). Ainsi, une seule séquence à la fois est analysée. Brièvement, la région à analyser est amplifiée par PCR puis marquée par fluorescence. Le FT soupçonné de se lier dans cette région est rajouté à une partie des échantillons ("positifs"), les autres servant de contrôle négatif. La DNAse1 est ajoutée à chaque échantillon (positifs et négatifs) et va digérer l'ADN. Les nucléotides directement liés à la protéine sont protégés de la dégradation et laissent ainsi une "empreinte" révélée par la comparaison de la migration sur gel des deux échantillons. Enfin, la technique de Maxam et Gilbert (5) permet de lire directement la séquence des segments d'intérêt sur le gel.

La méthode de SELEX (6) est basée sur l'utilisation d'une banque d'oligonucléotides générés aléatoirement qui sont mis en présence du facteur d'intérêt. Les séquences liées par la protéine sont récupérées, éluées puis amplifiées. Un nouveau jeu de séquences enrichi en sites de liaison est ainsi généré, qui sera à son tour amplifié et enrichi. Cette procédure est utilisée itérativement (typiquement 4 ou 5 cycles d'amplification-sélection) et permet de sélectionner

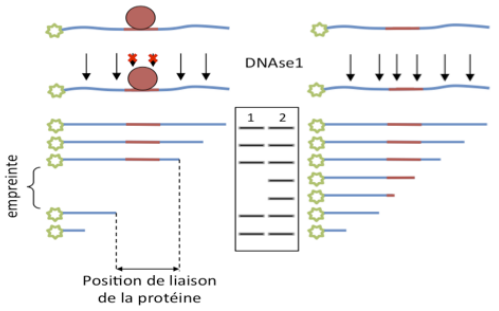
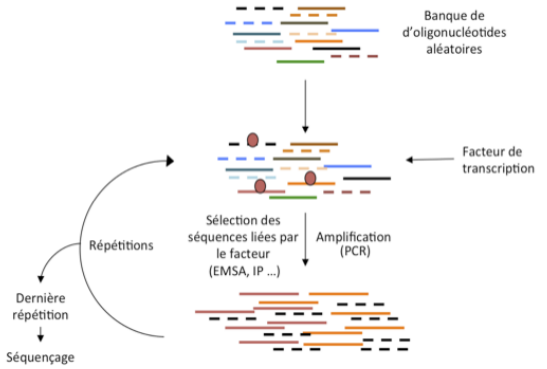
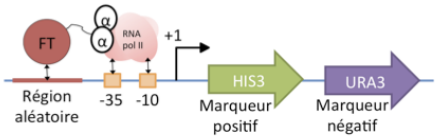
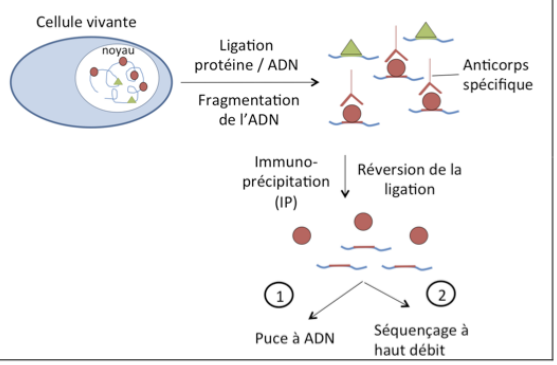
Technique	Caractéristiques	Principe
Empreinte à la DNAse1	Séquence régulatrice unique correspondant à un promoteur donné Détection de sites de liaison de différentes affinités Peut-être peu précis *In vitro*	
SELEX	Banque d'oligonucléotides aléatoires Détection de sites à haute affinité *In vitro*	
Système bactérien à hybridation simple	Banque d'oligonucléotides aléatoires Détection de sites à haute affinité Limitation de la longueur des oligonucléotides testés Problème de compatibilité eucaryote/procaryote *In vitro*	
ChIP + hybridation sur puce (1) ChIP + sequençage à haut débit (2)	Localisation sur le génome entier Détection de sites de liaison de différentes affinités Production d'anticorps spécifiques *In vivo*	

FIGURE 1.1: Représentation schématique des principales méthodes pour la collection de sites de liaison de facteurs de transcription (FT) - l'étoile verte indique un marquage de la séquence (e.g. fluorescence)

un petit nombre de séquences ayant une forte affinité avec la protéine. Les oligonucléotides sélectionnés sont ensuite séquencés (7). Cette méthode ne permet donc pas d'identifier tous les sites de liaison possibles du FT. D'une part, la méthode repose sur des oligonucléotides artificiels et ne permet donc pas d'identifier les positions génomiques de sites fonctionnels. D'autre part, les collections de sites obtenues par SELEX présentent généralement une affinité "optimale" pour le facteur, et ne représentent pas la diversité des séquences reconnues in vivo par la protéine.

Le système bactérien simple hybride (8) est également basé sur la génération d'oligonucléotides aléatoires. Deux constructions sont utilisés : chaque oligonucléotide est fusionné à un promoteur liant l'ARNpolII suivit de deux régions codant HIS3, impliqué dans la biosynthèse de l'histidine (pour la sélection positive), et URA3 (pour la sélection négative). La seconde construction contient la séquence codante pour le FT à analyser fusionnée à la sous-unité alpha de l'ARN polymérase. Une première expérience de sélection négative est opérée pour tester si les oligonucléotides générés n'activent pas seuls la transcription. La construction contenant HIS3 et URA3 est transfectée seule dans la bactérie en présence d'un substrat de URA3 dont la modification est toxique pour la cellule. Ensuite, les constructions contenues dans les bactéries survivantes sont récupérées et les deux constructions sont transfectées simultanément. Les bactéries sont mises en culture dans un milieu riche en un compétiteur de l'histidine (3-AT). Si le FT reconnaît l'oligonucléotide, il recrutera l'ARNpolII et permettra la synthèse d'histidine et ainsi permettra la survie des bactéries. Même si la concentration en 3-AT peut être modulée pour rechercher des sites de plus ou moins haute affinité, cette méthode détecte difficilement les sites de basse affinité. En effet, il existe une compétition entre les sites présents dans le génome de la bactérie et la séquence testée qui est en unique exemplaire (9). Le choix du système bactérien, par rapport à un système levure, permet de transfecter des plasmides facilement (10). Cependant, ce choix pose quelques problèmes pour étudier le comportement de protéines eucaryotes car les protéines peuvent être mal repliées (modifications post-traductionnelles défficientes), non exprimées ou vite dégradées (9).

Les méthodes combinant l'immuno-précipitation de la chromatine (ChIP) et la détection des séquences nucléotidiques par hybridation sur puce à ADN (11) (ChIP-chip) ou par séquençage à haut débit (12) (ChIP-seq), permettent de sélectionner tous les sites liés par le facteur d'intérêt en procédant à l'extraction de protéines liées à l'ADN à partir d'une cellule vivante. Cette méthode est probablement celle qui permet d'obtenir la plus grande diversité de séquences reconnues par le facteur d'intérêt. De plus, l'utilisation de cellules vivantes permet

d'étudier la liaison du facteur dans différents tissus et à différents stades du développement. Cependant, ces techniques reposent sur l'utilisation d'anticorps spécifiques qui ne sont pas forcément disponibles et dont la production est coûteuse.

A titre d'illustration, la figure 1.2 montre les sites résultant d'expérience de DNAse1 footprinting avec le facteur Hunchback (Hb) sur les séquences promotrices du gène even-skipped (13). Hb est un facteur impliqué dans la polarisation antéro-postérieure de l'embryon.

1.1.2.3 Représentations de la spécificité de liaison des facteurs transcriptionnels : les motifs de liaison

La représentation de la spécifité de liaison des FT est problème abordé depuis plusieurs décénies en bioinformatique ((14, 15, 16), voir (3) pour une revue).

À partir des collections de sites générées par une des méthodes citées ci-dessus, nous pouvons procéder à l'alignement de ces sites (figure 1.2A) pour définir un motif de liaison, qui représente la spécificité de liaison d'un facteur transcriptionnel. Un type familier de représentation est la séquence consensus (figure 1.2B) qui donne une idée de la conservation et des variants possibles au moyen du code IUPAC (figure 1.3). Cependant la génération de telles séquences est relativement arbitraire (3, 17), en particulier les pourcentages retenus pour considérer qu'une position donnée sera représentée par un nucléotide unique, une paire ou un triplet. Dans le cas de Hb, il est intéressant de noter que la séquence consensus (figure 1.2B) ne permet d'identifier aucun des 16 sites de la collection (figure 1.2A). Pour pouvoir l'utiliser à des fins de reconnaissance, il faudrait réduire le consensus et le rendre plus générique (en admettant plus de variations), mais il perdrait alors sa spécificité et détecterait des sites non liés par le facteur.

Une alternative pour la représentation des motifs de liaison est la matrice occurrences-positions (figure 1.2C) dans laquelle nous pouvons voir le nombre de chaque nucléotide (ligne) à chaque position (colonne) d'un alignement de sites. On peut en dériver la matrice poids-positions (figure 1.2D). Le poids d'un nucléotide est calculé en fonction de sa fréquence relative au sein de l'alignement, corrigée par un pseudo-poids pour éviter les biais dus aux faibles nombres d'observations, divisée par la probabilité a priori de chaque nucléotide (fréquence de chaque nucléotide dans le modèle de background, par exemple l'ensemble des séquences non-codantes d'un génome). La formule sera détaillée dans la section suivante (figure 1.5). Le motif correspondant peut être représenté sous forme de logo, (figure 1.2E), grâce à l'algorithme weblogo (18) (http ://weblogo.berkeley.edu/). Le logo consiste en un empilement des lettres

A,C,G,T dont la hauteur totale représente le contenu informationnel associé à chaque position, tandis que la hauteur de chaque nucléotide représente sa fréquence relative (19). La représentation d'un motif par une matrice est beaucoup plus informative que la séquence consensus, et le logo en découlant donne une bonne impression subjective de l'importance de chaque résidu pour la liaison protéine-ADN. Le contenu informationnel de la matrice est calculée en additionnant l'information de toutes les colonnes. Il indique la capacité de la matrice, à discriminer des sites réels de leur contexte ("background", estimé sur base de la fréquence de chaque nucléotide dans un jeu de séquences de référence) (20). Nous verrons par la suite que certains algorithmes optimisent cette grandeur pour découvrir des motifs surreprésentés dans des jeux de séquences particuliers.

1.1.2.4 Découverte de motifs

La découverte de motifs est généralement utilisée pour détecter des motifs exceptionnels présents dans des jeux de séquences fonctionnellement reliées (ex. promoteurs de gènes co-exprimés, collection de sites identifiés expérimentalement). Il existe deux grands types d'approches : les méthodes heuristiques basées des techniques d'optimisation des matrices et les méthodes énumératives basées sur le comptage de mots.

Un premier groupe de méthodes optimise un score (contenu informationnel ou autre) pour une série de paramètres définis par l'utilisateur (longueur de matrice, nombre de sites à aligner). Pour chaque motif généré, le score est calculé et les motifs ayant les meilleurs scores sont retenus. Le nombre de matrices possibles dépend du nombre et de la taille des séquences en entrée, de la taille du motif recherché et du nombre et de la répartition attendus des sites. Le nombre de matrices possibles augmente de façon exponentielle en fonction du nombre de séquences si l'on cherche exactement un site par séquence (par exemple pour les collections de sites), mais ce nombre devient encore plus énorme si l'on estime qu'une séquence peut contenir zéro ou plusieurs sites (ex. séquences non-codantes de gènes co-exprimés). En effet, cela revient à chercher toutes les combinaisons possibles d'une longueur données (taille de la matrice) parmi toutes les positions d'un jeu de séquences. Afin de pallier ce problème, les algorithmes basés sur les matrices ne testent qu'une partie des possibilités et retournent le meilleur score obtenu. Je vais brièvement décrire trois algorithmes : *MEME* (21), *CONSENSUS* (22) et *info-gibbs* (23).

MEME (Multiple Expectation Maximisation for Motif Elicitation) teste chaque oligonucléotide d'une taille donné en construisant une matrice "semence" et scanne les séquences afin

A

site1	G	C	A	T	T	A	A	A	A	A
site2	C	C	G	T	A	A	A	G	A	G
site3	T	C	G	T	A	A	A	A	A	C
site4	G	C	A	C	A	A	A	A	A	T
site5	T	C	A	T	A	A	A	A	A	C
site6	C	A	C	T	A	A	A	A	A	T
site7	G	A	A	T	T	A	A	A	A	A
site8	G	C	C	A	A	A	A	A	A	A
site9	C	A	A	T	A	A	A	A	A	G
site10	C	C	G	A	A	A	A	A	C	T
site11	G	G	G	A	A	A	A	A	A	A
site12	G	C	A	G	C	A	A	A	A	A
site13	G	A	A	C	A	A	A	A	A	A
site14	G	A	A	C	A	A	A	G	A	A
site15	C	G	C	A	A	A	A	A	A	A
site16	A	A	A	T	A	A	A	A	A	A

B

Consensus	S	M	R	W	A	A	A	A	A	A

C

PO	1	2	3	4	5	6	7	8	9	10
A	1	6	9	4	13	16	16	14	15	9
C	5	8	3	3	1	0	0	0	1	2
G	8	2	4	1	0	0	0	2	0	2
T	2	0	0	8	2	0	0	0	0	3

D

PO	1	2	3	4	5	6	7	8	9	10
A	-1.2	0.4	0.8	0.0	1.1	1.3	1.3	1.2	1.3	0.8
C	0.2	0.7	-0.3	-0.3	-1.2	-2.8	-2.8	-2.8	-1.2	-0.6
G	0.7	-0.6	0.0	-1.2	-2.8	-2.8	-2.8	-0.6	-2.8	-0.6
T	-0.6	-2.8	-2.8	0.7	-0.6	-2.8	-2.8	-2.8	-2.8	-0.3

E

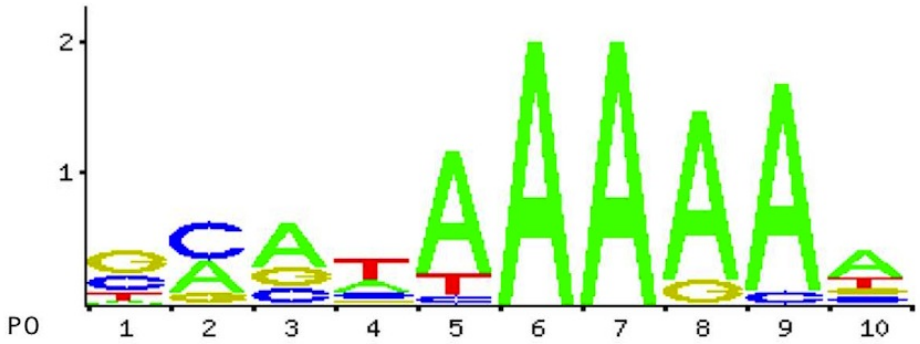


FIGURE 1.2: Construction de la matrice poids-positions de Hunchback à partir d'une collection de sites obtenus par expérience de DNAse1 footprinting disponible dans la base de données Jaspar - (identifiant de la matrice : P05084). A. Alignement de sites. B. Séquence consensus. C. Matrice de comptage en format tabulé. D. Matrice poids-positions en format tabulé. E. Logo. PO : positions des nucléotides.

Symbol	Nucleotide(s)	Description
A	A	Adenosine
C	C	Cytidine
G	G	Guanosine
T	T	Thymidine
R	= A or G	puRines
Y	= C or T	pYrimidines
W	= A or T	Weak hydrogen bonding
S	= G or C	Strong hydrogen bonding
M	= A or C	aMino group at common position
K	= G or T	Keto group at common position
H	= A, C or T	not G
B	= G, C or T	not A
V	= G, A, C	not T
D	= G, A or T	not C
N	= G, A, C or T	aNy

FIGURE 1.3: Code d'ambiguité IUPAC -

de détecter les meilleurs sites (en fonction du poids). À chaque itération, un score de vraisemblance de la matrice par rapport à un modèle de référence (modèle de Markov) est calculé. Pour chaque position des séquences d'entrée, on calcule un score de vraisemblance d'appartenir soit au motif soit au modèle de fond. Une nouvelle matrice est alors construite en collectant les sites les plus vraisemblables. Les itérations continuent jusqu'à convergence vers un score localement maximal.

Le programme *CONSENSUS* repose sur un algorithme glouton : il crée toutes les matrices de taille w possibles à partir des deux premières séquences du jeu puis va retenir les matrices les plus informatives. À l'étape suivante, toutes les combinaisons possibles entre ces matrices et tous les sites de la séquence suivante sont calculées et comme précédemment, les matrices les plus informatives sont gardées et ainsi de suite jusqu'à que toutes les séquences aient été prises en compte. Cet algorithme est sensible à l'ordre des séquences en entrée.

Comme *MEME*, *info-gibbs* est basé sur l'optimisation du score de vraisemblance mais incorpore une composante stochastique, ce qui réduit le problème de convergence vers des maxima locaux. Un site ayant servi à construire la matrice est retiré et remplacé par un autre site tiré au hasard avec une probabilité proportionnelle à sa vraisemblance d'appartenir au motif. Ensuite, la matrice est actualisée et les scores des positions des séquences sont recalculés. ces deux étapes sont répétées jusquà ce que le critère de convergence soit réalisé ou que le nombre de cycles maximum soit atteint. Cette méthode permet d'éviter les optimums locaux mais n'est pas exhaustive.

Les méthodes basées sur le comptage de mots sont présentées dans la figure 1.4. Le principe de ces méthodes et de compter les occurrences de chaque oligonucléotide d'une taille donnée et

de détecter ceux qui ont une fréquence exceptionnellement élevée (ou faible, selon les options choisies) (programmes *oligo-analysis* et *dyad-analysis* de la suite logicielle RSAT). Les mots les plus fréquents ne sont pas forcément les plus pertinents puisqu'ils reflètent les biais de composition des séquences analysées (par exemple une richesse générale en A+T). Afin de détecter les mots surreprésentés, il faut comparer les fréquences obtenues à des fréquences attendues au hasard. Pour définir un modèle de background, deux stratégies alternatives peuvent être utilisées. La première consiste à estimer un modèle de Markov sur base des séquences analysées elles-mêmes. Le principe consiste à calculer la fréquence attendue d'un mot de taille k (la longueur des oligonucléotides analysée) en fonction de la composition des séquences en mots de taille plus petite. Alternativement, on peut utiliser un jeu de séquences de référence (ex. l'ensemble des promoteurs d'un organisme lors de l'étude de promoteurs de gènes co-exprimés) à partir du quel on calcule la fréquence attendue de chaque oligonucléotide de taille k. La significativité de la sur-représentation est calculée sur base d'une distribution binomiale. Afin de ne pas surestimer la fréquence des mots auto-chevauchants ('GGGGGG', 'TATATATA', etc.), on ne compte que les occurrences renouvelantes. La p-valeur binomiale est alors calculée puis multipliée par le nombre de tests effectué (le nombre d'oligonucléotides analysés) pour obtenir un e-valeur. La significativité est alors calculée en prenant le logarithme en base 10 de cette e-valeur ($sig = -log_{10}(e-valeur)$), ce qui permet d'apprécier directement la surreprésentation du motif.

Le programme *position-analysis* (figure 1.4B) estime le biais positionnel des occurrences de chaque motif en comparant le nombre d'occurrences observées par rapport à un point de référence (milieu, début, fin de séquences) au nombre d'occurrences attendues selon une distribution homogène.

Enfin, le programme *local-word-analysis* (figure 1.4C) combine les approches de *oligo-analysis* et *position-analysis* en appliquant le test binomial pour détecter des mots exceptionnels dans des fenêtres de tailles fixes ou variables. Comme précédemment, les occurrences chevauchantes sont écartées du comptage.

La découverte de motifs retournant des matrices permet d'apprécier la variabilité de certaines positions des séquences reconnues par un FT, alors que les mots ne le permettent pas. Cependant, les algorithmes de découverte de mots détectent régulièrement des mots surreprésentés chevauchants, qui révèlent différents fragments d'un même motifs. Par ailleurs, il est fréquent d'obtenir des mots qui diffèrent par une seule lettre, reflétant les positions variables. Les mots retournés par *oligo-analysis*, *position-analysis* ou *local-word-analysis* peuvent être

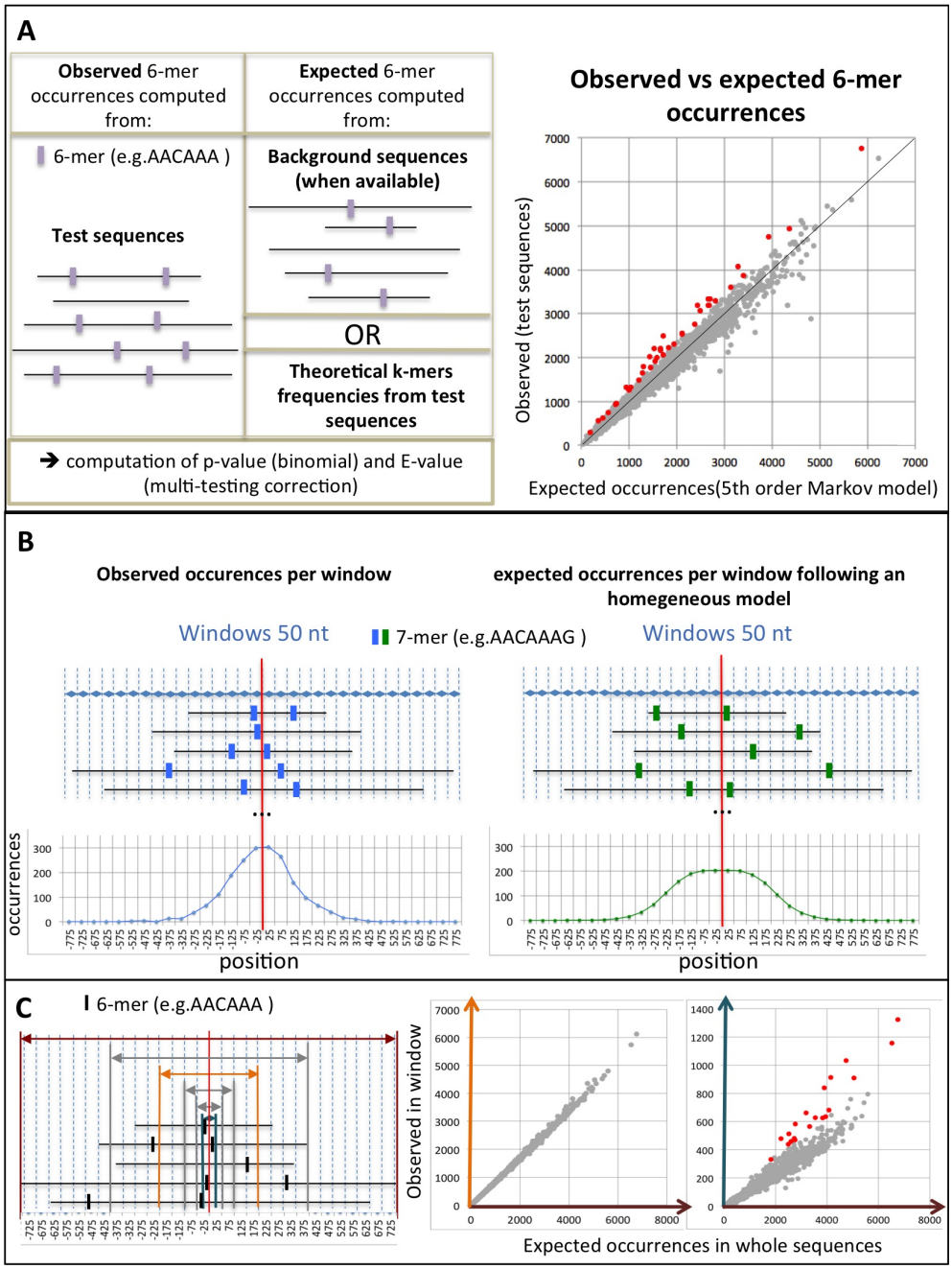


FIGURE 1.4: Méthodes de découverte de motifs - A. *oligo-analysis*. Gauche : vision schématique du principe du test de sur-représentation pour un mot donné. Droite : Nombre d'occurrences observées pour chaque mot dans le jeu test (axe des ordonnées) comparé au nombre d'occurrences attendues selon le modèle de référence (axe des abscisses). Chaque point représente un mot et les mots significatifs sont indiqués en rouge (significativité binomiale >= 10). B. *position-analysis*. Gauche : Les séquences sont alignées par rapport à leur centre. Les occurrences de chaque mot sont comptées dans des fenêtres non chevauchantes de taille fixe. Droite : Distribution homogène des occurrences. C. *local-word-analysis*. Gauche : Le nombre d'occurrences dans des fenêtres centrées sur le milieu des séquences est comparé au nombre attendu si la distribution était homogène.

assemblés (*pattern-assembly*, figure 1.5A) afin de construire des matrices de significativité (figure 1.5B). Les séquences sont scannées afin d'obtenir des matrices de comptage en appliquant un seuil sur le poids des sites retenus pour sa construction avec l'algorithme *matrix-from-patterns* (figure 1.5). Ici le poids de chaque site est calculé par rapport à la fréquence de bases dans le jeu de séquences testées. Par défaut, le poids minimum accepté pour que le site participe à la construction de la matrice est de 7. La séquence 1 (Seq1) de la figure 1.5G représente le site ayant le score maximum qui peut être obtenu. Nous pouvons noter que les matrices de significativité formées avec des mots faiblement significatifs ne détecteront aucun site et seront ainsi écartées. La séquence 2 (Seq2) est un exemple de site ayant un score inférieur au seuil de détection et qui ne sera donc pas utilisé pour la construction de la matrice de comptage.

1.1.2.5 Prédiction de sites

Nous pouvons prédire des sites à partir de motifs de liaison connus ou découverts sous forme de mots ou sous forme de matrices. Afin de prédire les sites correspondant aux motifs recherchés, il faut calculer un score à chaque position testée qui indiquera la qualité du site testé. Pour les motifs sous forme de mots, le score consistera à compter le nombre de substitutions qu'il faut appliquer au site pour correspondre au motif. Pour la prédiction à partir de matrices, le score de chaque site testé correspondra à la somme des poids des nucléotides présents dans la matrice. Dans le cas de la construction de matrices à partir d'assemblage de mots surreprésentés, le score du site correspondra alors la somme des significativités. Le scan des séquences avec l'une ou l'autre représentation des motifs va retourner un nombre important de faux positifs. Nous avons vu précédemment, dans l'exemple de la figure 1.2B, que les séquences consensus n'étaient pas efficaces pour la détection de sites et Day et McMorris (1992) (17) ont montré que le choix de la méthode de génération de séquences consensus devait être fait cas par cas suivant la question posée. Je ne parlerai ici que de la prédiction de sites à partir de matrices.

J'ai déjà un peu abordé ce sujet lorsque j'ai parlé de la transformation d'alignement de mots en matrice de comptage (figure 1.5). Durant cette étape, les sites étaient retenus si ils avaient un score au moins de 7 en utilisant un modèle de référence de Bernouilli qui assume l'indépendance entre les positions successives. Cependant, nous recherchons des séquences particulières, l'utilisation d'un modèle se basant sur la probabilité de dépendance entre les positions est alors plus appropriée. Pour cela, il convient d'utiliser des modèles de markov permettant de calculer la probabilité de trouver un nucléotide donné sachant la séquence précédente. Ainsi le poids

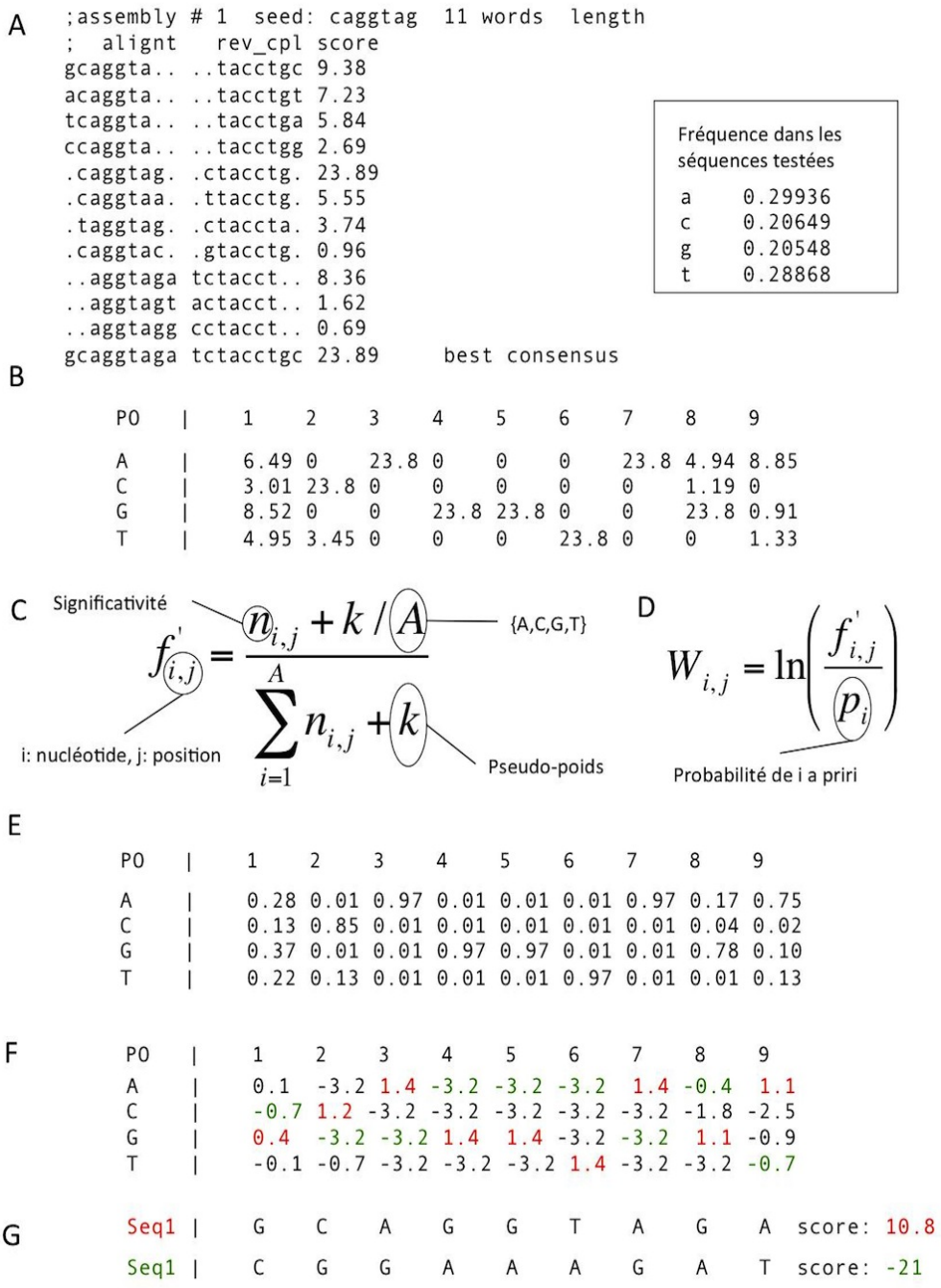


FIGURE 1.5: Transformation d'un alignement de mots chevauchant en matrices avec *matrix-from-patterns* - A. alignement de mots découverts chevauchants dont la significativité est indiqué dans la troisième colonne. B. Matrice de significativité. C. Calcul de la fréquence corrigée avec un pseudo-poids. D. Calcul du poids de chaque nucléotide i à chaque position j. E. Matrice poids-position. G. Alignement de deux séquences Seq1 et Seq2 avec la matrice, avec les scores correspondants en rouge et vert. Le score est calculé en additionnant les nombres de même couleur dans la matrice (significativités des nucléotide alignés).

calculé à chaque position est dépendant des positions précédentes. Le poids du site correspond alors au produit des probabilités de chaque position (figure 1.6B). La figure 1.6C montre la distribution des scores obtenus par la matrice de Hb sur la séquence de 2 Kb en amont du TSS du gène even-skipped (eve), qui contient trois des sites ayant servi à construire la matrice (étoiles rouges).

L'utilisation d'un seuil sur le poids est délicat lors la recherche de sites, en particulier si l'on cherche par la suite des régions enrichies en sites. En effet, les matrices courtes ou avec peu de positions informatives auront tendance à détecter des sites avec de plus faibles scores que des matrices plus longues. Il est donc préférable de se baser sur des critères statistiques s'appuyant sur la distribution théorique des scores de chaque matrice (*matrix-distrib*). Ainsi nous pouvons limiter le nombre de faux sites (faux positifs) détecter par la matrices. La figure 1.7 montre la distribution théorique des scores de deux matrices. Nous pouvons voir que les deux matrices ont des comportements très différents. En effet, la matrice Hb (figure 1.7A) atteint un score maximal de 7.9, alors que la seconde matrice (figure 1.7B) atteint un score maximal de 12.7. Mais plus important, si l'on considère un score de poids de 5, nous pouvons voir que la p-valeur correspondante varie d'un facteur proche de 1000. Une p-valeur de 10^{-3} indique que l'on attend un faux positif toutes les 1000 positions. Il est donc préférable d'utiliser un seuil sur la p-value plutôt que sur le poids. Plusieurs algorithmes de détection de motifs sont disponibles gratuitement, dont *patser* (20), *MotifLocator* (24), et *matrix-scan* (25). Ces outils ont fait l'objet d'une comparaison (26) qui est présentée dans la table 1.1. Pour la détection de sites uniques, *matrix-scan* et *MotifLocator* utilise des modèles de Markov comme modèle de référence pour le calcul des poids des sites, *matrix-scan* fournit en plus la p-valeur associée. Quant à *patser*, le poids des sites et leur p-valeur associée sont calculés mais le modèle de référence implique lindépendance entre les sites (modèle de Bernouilli).

Lors de la prédiction de sites, l'application de seuils statistiques pour limiter le nombre de faux positifs permet d'identifier des sites probablement reconnus par les facteurs de transcription in vitro. Cependant, cela n'implique pas que les sites identifiés soient vraiment fonctionnels *in vivo* ("*Futility theorem*" (27)). L'utilisation de critères discriminants est alors nécessaire. Deux critères, non exclusifs, peuvent être utilisés : la conservation de séquence entre espèces phylogénétiquement proches, ou "empreinte phylogénétique" ("*phylogenetic footprinting*" en anglais), et la colocalisation de sites de fixation de FT au sein de des régions relativement courtes, les "modules cis-régulateurs" (CRM). Le deuxième critère sera détaillé dans la section suivante.

A

PO	1	2	3	4	5	6	7	8	9	10
A	0.1	0.4	0.5	0.2	0.8	1.0	1.0	0.8	0.9	0.5
C	0.3	0.5	0.2	0.2	0.1	0.0	0.0	0.0	0.1	0.1
G	0.5	0.1	0.2	0.1	0.0	0.0	0.0	0.1	0.0	0.1
T	0.1	0.0	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.2

B

$$P(S \mid M) = \prod_{j=1}^{w} f'_{r_j,j}$$

P(S|M) probabilité du site S d'être une instance générée par la matrice.

$$P(S \mid B) = \prod_{j=1}^{w} p_{r_j}$$

P(S|B) probabilité du site S d'être une instance générée par le modèle de référence.

$$W_S = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right)$$

W_s poids: log ratio des deux probabilitités ci-dessus

C

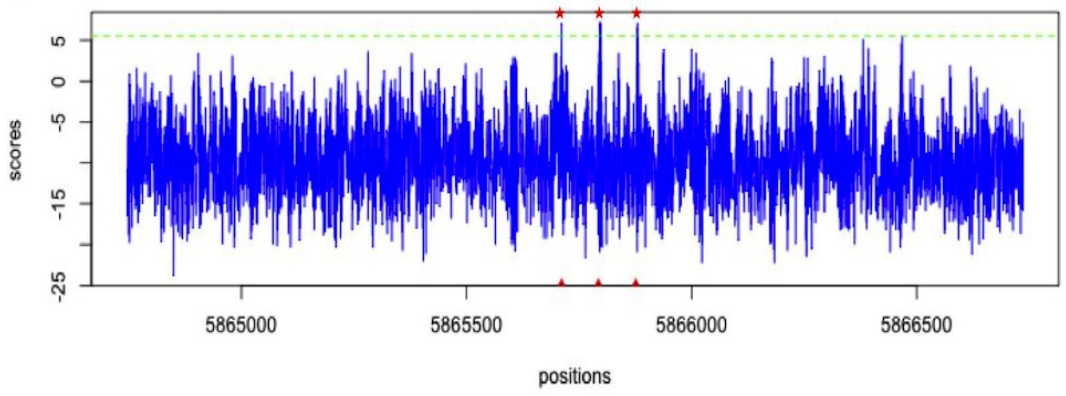


FIGURE 1.6: Calcul et distribution de scores de sites attribués en considérant une matrice donnée - A. Matrice de fréquences corrigées de Hunchback obtenue à partir d'une collection de sites provenant d'expérience de DNAse1 footprinting sur le promoteur de eve (cf. 1.2). B. Calcul du poids d'un site étant donné un modèle de référence. C. Distribution des scores obtenus pour chaque position de la séquence de longueur 2kb en amont du TSS du gène eve. l'axe des abscisses indique les positions génomiques. Les étoiles indiquent les sites identifiés expérimentalement. La ligne verte pointillée indique un seuil de p-valeur de 10^{-3}.

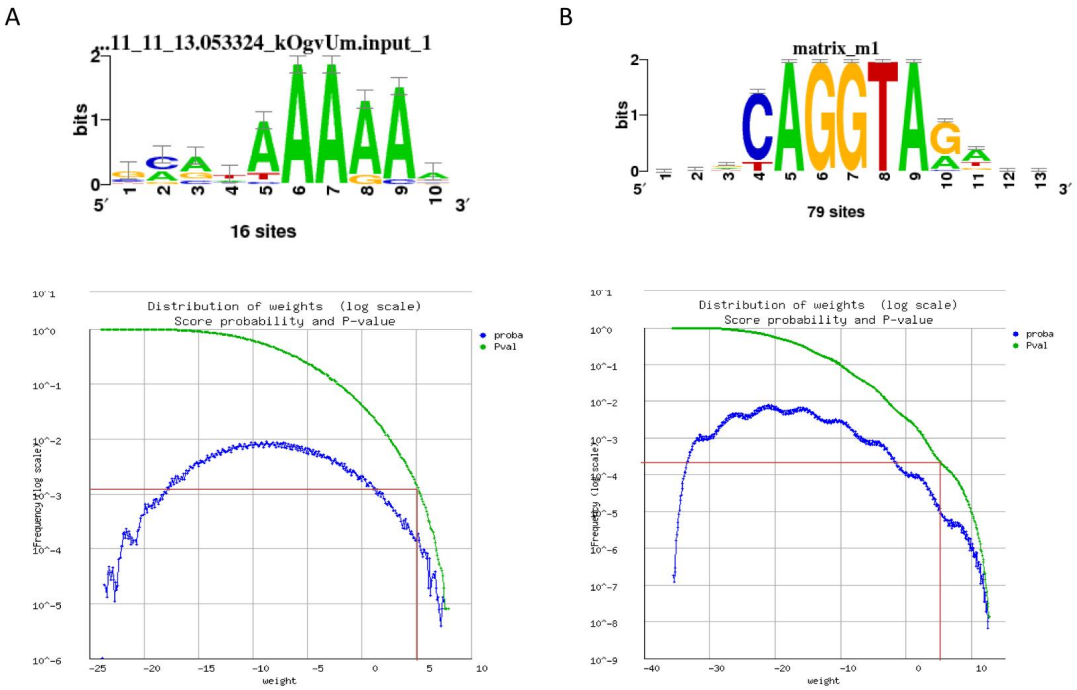


FIGURE 1.7: Distribution théorique des scores de probabilité de la matrice Hb (A) et d'une matrice construite à partir de l'alignement de mots chevauchants sur-représentés dans un jeux de données de drosophile (B) - La distribution des scores théorique a été calculée par *matrix-distrib*. L'ordonnée indique la probabilité d'obtenir chaque score possible avec la matrice étudiée (transformé en log, courbe bleue) et la p-valeur associée (courbe verte). La ligne rouge indique un score de 5.

La notion d'empreinte phylogénétique est basée sur l'hypothèse que les séquences régulatrices évoluent plus lentement que les régions non codantes voisines, à cause de la pression de sélection. Les algorithmes se basent sur l'hypothèse que les séquences orthologues sont soumises à des pressions sélectives similaires. Pour détecter les régions conservées (significativement similaires), les séquences doivent être alignées. Cette étape n'est pas triviale et de nombreux algorithmes ont été développés pour réaliser des alignements locaux ou globaux (voir Wasserman et Sandelin (27) pour une revue). Ces approches permettent de diminuer considérablement le taux de faux positifs en ne diminuant que faiblement la sensibilité ((28)). Bradley et collaborateurs (29) ont étudié les localisations génomiques des sites liés par six facteurs de transcription impliqués dans la segmentation antéo-postérieure de l'embryon de *D. melanogaster* et les localisations des sites liés par les facteurs orthologues chez *D. yakuba*. Ces espèces ont divergé il y a environ 5 millions d'années, ce qui est relativement court. Cette étude a mis en évidence des pertes, des gains et des changements de localisation de sites de liaisons des FT. La sélection des sites par empreinte phylogénétique présente donc des limites. Dans les travaux menés durant cette thèse, j'ai choisi de ne pas appliquer ce critère, mais de visualiser la conservation des CRM prédits en fin de processus, en utilisant les fonctionalités du navigateur de génome de l'UCSC (voir plus loin).

1.1.2.6 Prédiction de modules cis-réguleurs

Plusieurs facteurs de transcription agissent généralement de concert pour réguler la transcription de gènes, en se liant sur des régions proximales ou distales, appelées "modules cis-régulateurs" (CRM) (30, 31). Nous pouvons identifier de telles régions en recherchant des enrichissements en sites de liaison de FT (soit du même facteur, on obtient alors des CRM homotypiques ; soit de différents facteurs, on obtient alors des CRM hétérotypiques). De nombreux algorithmes ont été développés afin de détecter les CRM. Certains d'algorithmes utilisent la conservation entre espèces afin de prédire des CRM fonctionnels (Modulefinder (32), Target Explorer (33) ; TFBScluster (34), eCis-analyst (31, 35), etc. D'autres algorithmes sont basés sur la significativité d'enrichissement en sites par rapport à un modèle de référence calculé sur les séquences à analyser (ClusterBuster (30) et *matrix-scan* (25)).

Dans une revue récente, Van Loo et al. (36) ont mis en évidence trois types d'algorithmes de détection de CRM et résument les particularité d'une trentaine d'outils. Brièvement, la première classe d'algorithmes, appelée "*CRM scanners*", est basée sur l'utilisation d'un modèle de CRM prédéfini : jeu de facteurs de transcription connus pour être impliqués dans un même

processus biologique, jeu de séquences et d'autres paramètres suivant l'outil considéré, telle qu'une taille prédéfinie de CRM, un nombre et/ou un ordre des sites, la conservation entre espèces, etc. La seconde classe regroupent les algorithmes de type "*CRM builders*", en général appliqués à un jeu de gènes co-exprimés (ou co-régulés) et détectent des CRM similaires (e.g. sites prédits à partir d'un même jeu de matrices). Enfin, la troisième classe regroupe les algorithmes "*CRM screeners*", qui n'utilisent aucune connaissance a priori et recherchent des régions enrichies en sites prédits à partir d'un jeu de matrices. Les deux premières approches sont complémentaires, puisque les modèles de CRM construits par les CRM builders peuvent être utilisés par les CRM scanners pour détecter des CRM similaires supplémentaires. Les CRM screeners ne permettent pas d'inférer une fonction pour les CRM prédits, ce qui conduit Van Loo et collaborateurs à conclure que cette classe de méthodes est moins performante que les deux autres classes. Cependant, en l'absence d'information sur l'expression génique ou sur les mécanismes de régulation, les CRM screeners sont les seuls outils utilisables.

Turatsinze (26) a comparé de nombreux outils de prédiction de sites et de CRM en regardant différents paramêtres (possibilité d'utiliser une ou plusieurs matrices à la fois, le choix du modèle de background pour la détection de sites, l'évaluation statistique des prédictions et l'utilisation de plusieurs génomes). La table 1.1 résume cette étude. Parmi les outils de détection de CRM comparés, *matrix-scan* (intégré dans la suite d'outils RSAT) est le seul qui permet l'utilisation de modèle de Markov d'ordre supérieur pour la détection des sites.

Durant ma thèse, j'ai ainsi sélectionné le logiciel *matrix-scan* pour prédire des CRM impliqués dans l'AGZ. La prédiction de CRM opérée par *matrix-scan* est basée sur la détection de régions enrichies en éléments cis-régulateurs ("*Cis-Regulatory element Enriched Region*" ou "CRER"), c'est à dire des régions qui contiennent une densité en motifs significativement supérieure à celle attendue au hasard. Un CRER est obtenu en comptant le nombre de sites individuels obtenus au sein d'une fenêtres de taille variant entre une taille minimale et une taille maximale définies par l'utilisateur (e.g. 30pb - 800pb). La significativité de l'enrichissement est calculée en appliquant le test binomial. La statistique binomiale repose sur une hypothèse de probabilité constante de détecter un site tout au long de la fenêtre analysée. La p-valeur obtenue est alors directement interprétable en termes de risque de faux positifs, et l'utilisateur peux donc choisir un seuil sur la p-valeur pour contrôler ce risque (cf. section 1.1.2.5). La statistique binomiale repose également sur l'indépendance entre les tests, ce qui peut être problématique si les motifs utilisés sont des motifs répétés (e.g. GAGAGA, CAGCAGCAG etc.). Ce problème

TABLE 1.1: Comparaison de programmes de détection d'éléments et modules cis-réguleurs. Les signes + et - indiquent la présence ou l'absence d'un paramètre dans un programme. D'après la thèse de Turatsinze (26)

Comparison of main features of pattern matching programs							
Program name	Individual TFBS	Multiple PSSM	CRM prediction	Adaptive background	High-order Markov models	P-value computation	Required multi-genomes sequences input
MotifLocator	+	−	−	−	+	−	−
MATCH	+	−	−	−	−	−	−
patser	+	−	−	−	−	Bernoulli	−
TargetExplorer	−	+	−	−	−	−	−
CAST	−	+	+	−	−	Bernoulli	−
Cluster-Buster	+	+	+	Bernoulli	−	Bernoulli	−
CIS-ANALYST	+	+	+	−	−	+	−
ModuleFinder	+	+	+	−	−	−	+
TFBScluster	+	+	+	−	−	−	+
ModuleMiner & ModuleSearcher	+	+	+	−	+	−	−
MOODS	+	+	−	−	−	Bernoulli	−
matrix-scan	+	+	+	Markov	+	Markov	−

été résolu en imposant une distance minimale entre les débuts de chaque position testée lors de la prédiction de sites. Ainsi les occurences chevauchantes sont écartées.

L'utilisation de *matrix-scan* se situe dans la continuité des analyses de séquences régulatrices menées durant cette thèse : récupération des séquences non codantes d'intérêt, découverte de motifs et construction de modèles de référence adaptés, prédiction et visualisation d'éléments et modules cis-régulateurs. La modularité de la suite RSAT et la possibilité d'utiliser les différents outils en ligne de commande m'ont permis d'établir des schémas d'analyses complets ("*workflows*"), où chaque étape a pu être contrôlée et chaque paramètre choisi avec soin.

1.1.2.7 Bases de données

Plusieurs bases de données mettent à disposition, de façon publique ou commerciale, des jeux de motifs de laison de FT, la plupart du temps sous forme de matrices occurrence-position [1] (voir tableau 1.2). Il existe deux bases de donées générales contenant des motifs de liaison pour plusieurs organismes.

1. Les matrices obtenues à partir de sites provenant d'expériences de type SELEX sont plus souvent représentées sous forme de matrices de fréquences relatives (plutôt que d'occurrences) pour masquer le grand nombre de séquences obtenues par l'étape d'amplification.

TRANSFAC (37), qui est une base de données commerciale avec un accès public à certaines données, qui rassemble 1308 matrices (version 2010.1) dont 398 sont publiques, appartenant à six grands groupes taxonomiques (vertébrés, plantes, insectes, nématodes, champignons, bactéries). Les insectes sont peu représentés dans cette base puisque seulement 68 matrices sont disponibles (dont 38 publiques). Le format de matrices fourni par TRANSFAC est très informatif. En effet, différents champs sont renseignés en plus de la matrice elle-même. Les principaux champs sont : un identifiant unique indiquant le groupe taxonomique, le nom du facteur se liant au motif et la qualité des sites qui dépend de leur provenance ; un numéro d'accession ; une description courte du facteur ; les sites à partir desquels la matrice a été construite (pas toujours renseigné) ; et toute une série de champs renseignant sur la publication (auteurs, titre) d'où proviennent les donnée. Les sites proviennent préférentiellement d'expérience de DNAse1 footprinting.

JASPAR (38) est une base de données publique regroupant également des motifs provenant de plusieurs groupes taxonomiques. Elle contient 457 motifs dont 123 correspondent à des facteurs de drosophile. La source des sites utilisée est variée et les motifs ont été générés avec l'algorithme *MEME*.

D'autres bases de données spécialisées regroupent des motifs provenant d'espèces particulières. Durant mes travaux, j'ai surtout utilisé les bases de données spécialisées pour la drosophile. Fly Factor Survey (39) et DMMPMM (40) utilisent des sites provenant d'expériences diverses et exploîtent des algorithmes de découverte de motifs différents (*ChIP-Munk* et *MEME* ou *Consensus*, respectivement) afin d'en extraire les motifs. Des bases spécifiques pour d'autres organismes sont disponibles (RegulonDB (41) pour Escherichia coli par exemple).

1.1.3 Les co-régulateurs

La régulation du recrutement de la machinerie basale de transcription au TSS par les TF est possible grâce à des complexes protéiques, appelés co-régulateurs, permettant soit la transduction des signaux entre les TF et la machinerie basale, soit une conformation de la chromatine favorable.

1.1.3.1 Le complexe médiateur

Le complexe médiateur a été identifié chez la levure par Kelleher et collaborateurs au début des années 1990 (51). Conservé chez les eucaryotes, ce complexe est constitué de plusieurs

TABLE 1.2: Description de bases de motifs de liaison générales (TRANSFAC, JASPAR) et spécialisées pour la drosophile (Fly Factor Survey, DMMPMM). Les nombres entre parenthèses indiquent le nombre de motifs publics dans la base de données TRANSFAC. HMM : Hidden Markov Model ; B1H : Bacterial one-hybrid. 1. PAZAR, http ://www.pazar.info/, Portales-Casamar (2007) (42, 43) ; 2. FlyReg, http ://www.flyreg.org/, Bergman et al. (2004) (44) ; 3. Bergman Lab, http ://bergmanlab.smith.man.ac.uk/, Down et al. (2007) (45) ; 4. RedFly, http ://redfly.ccr.buffalo.edu/, 5. Halfon et al. (2008) (46) ; 6. Noyes et al. (2008) (47), 7. Noyes et al. (2008) (48) ; 8. BDTNP, bdtnp.lbl.gov/, (49, 50).

Base de données	Espèce(s) / Taxon(s)	Nombre de matrices	Autres informations
TRANSFAC 7.0 Version 2010.1	Vertébrés	913 (293)	- sites compilés depuis la littérature, préférence pour les expériences DNAse1 footprint - familles de motifs (HMM) - format de matrices descriptif
	Champignons	194 (31)	
	Plantes	124 (31)	
	Insectes	68 (38)	
	Nématodes	7 (4)	
	Bactéries	2 (1)	
	Total	1308 (398)	
JASPAR 4.0	Vertébrés	130	- sites : compilés depuis la littérature (DNAse1), SELEX, B1H, ChIP-chip, ChIP-seq, familles (PAZAR,[1]) - découverte de motifs avec MEME
	Champignons	177	
	Plantes	21	
	Insectes	123	
	Nématodes	5	
	Urochordés	1	
	Total	457	
Fly Factor Survey Version 09/05/11	*Drosophila melanogaster*	470	- sites : DNAse1 footprint and SELEX[2,3,4], B1H[6,7] - découverte de motifs avec MEME ou Consensus - redondance
DMMPMM/iDMMPMM	*Drosophila melanogaster*	41/39	- sites : DNAse1 footprint[2], SELEX[2,5], B1H[4], ChIP-chip[8] - découverte de motifs avec ChIPMunk

sous-unités protéiques (30 chez l'homme, 21 chez la levure et au moins 25 chez la drosophile) organisées en trois modules distincts : module "tête", module "milieu" et module "queue". Le module "tête" intéragit directement avec l'ARNpolII, le module "queue" représente une plateforme d'interaction avec les FT et le module "milieu" transmet le signal via deux sous modules med9 (requis pour l'activation) et med10 (requis pour la répression) (52). Il existe plusieurs complexes médiateurs dont la composition varie en fonction du FT qui le recrute et du type cellulaire. Par exemple, chez la drosophile, le complexe médiateur contenant la sous-unité MED31 n'est pas requis pour la régulation globale de la transcription mais plutôt pour une régulation adaptée des gènes spécifiques de la segmentation au tout début du développement embryonnaire (53).

1.1.3.2 Les régulateurs de la chromatine

Dans le noyau, l'ADN est enroulé autour de complexes d'histones formant ainsi la chromatine. La chromatine peut arborer différents niveaux de compaction. Luger et collaborateurs (54) ont défini la structure du premier niveau de compaction consistant en 146 pb d'ADN entourant un octamère d'histones et appelé "nucléosome". Les nucléosomes sont les déterminants primaires de l'accessibilité de l'ADN. Les octamères d'histones sont composés de deux copies du tétramère H2A-H2B/H3-H4. Les extrémités N- et C- terminales (ou queues) des histones passent entre les tours d'ADN étant ainsi accessibles pour interagir avec d'autres protéines. Les interactions ADN/histone ou histone/histone dépendent du repliement de chaque histone qui peut-être modifié par l'acétylation, la méthylation ou la phosphorylation d'acides aminés spécifiques. La stabilité de ses intéractions peut être également altérée par l'action de complexes de remodelages dépendant de l'ATP qui permettent le glissement des nuclésomes. Ce niveau de compaction permet un changement rapide entre l'état répressif et actif de la chromatine. La chromatine est alors appelée euchromatine. Le niveau de compaction supérieur implique la reconnaissance de certaines modifications d'histones par les protéines HP1 (hétérochromatin protein 1) (55). À ce niveau la chromatine est appelée hétérochromatine. La compaction de la chromatine affecte toutes les étapes de la transcription depuis la liaision des FT et la formation du PIC jusqu'à l'élongation. La compaction de la chromatine est un processus dynamique et implique de nombreux facteurs. L'impact de différentes modifications d'histones sur l'accessibilité de la chromatine sera présenté dans le chapitre 4 de ce manuscrit. J'y détaillerai également l'action de différents facteurs.

1.2 L'activation du génome zygotique (AGZ) durant l'embryogenèse de la drosophile

Durant la première moitié du 20^{me} siècle, les biologistes ont effectué des observations anatomiques détaillées du développement embryonnaire de plusieurs espèces. L'embryologie comparée a mis en lumière de nombreuses similarités dans les étapes du développement précoce des embryons. Le développement des métazoaires passe toujours par une étape où l'embryon est formé de trois feuillets, la gastrula comprenant l'ectoderme, l'endoderme et le mésoderme, nettement caractérisés, dont dérivent des parties analogues chez tous les animaux. L'utilisation d'organismes modèles est alors pertinente pour étudier ces différents phénomènes. En analysant des mutants de drosophile, Ed Lewis mis en évidence le rôle important de certains gènes au cours du développement (56). Vers la même époque, Erik Wieschaus et Christianne Nüsslein-Volhard ont caractérisés 120 gènes essentiels à une segmentation normale (57). L'embryon précoce des métazoaires connaît des changements morphologiques et une reprogrammation transcriptionnelle drastiques. La compréhension de ces phénomènes reste encore un enjeu fondamental et la drosophile semble être un bon système modèle pour leur étude. Durant ce chapitre, je vais décrire les mécanismes impliqués particulièrement dans la reprogrammation transcriptionnelle de l'embryon, ou "transition maternelle-zygotique".

1.2.1 L'embryogénèse précoce chez les métazoaires : mécanismes communs

1.2.1.1 Activation de l'œuf

Au cours de la phase dite "d'activation", l'œuf initie la synthèse de protéines et d'ADN, et subit des changements structuraux au niveau du cortex et du cytoplasme. Cette activation se traduit par une augmentation de la concentration en calcium intracellulaire par vagues, qui initie des cascades de transduction de signaux induisant la reprise et la fin de la méiose (jusque là le cycle cellulaire était en pause à un stade dépendant de l'espèce), l'activation des processus métaboliques, des réarrangements du cytosquelette, la régulation de la traduction des ARNm (par blocage ou dégradation), et enfin la régulation de la synthèse d'ADN.

1.2.1.2 Transition mid-blastuléenne (TMB)

Chez les métazoaires, après la fécondation, les cycles mitotiques sont rapides et dépourvus de phases gap. La TMB correspond à un allongement de ces cycles par l'introduction des

phases gap.

1.2.1.3 Transition maternelle-zygotique (TMZ)

La TMZ est le passage du contrôle du développement de la mère vers l'enfant. Chez les métazoaires, mais aussi chez les plantes, ARNs et protéines sont synthétisés et déposés dans l'œuf durant l'oogenèse. Ces produits "maternels" assurent le contrôle du développement embryonnaire depuis la fécondation jusqu'à l'activation transcriptionnelle du zygote (figure 1.8). En effet, le génome zygotique est silencieux durant les premières divisions mitotiques. L'activation du génome se passe en deux phases, une première phase dite "mineure" (bleu clair, figure 1.8) et une seconde dite "majeure" (bleu foncé, figure 1.8). Alors que les nouveaux ARNs "zygotiques" sont synthétisés, les ARNs maternels sont dégradés. L'embryon prend alors le contrôle de son développement.

1.2.2 L'embryogénèse précoce chez *Drosophila melanogaster*

1.2.2.1 Transition mid-blastuléenne (TMB)

Chez la drosophile, dès la fécondation, l'embryon présente deux zones identifiables : le vitellus, sombre et opaque, qui occupe quasiment la totalité de l'embryon, et le périplasme, plus clair, formant une couche périphérique. Les clivages mitotiques des noyaux ne sont pas accompagnés de cytokinèse (division du cytoplasme). L'embryon se développe donc sous forme d'un syncytium (cytoplasme contenant plusieurs noyaux) durant 13 cycles. Les dix premières divisions mitotiques durent environ huit minutes et sont synchrones. Dès le cycle 8, les noyaux jusque là répartis de façon homogène dans le vitellus, entament une migration radiale vers le périplasme. Les noyaux à la périphérie de l'embryon forment le blastoderme syncytial au cycle 10. Cette transition est marquée par la cellularisation des cellules polaires (précurseurs de la lignée germinale). À partir du cycle 7, on observe un ralentissement progressif des divisions nucléaires (10 minutes pour le cycle 10 jusqu'à 20 minutes au cycle 13) qui se produisent maintenant de façon métasynchrone (différences de vitesse de division entre les noyaux aux pôles (rapides) et dans le reste de l'embryon (59). Ce ralentissement est dû à l'activation des points de contrôle de la réplication assurée par les protéines Chk2 et Chk1 et à la dégradation des ARNm cycline B (59, 60). L'introduction des phases gap au cours du cycle 14 entraîne un ralentissement important des mitoses en allongeant considérablement l'interphase. Cette pause permet la cellularisation des noyaux périphériques par invagination de la membrane plasmique

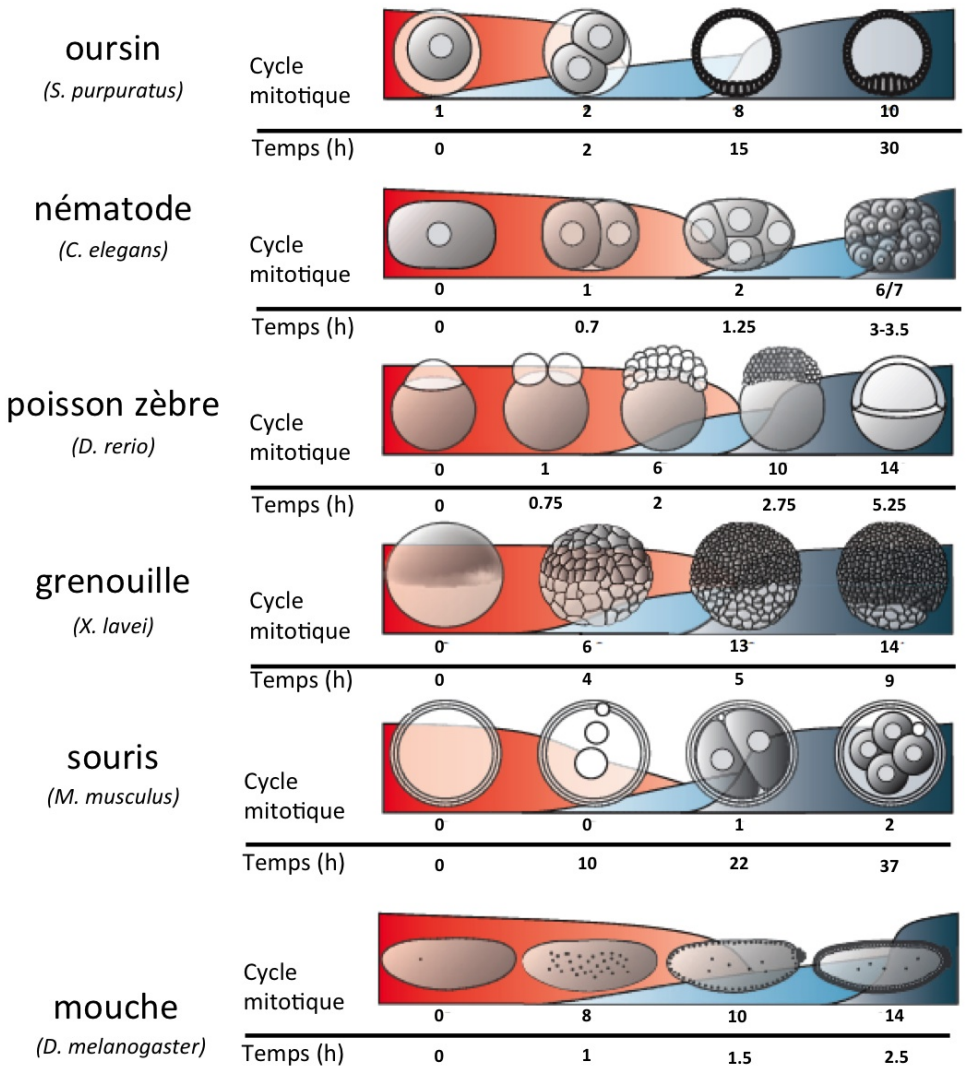


FIGURE 1.8: Vue comparative de l'embryogenèse précoce chez les métazoaires. - Représentation schématique des étapes embryonnaires clés pour plusieurs organismes modèles (oursin, nématode, poisson zèbre, grenouille, mouche et souris) précisant les cycles mitotiques et le temps après la fécondation correspondant. Quantités d'ARNm maternels (profil rouge) et d'ARNm nouvellement synthétisés lors des vagues mineure (profil bleu clair) et majeure (profil bleu foncé) d'activation du génome zygotique (AGZ). Le temps 0h correspond à la fécondation. (Figure modifiée à partir de Tadros et al. (58)).

à la fin du cycle 14 (figure 1.9A). Le blastoderme cellulaire subira la gastrulation dès le début du cycle 15. Il est à noter que la TMB est le premier changement morphologique assuré par le génome zygotique.

1.2.2.2 Mise en place de la polarité de l'embryon

Durant la première heure de développement, le génome de l'embryon de la drosophile est silencieux (figure 1.9B) et le développement est donc assuré par les ARNm dits « maternels », accumulés dans l'oocyte durant l'oogenèse. Les gènes produisant ces ARNm sont appelés gènes à "effet maternel". Les gènes à effet maternels les plus étudiés sont ceux responsables de la mise en place des axes embryonnaires. La polarisation de l'embryon est initiée dès les premiers stades de développement de l'oocyte par quatre groupes de gènes : groupe A (spécification de la tête et du thorax), le groupe P (spécification de l'abdomen), le groupe T comprenant les gènes transcrits au pôle antérieur (acron), postérieur (telson) et dans les régions terminales respectivement, et enfin le groupe D impliqué dans la mise en place de l'axe dorso-ventral (DV). Le groupe T permet la différenciation des régions terminale mais n'entre pas directement dans la spécification des axes embryonnaires. La segmentation antéro-postérieure (AP) commence dans l'oocyte avec trois acteurs principaux gurken (groupe D), oskar (groupe P) et bicoid (groupe A) (figure 1.10). Produit par le noyau de l'oocyte, l'ARNm Gurken est déterminant pour la mise en place de l'axe AP durant les stades 2 à 6 du développement de l'oocyte permettant la différenciation des cellules terminales en cellules postérieures (figure 1.10A). Les cellules postérieures vont induire à leur tour la réorganisation des microtubulles (MTs) du cortex de l'oocyte, ce qui va entraîner la migration du noyau au futur coin antéro-dorsal (figure 1.10B). Une deuxième vague de production de Gurken par le noyau va permettre la spécification de la face dorsale en inhibant la synthèse de Pipe, qui va être sécrété par les cellules épithéliales de la face ventrale, ainsi que l'accumulation d'un produit encore non-identifié dans l'espace péri-vitellin (figure 1.10C). Pipe est le premier signal de ventralisation. Un second se produira après la fécondation. Ainsi l'axe DV est mis en place. Les MTs, polarisés sur l'axe AP, permettent le transport des ARNm bicoid et oskar grâce à leur association à des moteurs moléculaires (respectivement dynéines et kinésines). Les protéines Oskar, au pôle postérieur, permettent la localisation restreinte des protéines Nanos (groupe P) en évitant le blocage de leur traduction induit par Smaug dans le reste de l'oocyte (61). Les ARNm hunchback et caudal, tous deux déterminants dans la segmentation AP de l'embryon, sont également accumulés dans l'oocyte mais distribués de façon homogène (figure 1.11A).

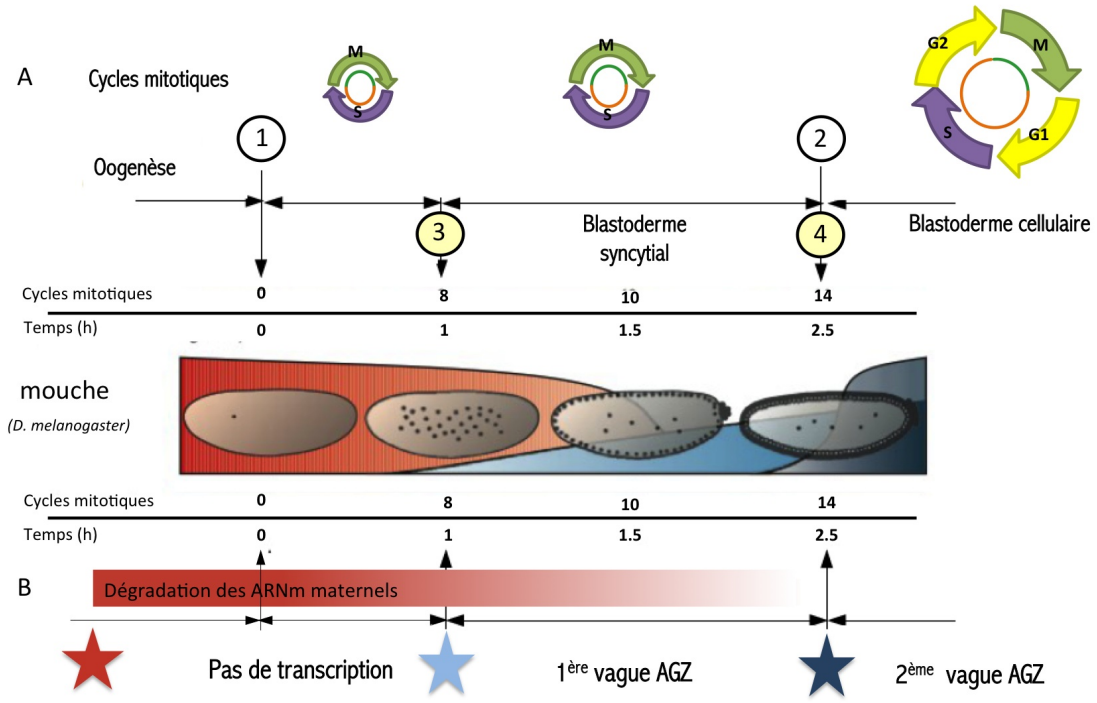


FIGURE 1.9: Développement précoce de l'embryon de la drosophile entre 0h et 3h après la fécondation. - A. La première ligne indique les phases des cycles mitotiques aux cours du temps. M : mitose (vert clair) ; S : réplication (violet) ; les phases gap G1 et G2 (jaune). Les temps de mitose et d'interphase sont indiqués en vert foncé et orange respectivement. **(1)** Fécondation, **(2)** Transition mid-blastuleenne, **(3)** début de migration des noyaux à la périphérie de l'embryon, léger ralentissement des cycles mitotiques, divisions méta-synchrones des noyaux, **(4)** cellularisation. Les axes horizontaux indiquent l'évolution temporelle des événements en fonction du temps en heures et du nombre de cycles mitotiques. L'image au centre est identique à la figure 1.8. B. Représentation schématique de la transition maternelle-zygotique. Les étoiles correspondent aux événements principaux influant quantitativement et qualitativement sur la présence des ARNm. L'activation de l'oeuf (étoile rouge) correspond au début de la dégradation des ARNm maternels ; la première et seconde vague d'activation de la transcription sont symbolisées par les étoiles bleu clair et bleu foncé respectivement. AGZ : Activation du Génome Zygotique.

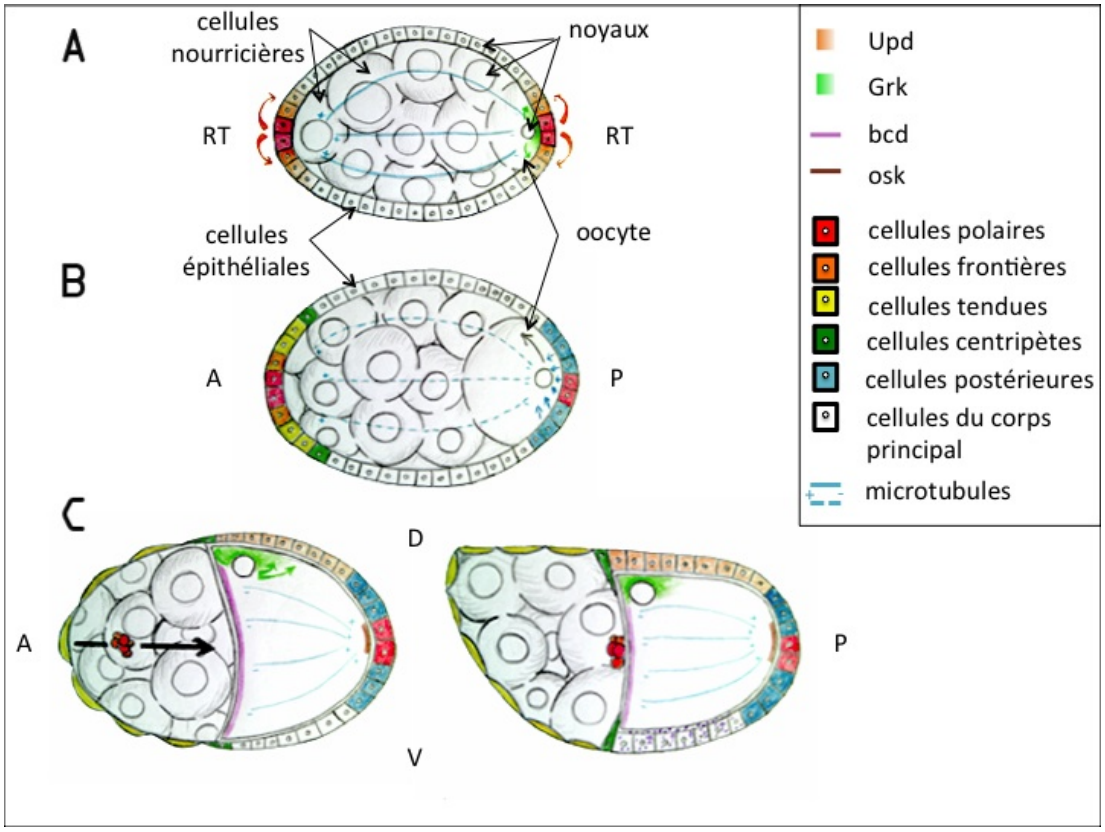


FIGURE 1.10: Mise en place des axes embryonnaires durant l'oogenèse. - (A) Stade 2-6 du développement de l'oocyte. Différents signaux moléculaires permettent la différenciation des cellules épithéliales de la région terminale (RT) : Upd en orange (voie JAK/STAT (62)), Gurken (Grk, vert clair). Le réseau de MT, résultant s'étend d'un pôle à l'autre du follicule permettant le transport des ARNm depuis les cellules nourricières (extrémité +) dans l'oocyte (extrémité -) (63). (B) Stade 7 du développement de l'oocyte : réorganisation des MT, différenciation des cellules RT en trois types (cellules frontières en orange, cellules tendues en jaune, cellules centripètes en vert) (64). (C) À gauche : début du stade 9, à droite : stade 10b. Stade 9 : migration des cellules polaires (flèche noire), spécification de l'axe dorso-ventral via Grk et qui induit la répression de la synthèse de Pipe (points violets) à la face dorsale de l'embryon. Déstabilisation et réorganisation des MTs du cortex de l'oocyte en un réseau orienté sur l'axe AP dont l'extrémité – se situe au pôle antérieur. Les ARNm de Bicoid (Bcd, rose) et Oskar (Osk, marron) vont être transportés de façon polarisée sur le réseau de MTs respectivement au pôle antérieur par les dynéines (+ -> -) et au pôle postérieur par les kinésines (+ -> -).

À la fécondation, ces ARNm peuvent être traduits. Les protéines Bicoid, qui sont des facteurs de transcription mais aussi des répresseurs de la traduction, vont diffuser en un gradient de concentration depuis le pôle antérieur et vont empêcher la traduction des ARNm caudal restreignant la synthèse des protéines Caudal au pôle postérieur (65). De façon opposée, la traduction des ARNm nanos induit un gradient postéro-antérieur des protéines Nanos, ces dernières, associées aux protéines Pumilio (groupe P), inhibent la traduction des ARNm hunchback au pôle postérieur (figure 1.11B et C).

La fécondation fournit aussi le second signal de spécification de l'axe DV. En effet, une cascade de phosphorylation va être initiée par l'activation des récepteurs Toll (groupe D) présents sur la face ventrale de l'embryon (produits déposés préalablement dans l'espace péri-vitellin). Cette cascade mène à la libération de la protéine Dorsal dans le cytoplasme, qui pourra alors être transportée dans les noyaux et former un gradient ventro-dorsal. Bicoid, Hunchback, Dorsal et Caudal sont des facteurs de transcription appelés aussi morphogènes car ils vont pouvoir activer ou réprimer leurs cibles en fonction de leur concentration.

1.2.2.3 Transition maternelle-zygotique (TMZ)

Durant la première vague d'AGZ, les gènes de segmentations sont activés par les facteurs maternels. Bosveld et collaborateurs (53) ont mis en évidence l'implication d'un complexe médiateur comportant spécifiquement la sous-unité dMED31, synthétisée au niveau maternel, dans l'activation des gènes gap et pair-rule. Ce complexe médiateur constitue un interface entre les morphogènes maternels et la machinerie basale de transcription, permettant ainsi l'expression zygotique des gènes primordiaux pour la détermination de l'identité des segments. Ce complexe est d'autant plus intéressant qu'il est conservé chez les métazoaires et pourrait jouer un rôle similaire dans l'AGZ de ces organismes. Cependant, l'action de ce médiateur n'explique qu'en partie la régulation de l'AGZ puisqu'il semble cibler spécifiquement les gènes de segmentation. Or plusieurs centaines de gènes sont activés durant les vagues de l'AGZ (67, 68, 69, 70, 71).

Le cas spécifique des gènes de la segmentation met en évidence le rôle crucial des gènes à effet maternel et leur contrôle post-transcriptionnel. Je n'aborderai cependant que brièvement les mécanismes de dégradation des ARNm maternel et me focaliserai sur l'activation du génome zygotique. Je détaillerai les mécanismes impliqués dans le contrôle temporel de l'AGZ et présenterai les quelques acteurs moléculaires connus.

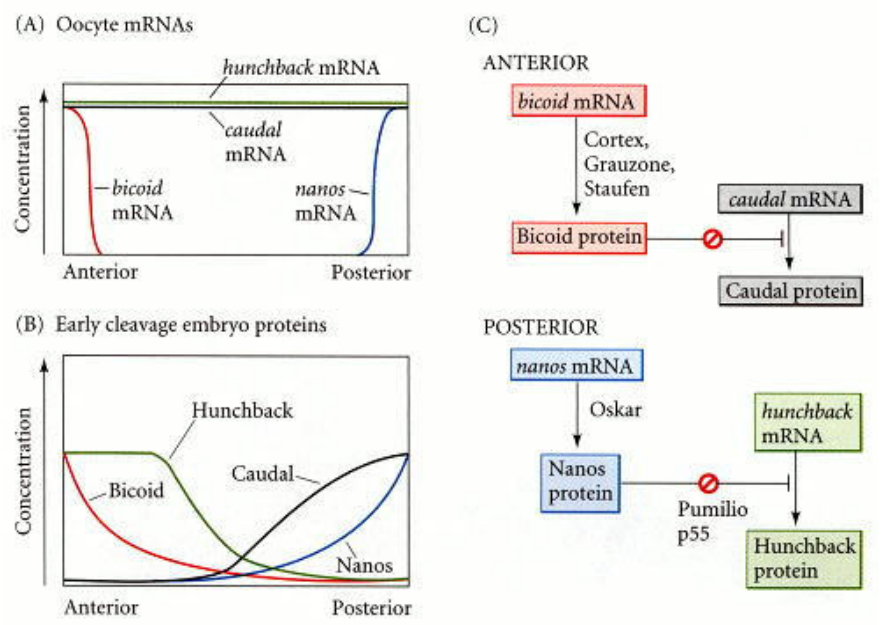


FIGURE 1.11: Modèle de la spécification antéro-postérieure (AP) par les gènes à effet maternel. - (A) Les ARNm bicoid, nanos, hunchback et caudal, déposés dans l'oocyte par les cellules nourricières durant l'oogenèse. Les ARNm bicoid et nanos sont respectivement transportés aux pôles antérieur et postérieur de l'oocyte. (B) La traduction des ARNm bicoid produit un gradient antéro-postérieur et celle de nanos un gradient postéroantérieur. Bicoid empêche la traduction des ARNm caudal au pôle antérieur et Nanos inhibe la traduction des ARNm huchback au pôle postérieur. (C) Interaction parallèle de la régulation traductionnelle induisant la spécification de l'axe AP (d'après Gilbert, Developmental Biology. 6ème édition 2000) (66).

– La dégradation des ARNm maternel

Avant l'activation de l'oeuf, les transcrits d'environ 50% des gènes codant des protéines (environ 6000 gènes) sont présents l'oocyte (68, 70, 72). Thomsen et collaborateurs (73) ont récemment estimé que presque 4000 espèces d'ARNm sont dégradées durant la TMZ, dont un millier dès l'activation de l'oeuf (72). Deux voies indépendantes mais non exclusives sont impliquées dans la dégradation des ARNm. Une voie catalysée par les facteurs maternels et une voie zygotique catalysée par les nouveaux ARNm produits lors de l'AGZ. La dégradation de certains ARNm peut dépendre de l'action des deux voies (73). Le facteur maternel le plus connu, Smaug, a été identifié pour son implication dans le blocage de la traduction des ARNm nanos lors de l'activation de l'oeuf. Tadros et al. (2007) lui ont attribué la responsabilité de la dégradation d'environ 2/3 des ARNm dégradés dès l'activation de l'oeuf. La dégradation induite par Smaug semble être essentielle dans le ralentissement des cycles mitotiques syncytiaux et participerait à l'activation du point de contrôle de la réplication à partir du cycle 10, ainsi que de la dégradation de la cycline B (60). Bushati et al. (74) ont mis en évidence l'élimination de 410 ARNm par un système purement zygotique impliquant le cluster de microARNs (miARNs) miR-309. Parmi ceux-ci, 138 ARNm sont dégradés dès l'activation de l'oeuf dont 92 sont des cibles de Smaug. De plus, dans les mutants *smg*, environ 85% des transcrits dégradés par cette classe de miARNs sont stabilisés (60). Ceci suggère que Smaug collabore avec les facteurs zygotiques pour l'activation de l'expression zygotique de ces miRNAs, probablement indirectement via la dégradation de suppresseurs. Smaug et le cluster miR-309 n'expliquent pas la totalité de la dégradation des ARNm maternels. De Renzis et al. (2007) et Thomsen et al. (2010) ont mis en évidence l'implication de la protéine Pumilio, connue pour être impliqué dans le blocage de la traduction des ARNm maternels dans l'oeuf activé non-fécondé en s'associant avec Nanos (figure 1.11), mais aussi dans leur dégradation (75). D'autre part, les travaux menés par ces deux groupes et ont également mis en évidence l'enrichissement d'une famille de motifs riches en AU (AU Rich Elements, AREs) dans les régions 3'UTR des gènes maternels. Ce type de motif est reconnu par des miARN et a été impliqué dans la dégradation de certains ARNm dans des cellules S2. L'action de ces miARNs serait exclusivement zygotique (72, 73).

La dégradation des ARNm maternels est un processus dynamique et spécifique (73) qui commence dans l'oocyte dès son activation. Trois raisons ont été proposées pour expliquer l'importance de la dégradation des ARNm maternels. Dans un premier temps, la dégradation

pourrait compenser un dosage anormal des ARNm dans l'embryon (cependant cela devrait impliquer une dégradation globale plutôt que spécifique). Un deuxième modèle suggère que la dégradation spécifique de certains ARNm maternels distribués de façon ubiquitaire permettrait une expression localisée pour la mise en place de régions spécialisées (68, 73). Le dernier modèle propose une dégradation instructive plutôt que permissive. Par exemple, le niveau des ARNm impliqués dans le cycle cellulaire est responsable du ralentissement des cycles mitotiques jusqu'à la pause en interphase du cycle 14, qui permet la cellularisation. L'exemple du gène string, codant pour un régulateur des cycles mitotiques, permet d'illustrer ces 2 derniers modèles. En effet, l'élimination des ARNm maternels string à la fin de la TMZ laisse place à une expression zygotique régionalisée correspondant aux domaines mitotiques de l'embryon en gastrulation, ce qui est consistant avec le modèle permissif. Edgar et Datar (1996) (76) ont montré qu'une augmentation (ou une diminution) des ARNm string et twine entraîne une augmentation (ou une diminution) du nombre de cycles nucléaires avant la cellularisation. La dégradation est alors instructive.

– Régulation temporelle de l'AGZ

Trois modèles ont été proposés pour la régulation de la répression et l'activation de l'AGZ au niveau temporel : le ratio nucléo-cytoplasmique (diminution du rapport ADN/cytoplasme au fil des cycles mitotiques dans le syncytium), l'horloge maternelle (temps absolu après la fécondation) et l'avortement de la transcription dû à la rapidité des premiers cycles mitotiques (cycle 1 à 8, cf. figure 1.9A).

Jusqu'à la TMB, le volume global de l'embryon est stable, alors que la quantité d'ADN augmente exponentiellement, modifiant ainsi le ratio ADN/cytoplasme (ratio NC). Edgar et al. (1989) (77) ont montré que la diminution de moitié du contenu en ADN dans des embryons mutants haploïdes engendrait un cycle mitotique supplémentaire avant la TMB, qui se produit alors au cycle 15 plutôt que 14. À partir de ces observations, les auteurs ont suggéré que la titration de facteurs maternels par la quantité d'ADN en constante augmentation pourrait expliquer un tel phénomène. Pritchard et Schubiger (1996) (78) ont montré que la répression de fushi tarazu (ftz) dépend de la quantité du répresseur maternel Tramtrack (Ttk), elle même dépendante du ratio N/C. Ils ont également suggéré qu'il existait d'autres répresseurs maternels titrés expliquant la répression/activation d'autres gènes (comme krüppel, qui n'est pas régulé par Ttk mais semble être soumis au ratio NC).

Lu et collaborateurs (2009) sont partis de ces observations et ont mis en place une étude à grande échelle, basée sur l'utilisation de puces à ADN, pour détecter les gènes dont l'activation dépend du ratio NC. Pour cela, ils ont comparé l'abondance des ARNm dans des embryons sauvages et des embryons haploïdes durant la cellularisation qui se produit respectivement entre les cycles 13-14 et 14-15. Ils se sont particulièrement intéressés aux ARNm ayant une composante exclusivement zygotique (en partant des gènes purement zygotiques rapportés par (68)) et ont ainsi identifié 88 gènes purement zygotiques dont l'activation dépend du ratio NC.

Une grande majorité des gènes exprimés lors de l'AGZ ne répond pas au ratio NC. Au contraire, leur activation dépend du temps absolu après l'activation de l'oeuf et/ou après la fécondation, indépendamment du caractère diploïde ou haploïde de l'embryon. Ceci correspond au modèle impliquant l'horloge maternelle, dont Smaug semble être un des principaux acteurs (60, 72). Benoit et collaborateurs ont suggéré que Smaug est responsable de la dégradation de répresseurs de l'activation zygotique, qui corrèle avec l'augmentation de la transcription pendant la TMZ , indépendamment des points de contrôle de la réplication, permettant par exemple l'accumulation de la forme phosphorylée de l'ARN polymérase II.

Enfin, la rapidité des cycles cellulaires ne permettrait pas de terminer la transcription des ARNm. L'application d'inhibiteurs du cycle cellulaire résulte en une activation prématurée de la vague majeure d'AGZ. De plus, il a été démontré qu'effectivement la progression du cycle mitotique arrête les transcriptions en cours (58). De Renzis et al. 2007 ont montré que les premiers transcrits zygotiques synthétisés lors de la première vague (cycles ralentis mais encore rapides) sont relativement courts et dépourvus d'introns. L'arrêt en interphase du cycle 14 est donc nécessaire pour la transcription d'ARNm plus long.

Le ralentissement des cycles mitotiques semble indispensable mais pas suffisant pour expliquer l'activation de la transcription. En effet, lorsque les cycles sont bloqués artificiellement, l'AGZ ne peut être initiée avant le cycle 10 (77). Ceci indique que d'autres mécanismes sont impliqués pour empêcher l'activation trop précoce du zygote. Ces trois modèles ne sont pas exclusifs et semblent se compléter. Particulièrement, le cycles mitotiques sont régulés à la fois par le ratio NC et l'horloge maternelle. Le ratio NC est effectivement impliqué dans la titration de la machinerie de réplication (79) et de la Cycline B (59). Ceci permettrait l'introduction de points de contrôle, assurés par les protéines Chk1 et Chk2, entre la fin de la phase S (synthèse d'ADN) et le début de la phase M (mitose) à partir du cycle 8 (79), ce qui induirait le ralentissement des divisions cellulaires. D'autre part, Smaug est en partie responsable de la dégradation de la Cycline B qui réprime Chk1 et Chk2. En fait la dégradation de certains activateurs du

cycle cellulaire, comme nous venons de le voir, ou de la transcription, comme Ttk, dépend des deux mécanismes.

– Les facteurs activateurs de l'AGZ

Comme nous l'avons vu précédemment, l'activation d'un nombre restreint de gènes impliqués dans la segmentation s'explique en partie par le recrutement du complexe médiateur contenant dMED31 (53). Les régions régulatrices impliquées dans l'activation des gènes de la segmentation ont été intensivement étudiées à grande échelle (31, 35) ou de façon spécifique (régions régulatrices d'even-skipped (80)). Il en va de même pour les gènes impliqués dans la spécification de l'axe dorso-ventral (cibles de Dorsal (81)). Cependant, ces régions activent ne peuvent pas expliquer l'activation ubiquiste de nombreux gènes durant l'AGZ.

Plusieurs études un mis en évidence l'action d'un facteur maternel, Zelda, dans l'activation des gènes durant la première vague de l'AGZ (68, 82, 83). Ce facteur collabore avec Dorsal (81) et STAT (84), et pourrait jourer un rôle d'amplificateur général de l'action des FT précoces (Hunchback, Bicoid, Krüpple, Giant, Caudal, Knirps) (29, 50).

1.2.3 Apport des études transcriptomiques

J'ai cité précédemment plusieurs études qui ont permis d'identifier les gènes induits durant l'AGZ (67, 68, 71). Ces études ont été menées à large échelle au moyen de puces à ADN, qui permettent d'étudier les profils de transcription de tous les gènes simultanément dans différentes conditions. Ce genre d'étude permet d'évaluer des profils temporels d'expression à différents stades de développement, ou dans différents contextes génétiques (pour étudier par exemple l'effet de mutations de protéines régulatrices conduisant à une expression différentielle de gènes cibles directs ou indirects). Ce type d'expérience est particulièrement utilisé pour sélectionner des groupes de gènes coexprimés afin de découvrir des éléments fonctionnels partagés par ces gènes. Il existe plusieurs méthodes pour définir de tels clusters qui seront décrites et discutées dans la section 2.1 du manuscrit.

J'ai déjà décrit brièvement l'étude mené par Lu et collaborateurs (2009) dans le sous chapitre précédent. Brièvement, ils ont analysé les profils d'expression temporels dans des embryons sauvages et haploïdes afin de détecter les gènes dont l'activation est soumise au ratio NC.

L'étude menée par De Renzis et al. (2007) a consisté à analyser l'expression des gènes durant différentes classes temporelles entourant l'AGZ dans des embryons sauvages et des

embryons dont un chromosome (ou d'une partie de chromosome) a été retiré. Les résultats de ces expériences ont permis de détecter la contribution maternelle et zygotique de chaque espèce de transcrits et d'effectuer classement temporel précis des gènes par rapport au deux vagues de l'AGZ. Ils ont également mis en évidence la surreprésentation spécifique du motif de liaison de Zelda dans les régions promotrices des gènes activés dans le blastoderme pré-cellulaire (68).

Enfin, Pilot et collaborateurs (2006) ont produit une série temporelle transcriptomique afin de détecter les gènes impliqués dans la cellularisation.

L'ensemble de ses études sera décrit plus particulièrement dans la partie 2.

1.2.4 Objectif de la thèse

L'activation transcriptionnelle du génome zygotique (AGZ) est un événement fondamental dans le développement des embryons, car elle permet le remplacement des produits maternels dégradés et la synthèse de nouveaux transcrits essentiels pour la poursuite du développement. Cette activation concerne plusieurs centaines gènes et la régulation d'une petite partie d'entre eux, les gènes de la segmentation, est au mieux partiellement comprise. L'objectif général de cette thèse est d'approfondir les connaissances sur la régulation de l'AGZ en identifiant de nouveaux éléments et modules cis-régulateurs communs aux gènes induits durant cette période par une approche bioinformatique. A cet égard, il a d'abord fallu déterminer la meilleure méthode pour détecter les gènes induits durant l'AGZ à partir de données temporelles de transcriptome (Chapitre 2). En effet, la précision du groupe de gènes choisi est déterminante pour détecter des éléments cis-régulateurs pertinents, alors que la détection bioinformatique de tels éléments repose sur le partage significatif de caractéristiques par ces gènes (chapitre 3). Cette approche m'a permis de détecter et sélectionner les éléments et modules cis-régulateurs les plus pertinents. D'autres données génomiques fonctionnelles (par exemple concernant l'état de la chromatine) on ensuite été exploitées pour identifier des signatures particulières associées aux CRM prédits (Chapitre 4). Enfin, les résultats obtenus ont été intégrés dans un modèle mécanistique de l'AGZ (Chapitre 5).

Pour ne pas trop alourdir l'introduction, j'ai choisi de présenter les principales méthodes bioinformatiques utilisées, ainsi que des données biologiques complémentaires au fil des besoins, au sein des chapitres qui suivent. Par ailleurs, les principales publications auxquelles j'ai participé sont présentées en annexe.

2

Identification de gènes co-exprimés pendant l'activation transcriptionnelle du génome zygotique (AGZ)

Le projet trouve son origine dans les travaux de Fanny Pilot (67), qui a produit des données transcriptomiques à partir d'embryons précoces de drosophile minutieusement échantillonnés à la loupe binoculaire et répartis en cinq classes ou stades du développement (figure 2.1B). Elle a ainsi extrait les ARNm pour cinq classes temporelles. La première classe permet d'analyser le contenu en ARNm de l'embryon avant l'AGZ, ces ARNm correspondent aux ARNm maternels représentés en rouge dans la figure 2.1A. Les classes temporelles suivantes couvrent la phase d'activation majeure du génome zygote, dans le but de caractériser les transcrits nouvellement synthétisés à partir du génome zygotique (représentés en bleu foncé dans la figure 2.1A). Le but de cette étude était d'identifier des gènes variant de façon similaire entre classes temporelles consécutives. J'ai tout d'abord utilisé des approches "classiques" de clustering sur les profils temporels d'expression des gènes. Ces analyses ne m'ont pas permis d'extraire des groupes de gènes bien définis, me laissant juger arbitrairement de la définition de chaque cluster, et ne permettant pas d'évaluer directement les variations d'expression entre classes temporelles consécutives. J'ai alors décidé de développer une approche adaptée en discrétisant, sur base de critères statistiques robustes, les profils de transition. Cette méthode m'a permis d'obtenir des clusters compacts et cohérents biologiquement.

Durant mon doctorat, Lu et collaborateurs (71) ont analysé l'impact du ratio nucléo-cytoplasmique (NC) sur l'AGZ, en étudiant les profils temporels d'expression des gènes dans

des embryons diploïdes sauvages et des embryons haploïdes mutants (figure 2.1D). Leur étude ne se basant que sur un sous groupe de gènes, j'ai décidé de traiter ces données en adaptant la méthode de discrétisation développée précédemment. J'ai ainsi pu identifier les gènes dont l'activation suivait le modèle impliquant le ratio NC ou l'horloge maternelle. Un troisième groupe de gènes dont l'activation ne suit aucun des deux modèles connus à également été identifié.

Enfin, De Renzis et collaborateurs (68) ont étudié la contribution des ARNm maternels et zygotiques dans l'abondance des transcrits observée dans l'embryon au moment des deux vagues de l'AGZ (figure 2.1C). Pour cela ils ont utilisé des données temporelles couplées à des expériences de délétions chromosomiques. La méthode proposée précédemment n'étant pas adaptée à ce schéma expérimental, j'ai utilisé les clusters qu'ils avaient définis.

2.1 Mise en place d'un protocole d'analyse de séries temporelles à partir des données de Pilot et al.

2.1.1 Normalisation

La première étape dans le traitement de données issues des expériences de puces est la normalisation des données. Les puces Affymetrix fournissent pour chaque sonde un couple de valeurs PM ("*Perfect Match*") et MM ("*MisMatch*"). Les valeurs MM permettent d'estimer le bruit de fond dû à des liaisons non spécifiques. Dans les puces de type Affymetrix, chaque gène est représenté par huit sondes. Plusieurs algorithmes de traitement existent, dont les différences reposent principalement sur la façon d'identifier les hybridations non spécifiques. La méthode de normalisation MAS5.0 (85), incluse dans les logiciels fournis par la firme Affymetrix, repose sur la soustraction des valeurs MM aux valeurs PM, en adoptant un traitement spécifique pour les cas où les MM ont des valeurs plus élevées que les PM (gènes considérés comme "absents", plus précisément non-détectés). L'utilisation des valeurs PM-MM pour l'analyse de l'expression différentielle des gènes produit beaucoup de faux positifs, en particulier pour des échantillons contenant une faible concentration de molécules et pour les valeurs faibles (86). Des labels de détection (absent : 'A', présent : 'P' et marginal : 'M') ont été calculés (87) à partir des valeurs PM/MM permettant d'écarter les sondes non fiables. En effet, le retrait de ces valeurs diminue sensiblement le taux de faux positifs (FP) mais conduit à une augmentation des faux négatifs (FN) (88) [1]. Ces résultats montrent que les étiquettes "A" sont ambigues et ne

1. Les FP sont des gènes rapportés comme étant différentiellement exprimés alors qu'il ne le sont pas. À l'inverse, les FN sont des gènes qui n'ont pas été détectés alors qu'ils présentent une expression différentielle.

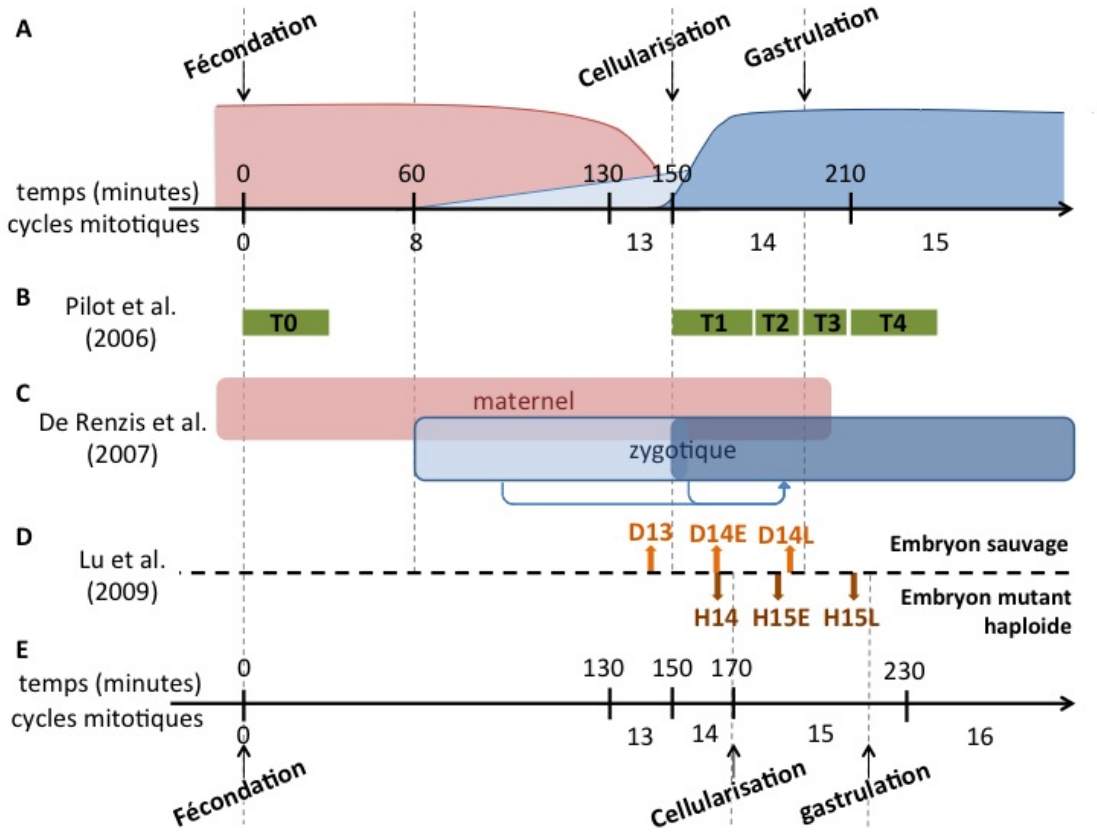


FIGURE 2.1: Répartition temporelle des données transcriptomiques utilisées pour l'étude de l'AGZ. - A. ARNm maternels (profil rouge) et ARNm zygotiques synthétisés durant la première et la seconde vague d'AGZ (profil bleu clair et foncé, respectivement). B. Intervalles temporels étudiés par Pilot et al. (2006). C. Représentation des clusters définis par De Renzis et al. (2007), les couleurs ont la même signification qu'en A. Les flèches indiquent l'activation secondaire des gènes zygotiques par des facteurs zygotiques. D. Schéma expérimental utilisé par Lu et al. (2009). Les préfixes D et H indiquent respectivement les génotypes diploïdes et haploïdes des embryons analysés. Les suffixes indique le cycle mitotique. E. Représentation temporelle des événements de cellularisation et gastrulation dans les embryons haploïdes.

permettent pas de déterminer si la faible hybridation est due à un artefact technique (mauvaise d'hybridation) ou à l'absence réelle du transcrit.

D'autres méthodes ont été proposées, en particulier les méthodes RMA (89) et gcRMA (90), où les valeurs MM ne sont pas utilisées et où les PM sont corrigées par le bruit de fond ajusté et transformées en logarithme base 2, alors que l'ensemble des puces [1] est normalisé par les quantiles afin d'assigner à chaque puces une distribution de valeurs identique à une distribution "étalon" calculée sur l'ensemble des puces. Les méthodes RMA rapporte un faible taux de FP (86, 88) mais l'association des labels de détection à la méthode MAS5.0 réduit le nombre de FP et donne même de meilleurs résultats que RMA sur ce critère. Par ailleurs, le taux de FN est sensiblement le même dans les deux approches. Harr et collaborateurs (91) ont comparé ces différentes approches en considérant les quatre étapes de normalisation de puces à ADN Affymetrix : correction du bruit de fond, normalisation entre puces, correction PM-MM, obtention d'une valeur unique d'expression par jeu de sondes. Ils ont conclu que les méthodes RMA/gcRMA était les meilleures méthodes de normalisation pour la détection de gènes différentiellement exprimés.

Sur la base de ces considérations, j'ai séléctionné la méthode RMA, qui semble la indiquée pour détecter des gènes différentiellement exprimés. De plus, la méthode de labélisation appliquée à MAS5.0 identifie des sondes comme absentes ou "non fiables" pour palier le défaut de MAS5.0, alors que ce label est ambigu. Pilot et collaborateurs (67) avait utilisé les labels de détection MAS5.0 et d'autres critères pour identifier les gènes différentiellement exprimés, puis déterminer les gènes co-exprimées. L'analyse des clusters obtenus par Pilot (annexe 6.1) a montré les limites des labels de détection. Du coup, j'ai choisi de considérer un jeu de données complet et robuste et d'appliquer par la suite des filtres basés sur des tests statistiques disponibles dans les packages R (92).

2.1.2 Sélection de gènes différentiellement exprimés

Avant de procéder au groupement des gènes co-exprimés, j'ai effectué un filtrage des données afin de ne retenir que les gènes présentant une variation d'expression significative entre deux points temporels consécutifs. Tout d'abord, j'ai calculé la valeur médiane $T_{i,j}$ pour chaque gène *i* à chaque point temporel *j*, en prenant ainsi en compte les trois réplicats par classe temporelle. J'ai ensuite calculé pour chaque gène *i* les valeurs de transition $X_{i,j}$, définies comme

1. Les hybridations produites pour analyser le contenu en ARNm dans les différentes conditions (classes temporelles, génotypes ect.) sont faites sur une série de puces de même type.

les log-ratios des mesures d'expression entre les points temporels consécutifs *j-1* et *j* (figure 2.2A).

$$X_{i,j} = log_2(\frac{T_{i,j}}{T_{i,j-1}}) \tag{2.1}$$

J'ai ensuite ajusté une courbe gaussienne à la distribution des log ratio pour chaque transition et j'ai calculé les z-scores $Z_{i,j}$ pour chaque gène *i* comme suit :

$$Z_{i,j} = \frac{X_{i,j} - \tilde{m}_j}{\hat{s}_j} = \frac{X_{i,j} - \tilde{m}_j}{(IQR/1.349)_j} \tag{2.2}$$

où $\tilde{m}_j$ est la médiane des valeurs de transition au temps *j* pour tous les gènes. Nous pouvons noter que j'ai utilisé les estimateurs robustes de la tendance centrale (médiane) et de la dispersion (IQR standardisé [1]) ce qui permet d'être moins sensible aux valeurs extrêmes. J'ai ensuite calculé, pour chaque gène *i* et à chaque transition *j*, la p-valeur nominale $Pval_{i,j}$ basée sur la distribution normale ajustée. La p-valeur a été corrigée pour les multi-tests en multipliant la p-valeur par le nombre de gènes *G*.

$$Eval_{i,j} = Pval_{i,j} * G \tag{2.3}$$

La E-valeur $E_{i,j}$ est ensuite convertie en significativité par transformation logarithmique.

$$sig_{i,j} = -log_{10}(Eval_{i,j}) \tag{2.4}$$

1. L'étendue interquartile (Inter Quartile Range, IQR) est la différence entre le troisième et le premier quartile. Elle couvre les 50% des données au centre de la distribution. L'IQR peut être utilisé comme estimateur robuste de l'écart-type, moyennant une opération de standardisation. Dans une distribution normale standard (moyenne=0, écart-type=1), l'IQR vaut 1.34898. Pour estimer l'écart-type d'une population sur base d'un échantillon, on divise donc l'IQR observé par un facteur de standardisation $\hat{s}$ = *IQR/*$(Q3_{norm} - Q1_{norm})$ = *IQR/1.34898*

J'ai choisi d'appliquer un seuil sur la E-value très stringent (< 0.01) afin de rapporter le moins de FP possibles, même si le taux de FN s'en trouve probablement augmenté. Le but de cette étape est de sélectionner des gènes qui seront ensuite groupés et analysés afin de détecter des caractéristiques communes. En limitant le taux de FP, je limite le bruit de fond qui pourrait noyer des signaux pertinents. L'augmentation du taux de FN, en revanche, m'empêche de sélectioner tous les gènes effectivement exprimés différentiellement. Cependant, j'ai supposé que les gènes présentant les plus fortes variations d'expression étaient représentatifs de l'ensemble des gènes présentant les mêmes variations à plus faible niveau. Le critère de sélection retenu est donc :

$$select_i = \sum_{j=1}^{n} (Eval_{i,j} \leq 0.01 = VRAI) \geq 1 \tag{2.5}$$

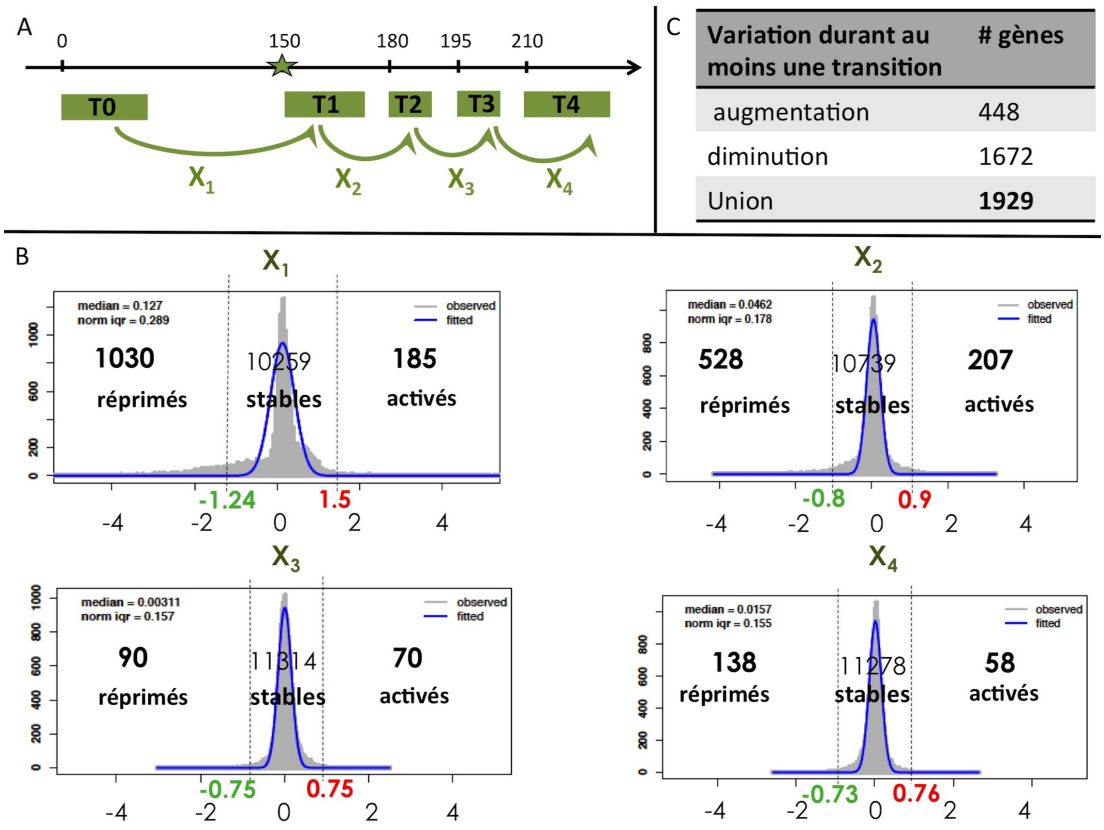
où n est le nombre de transitions présentes dans la série temporelle. Le gène *i* sera sélectionné s'il présente une E-valeur inférieure au seuil de 0.01 pour au moins une transition de la série temporelle. Comme nous pouvons voir dans la figure 2.2B, la distribution des log-ratios de la transition X_1 n'est pas normale, les calculs statistiques basés sur l'ajustement de la distribution normale renvoient des seuil d'autant plus stringents. Ainsi, 1929 gènes ont été sélectionnés (figure 2.2C) parmis 11474 gènes représentés par les sondes hybridées sur les puces utilisées par Pilot et al.

2.1.3 Développement d'une approche pour la sélection de signatures transcriptomiques discrètes

2.1.3.1 Approches classiques de clustering

Afin de regrouper les gènes présentant des profils d'expression similaires, j'ai testé les principales approches pour l'identification de groupes de gènes co-exprimés. Pour cela, j'ai utilisé l'outil MeV (93) (Multiexperiments Viewer) qui propose une série d'algorithmes pour le clustering, la visualisation et l'analyse statistiques de données transcriptomiques.

Tout d'abord, j'ai testé le clustering hiérarchique avec différentes métriques : distance euclidienne, corrélation de Pearson, produit scalaire moyen (*average dot product*) et différentes règles d'agglomération (simple, moyenne, complète). Sur base d'une analyse visuelle des profils de groupes (annexe 6.2), la distance euclidienne et l'agglomération simple donnent de très



C

Variation durant au moins une transition	# gènes
augmentation	448
diminution	1672
Union	**1929**

FIGURE 2.2: Distribution des log2-ratios pour chaque transition entre classes temporelles consécutives. - A. Représentation du schéma temporel des expériences de Pilot (67). L'axe horizontal représente le temps en minutes après la fécondation. L'étoile indique la cellularisation. B. Pour chaque distribution l'abscisse représente la valeur des transition $X_{i,j}$ (log2-ratio entre deux classes temporelles consécutives), l'ordonnée représente le nombre de gènes, les lignes pointillées représentent les seuils négatifs (en vert) et positifs (en rouge) calculés à partir de l'e-value maximale à 0,01. La courbe normale ajustée à la distribution des log2-ratios est indiquée en bleu. C. Résumé du filtrage.

mauvais résultats : les clusters générés présentent des profils dispersés. Je me suis intéressée au produit scalaire moyen qui permet de différencier la corrélation (caractérisée par des signes positifs) de l'anti-corrélation (signes négatifs), ce que ne permet pas la distance euclidienne. Le produit scalaire moyen prend également en compte l'amplitude des variations, ce qui permet de différencier entre les profils "plats" des profils présentant de fortes variations (ce que ne permet pas la corrélation de Pearson). Visuellement, le produit scalaire moyen donnait des regroupement au moins aussi satisfaisants que les autres métriques, en terme de cohérence des groupes. Cependant, le problème qui restait, était dans le choix du niveau de l'arbre à couper pour obtenir des clusters cohérents. Ce choix est assez arbitraire et délicat. La figure 2.3A montre un exemple de choix de cluster visuellement intéressant. Si l'on regarde le profil temporel médian, il y a effectivement une tendance d'augmentation du signal tout au long des points temporels. Cependant les profils de transition individuels apparaissent peu cohérents et informatifs. Par exemple, la transition $X_{.1}$ regroupe des valeurs de log ratio allant de - 1.5 à 3, ce qui correspond à un continuum entre la diminution et l'augmentation significative du signal à $X_{.1}$.

J'ai également testé la méthode des k-means (clustering supervisé), le choix préalable du nombre de clusters influence beaucoup les résultats. La figure 2.3B montre un cluster résultant d'un test avec la méthode k-means. L'observation de la heatmap donne déjà une indication sur l'hétérogénéité de la composition du cluster.

La représentation par des heatmap des profils temporels d'expression est trompeuse et parfois peu intuitive car les passages entre noir-rouge et vert-noir, par exemple, correspondent tous deux à une augmentation de l'intensité entre deux points temporels consécutifs. Les profils de transition regroupés par k-means manifestent donc, comme précédemment, un manque de cohérence par rapport à la question posée, à savoir regrouper les gènes dont l'intensité varie de la même façon tout au long du temps.

2.1.3.2 Clustering sur base de la discrétisation des profils de transition

Sur la base de ces constats, j'ai décidé de travailler directement avec les profils de transition qui permettent une meilleure représentation des variations de signaux entre points temporels consécutifs. J'ai décidé de me servir du calcul de significativité opéré sur les log-ratios durant l'étape de filtrage pour la sélection des gènes différentiellement exprimés. Pour chaque gène et chaque transition, j'ai discrétisé les z-scores $Z_{i,j}$ (équation 2.2) en considérant le seuil $\Theta_{0.01}$

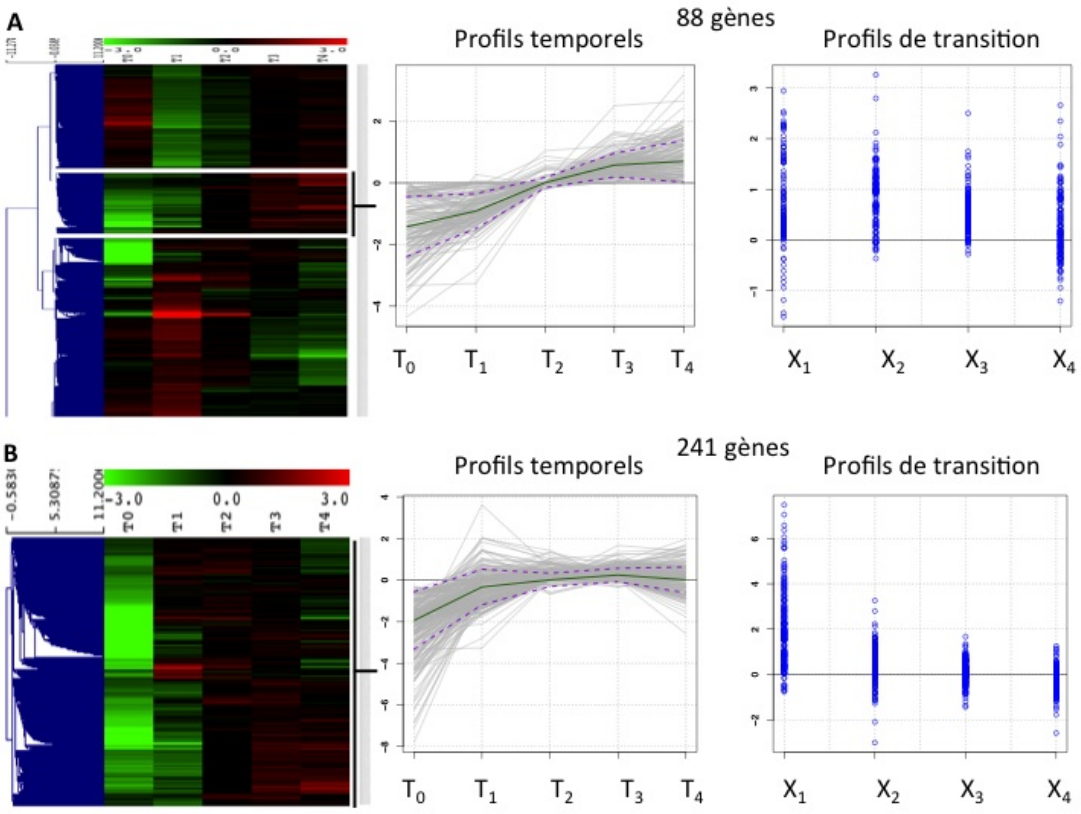


FIGURE 2.3: Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression ("clusters"). - À gauche : heatmap des profils temporels entre T0 et T4, vert, noir et rouge : intensité inférieure, égale ou supérieure à l'intensité médiane du gène au travers de tous les points temporels, respectivement. Au milieu : Profils temporels, où l'ordonnée indique la valeur d'intensité (log2) pour chaque gène (chaque courbe grise correspond à un gène) standardisée par l'intensité médiane du gène dans l'ensemble des classes temporelles. La ligne verte représente le profil médian et les lignes pointillées indiquent l'écart-type par rapport à la médiane. À droite : Profils de transition (X_{i1} à X_{i4}) entre points temporels consécutifs. L'ordonnée indique la valeur du log-ratio et chaque point représente un gène. A. regroupement par clustering hiérarchique, distance : produit scalaire moyen, règle d'agglomération : complète, la heatmap ne représente qu'une partie des données. B. K-means, un des dix clusters obtenus après 50 itérations. Données de Pilot et al. (67)

défini par la e-valeur à 0,01 (figure 2.4A). L'échelle discrétisée admet trois valeurs possibles : u ("up-regulated"), d ("down-regulated") et s ("stable"), définies comme suit :

$$\begin{aligned} D_{i,j} &= u \; si \; Z_{i,j} \geq \Theta_{0.01} \\ D_{i,j} &= d \; si \; Z_{i,j} \geq \Theta_{-0.01} \\ D_{i,j} &= s \; sinon \end{aligned} \tag{2.6}$$

Comme montré dans la figure 2.4B, à chaque gène est assigné un profil discret défini par un vecteur $\bar{D}$ dont les composantes prennent leurs valeurs dans l'ensemble {u, d, s}.

Pour des profils comportant un nombre restreint de points temporels, la discrétisation des valeurs de transition définit un nombre raisonnable de profils possibles. Par exemple, les données de Pilot comportent 5 classes temporelles (T_0,...,T_4) convertis en 4 transitions (X_{i1}, ... X_{i4}). Chaque gène *i* est donc représenté par un vecteur de 4 composantes discrètes (D_{i1}, D_{i2}, D_{i3}, D_{i4}) pouvant prendre trois valeurs (d, s, u). On peut donc a priori définir 3^4 profils discrets de transition distincts. Parmi ces 81 profils possibles, seulement 46 sont retrouvés dans les données analysées.

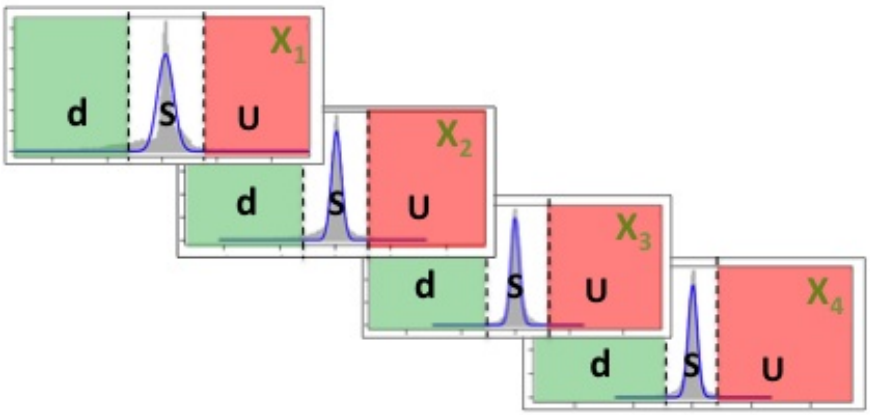
Ce nombre réduit de profils permet donc d'effectuer le clustering par simple regroupement des gènes ayant des profils discrets identiques. La figure 2.5A montre les profils de transition obtenus, qui s'avèrent particulièrement faciles à interpréter (figure 2.5 B). Par ailleurs, les profils temporels sont remarquablement compacts et cohérents (figure 2.5 C-E). À titre de comparaison, j'ai testé les méthodes classiques de clustering sur les profils temporels (annexe 6.3). Les clusters sont beaucoup moins compacts et sont plus difficiles à définir.

Comme montré dans le tableau 2.1, les profils de transition discrets présentent l'avantage d'être directement interprétables biologiquement. De façon intéressante, seulement 18 clusters contiennent plus de dix gènes.

2.2 Modèles de régulation de l'activation des gènes : Lu et al. (2009)

En 2009, Lu et collaborateurs (71) ont étudié l'impact du ratio NC sur l'activation des gènes au moment de la deuxième vague de l'AGZ (figure 2.1D). Ils ont procédé à un clustering hiérarchique sur un sous-ensemble de données d'expression (Affymetrix) pour les gènes purement zygotiques (68). Ils ont ainsi identifié 88 gènes dont l'activation dépend du ratio NC. Afin

A



B

FlyBase ID	X_1	X_2	X_3	X_4	profil
FBgn0001090	5.77 **u**	0.43 **s**	0.32 **s**	-0.17 **s**	**usss**
FBgn0000166	-2.17 **d**	-2.11 **d**	-0.20 **s**	0.06 **s**	**ddss**
FBgn0035572	3.82 **u**	-3.18 **d**	-0.44 **s**	-0.15 **s**	**udss**

FIGURE 2.4: Méthode de discrétisation des profils de transition obtenus à partir des données de Pilot (2006). - A. Distribution des log-ratios pour chaque transition ($X_{i1} - X_{i4}$) et indication de l'assignation des labels selon la significativité de la variation. Les lignes pointillées correspondent au seuil $\Theta_{0.01}$ comme indiqué dans la figure 2.2B. B. Exemple de discrétisation de profils de transition pour trois gènes.

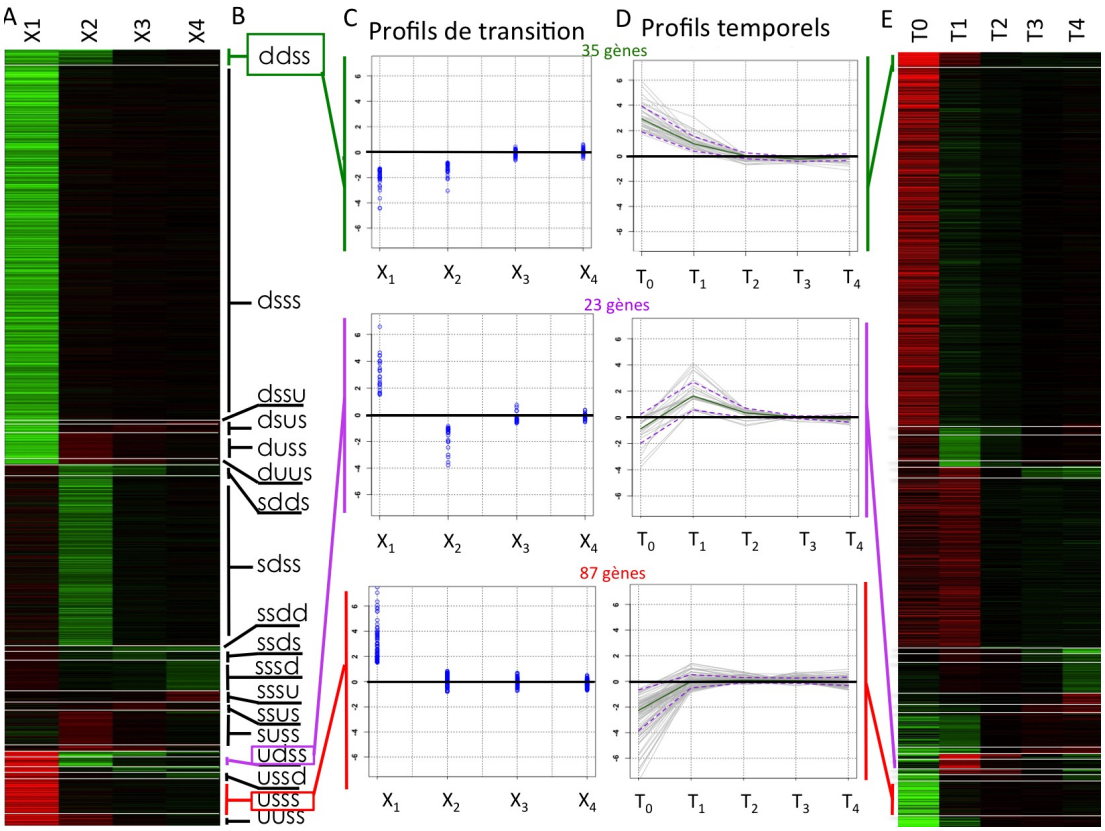


FIGURE 2.5: Clusters résultant du regroupement de profils discrets identiques obtenus à partir des données de Pilot (2006). - A. Heatmap des profils de transition des gènes regroupés. Les couleurs verte, rouge et noire représentent respectivement les variations positive, négative ou négligeable durant la transition. B. Profils de transition discrets. Profils de transition (C) et profils temporels (D) pour 3 clusters particuliers : la classe "ddss" regroupe des gènes maternels, dégradés durant toute la cellularisation ; la classe "usss" des gènes activés précocement puis exprimés de façon stable ; enfin, la classe "udss" des gènes activés de façon transitoire. E. Heatmap des profils temporels des gènes regroupés.

TABLE 2.1: Interprétation biologique des classes de profils discrets comportant au moins 10 membres, obtenues à partir des données de Pilot (2006).

profils	nombre de gènes	Interprétation biologique
ddss	35	ARNm maternels dégradés durant la cellularisation
dsss	885	ARNm maternels dégradés durant la phase lente de cellularisation
dssu	11	ARNm maternels dégradés durant la phase lente de cellularisation puis ARNm zygotiques synthétisés pendant la phase tardive de la gastrulation
dsus	13	ARNm maternels dégradés durant la phase lente de cellularisation puis ARNm zygotiques synthétisés pendant la phase précoce de la gastrulation
duss	66	ARNm maternels dégradés durant la phase lente de cellularisation puis ARNm zygotiques synthétisés pendant la phase rapide de la cellularisation
sdds	23	ARNm maternels dégradés à partir de la phase rapide de cellularisation
sdss	415	
ssdd	12	ARNm maternels dégradés durant la phase précoce de la gastrulation
ssds	22	
sssd	77	ARNm maternels dégradés durant la phase tardive de gastrulation
sssu	28	ARNm zygotiques synthétisés pendant la phase tardive de la gastrulation
ssus	21	ARNm zygotiques synthétisés pendant la phase précoce de la gastrulation
suss	75	ARNm zygotiques synthétisés à partir de la phase rapide de la cellularisation
suus	11	
udss	23	ARNm zygotiques présents de façon transitoire durant la cellularisation
ussd	16	ARNm zygotiques synthétisés dès le début de la cellularisation
usss	87	
uuss	23	

de retirer le maximum d'information de ce jeu de données, j'ai décidé de retraiter la totalité des données en appliquant le protocole défini précédemment : normalisation RMA, sélection des gènes différentiellement exprimés sur base d'un seuil de significativité par échantillon ("chip-wise") et groupement des gènes co-exprimés sur base des profils de transition discrets. La particularité des données (deux génotypes différents) et des interprétations possibles (implication ou non du ratio NC) m'ont amenée à modifier légèrement la procédure. J'ai donc appliqué un autre filtre en considérant la cohérence entre les profils de transition de deux génotypes sur base des deux modèles alternatifs de régulation du niveau d'ARNm, à savoir le ratio NC ou l'horloge maternelle. En effet, dans un embryon haploïde, la cellularisation et la pause en interphase sont retardées d'un cycle mitotique, c'est à dire qu'elles se passent au cycle mitotique 15 au lieu de 14 dans l'embryon diploïde.

Je vais tout d'abord définir la nomenclature utilisée pour les transitions. Ici, les classes temporelles sont définies par le génotype (D : diploïdes, H : haploïde) des embryons et le cycle mitotique pendant lequel les embryons ont été extraits. Le cycle mitotique correspondant à la cellularisation (cycle 14 et 15 chez les diploïdes et haploïdes respectivement) dure environ 1h, les auteurs ont extrait des embryons au début de ce cycle (E : "*early*") et à la fin (L : " *late*"). Ainsi, en suivant l'équation 2.2, la transition entre les classes temporelles consécutives D13 et D14E (embryons diploïdes au cycle 13 et au début du cycle 14) pour le gène *i* sera notée $X_{Di}14E$ et plus généralement pour l'ensemble des gènes X_D14E (figure 2.6A). À partir de ces transitions (diploïdes : X_D14E/X_D14L et haploïdes : X_H15E/X_H15L), j'ai obtenu 37 profils discrets sur les 81 possibles.

Pour les gènes régulés par le ratio NC, on s'attend à ce que les niveaux d'expression (et leurs transitions) du 14ème cycle cellulaire de l'embryon diploïde soit similaire à celui du 15ème cycle de l'embryon haploïde (X_D14E = X_H15E et X_D14L = X_H15L). Au contraire, les gènes insensibles au ratio NC, régulés par l'horloge maternelle, varieront de la même façon quelque soit le génotype des embryons utilisés. Les transitions X_D14L et X_H15E correspondent au même temps absolu (figure 2.6A) et on suppose alors que X_D14L = X_H15E. La figure 2.6B montre un exemple de gène pour chacun des cas. Seuls 24 de ces clusters comportent plus de 10 gènes. Parmi les 24 clusters contenant au moins dix gènes (tableau 2.2), 15 correspondent à l'un de ces deux modèles, dont trois clusters contiennent des transcrits dont la dégradation est dépendante du ratio NC, "ds_Dds_H", "sd_Dsd_H" et "dd_Ddd_H". Les 9 clusters restants contiennent des nombres de gène non-négligeables (montrés dans l'annexe 6.5). Il

serait donc étonnant que cela soit dû à des erreurs de manipulations expérimentales. Il est possible que ces gènes combinent une composante maternelle et une composante zygotique et que leur régulation soit dépendante différemment du ratio NC ou de l'horloge maternelle. Pour la suite de mes analyses, j'ai décidé d'écarter ces gènes en raison de ces ambiguïtés. L'ensemble des gènes retenus est néanmoins plus large et plus diversifié que celui des gènes purement zygotiques analysé par Lu et collaborateurs.

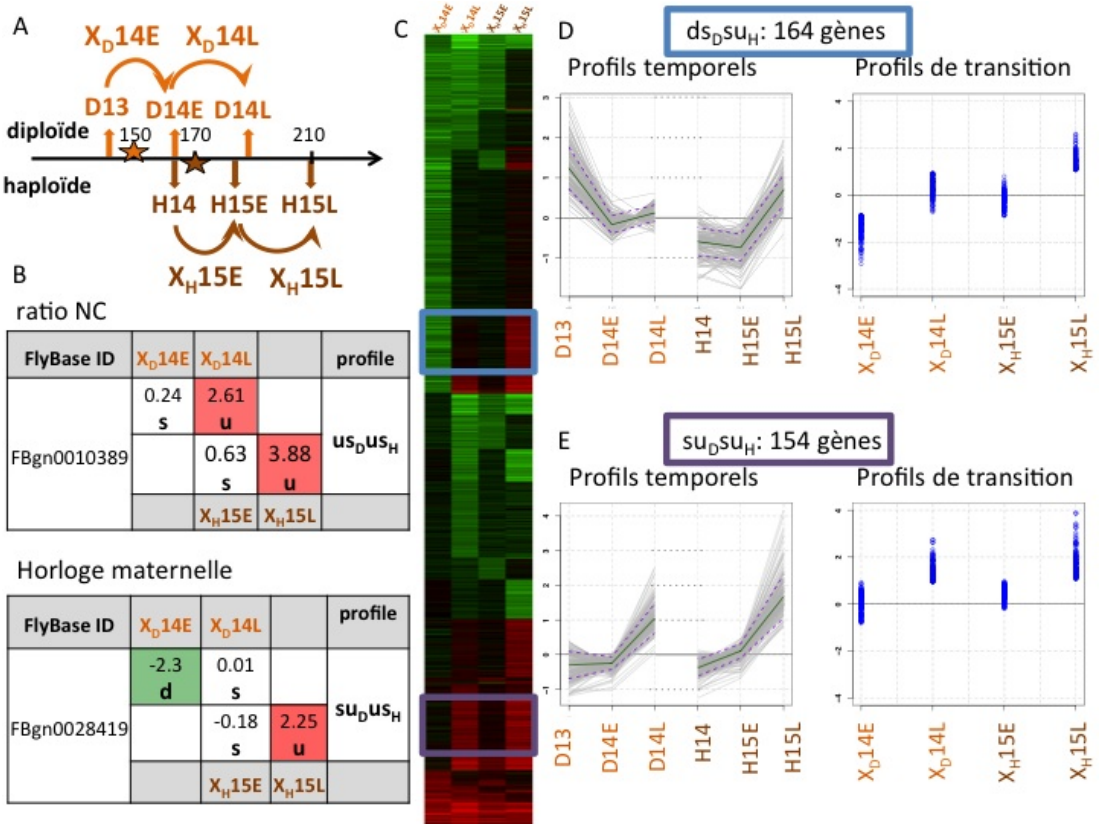


FIGURE 2.6: Application et résultats de la discrétisation des profils de transitions obtenus à partir de des données de Lu et collaborateurs. - A. Schéma expérimental. L'axe horizontal représente le temps en minutes après la fécondation, les étoiles correspondent au moment de la cellularisation observé chez les diploïdes (en haut, orange) et chez les haploïdes (en bas, en marron). B. Les profils de transition entre les deux génotypes sont comparés afin de déterminer si ils sont soumis (tableau du haut) ou non (tableau du bas) au ratio NC. C. Heatmap représentant les clusters de profils de transition discrétisés. D-E. Profils temporels et de transition du cluster ds_Dsu_H et du cluster su_Dsu_H respectivement. Les couleurs correspondent aux profils encadrés dans la heatmap (C).

2.3 Contributions maternelle et zygotique : De Renzis et al. (2007)

D'autres études fournissent des données transcriptomique à grande échelle concernant l'embryogenèse précoce de la drosophile. De Renzis et collaborateurs (68) ont procédé à l'analyse du transcriptome dans des embryons sauvages pour trois classes temporelles (entre 0 et 1h : ARNm maternels, entre 1 et 2h : première vague de l'AGZ ; entre 2 et 3h : deuxième vague de

TABLE 2.2: Interprétation biologique des classes de profils discrets présentant au moins 10 membres, obtenues à partir des données de Lu (2009).

profil	modèle	nombre de gènes	Interprétation biologique
dd_Ddd_H	NC	41	ARNm maternels dégradés durant la cellularisation
dd_Dds_H	horloge maternelle	165	
ds_Dds_H	NC	37	ARNm maternels dégradés durant la phase lente de cellularisation
ds_Dss_H	horloge maternelle	406	
ds_Dsu_H	horloge maternelle	163	ARNm maternels dégradés durant la phase lente de cellularisation puis ARNm zygotiques synthétisés pendant la phase précoce de la gastrulation
ds_Ddu_H	?	19	?
dd_Dss_H	?	104	
du_Dsu_H	?	48	
sd_Dsd_H	NC	91	ARNm maternels dégradés à partir de la phase rapide de cellularisation
sd_Ddd_H	horloge maternelle	61	
sd_Dds_H	horloge maternelle	97	
ss_Dsd_H	horloge maternelle	111	ARNm maternels dégradés durant la phase précoce de la gastrulation
sd_Dss_H	?	213	?
ss_Dds_H	?	58	
ss_Dsu_H	horloge maternelle	164	ARNm zygotiques synthétisés pendant la phase précoce de la gastrulation
su_Dsu_H	NC	154	ARNm zygotiques synthétisés à partir de la phase rapide de la cellularisation
su_Duu_H	horloge maternelle	47	
ss_Dus_H	?	12	?
su_Dss_H	?	37	
us_Dus_H	NC	14	ARNm zygotiques synthétisés dès le début de la cellularisation
us_Dss_H	horloge maternelle	24	
uu_Duu_H	NC	60	
uu_Dus_H	horloge maternelle	27	
uu_Dss_H	?	13	?

l'AGZ et cellularisation) ainsi que pour des embryons privés d'une partie ou de la totalité d'un chromosome pendant la cellularisation (2-3h). Ils ont combiné une analyse temporelle avec la méthode de détection MAS5.0 et à l'analyse des mutants. Cela a permis d'identifier les ARNm d'origine maternelle et ceux d'origine zygotique. Certains ARNm semblent être stables ou décroître dans l'analyse temporelle, alors que l'analyse des mutants indique un apport zygotique. D'autre part, l'analyse des mutants a permis d'identifier les messagers zygotiques provenant de l'activation des gènes par des facteurs zygotiques. Enfin, le choix des classes temporelles a permis de différencier les gènes activés pendant la première et la seconde vague de l'AGZ. Au final, cinq clusters ont été définis par De Renzis et collaborateurs (2007) : les ARNm résultant des transcriptions maternelle et zygotique combinées (nommés "maternel-zygotiques" et regroupant 827 gènes), les ARNm purement zygotique (325 gènes), les ARNm activés de façon très importante pendant la première vague de l'AGZ (60 gènes), les ARNm activés par des produits zygotiques ("cibles secondaires", 217 gènes) et les ARNm dont l'expression décroît mais qui ont pourtant une composante zygotique (653 gènes). Ces clusters sont chevauchants, en effet, les gènes purement zygotiques contiennent des gènes zygotiques précoces et aussi des cibles secondaires, le cluster contenant les ARNm dont l'expression décroît mais qui ont une composante zygotique est un sous ensemble du cluster maternel-zygotique. J'ai choisi de ne pas retraiter ces données car la méthode des profils discrets n'est pas adaptée au schéma expérimental. J'ai donc utilisé directement les clusters définis dans l'étude originale pour les analyses suivantes.

2.4 Analyse de la composition des clusters

J'ai identifié plusieurs groupes de gènes co-exprimés à partir de trois études transcriptomiques. Afin d'analyser la cohérence des clusters contenant des gènes exprimés de façon similaire entre les trois études, j'ai analysé le chevauchement de leur contenu en gènes. La figure 2.7 montre que les clusters regroupant des gènes activés dès le début de la cellularisation identifiés dans le jeu de données de Pilot et al. (profils "uxxx" où x représente n'importe quel type de transition) présentent des chevauchements significatifs avec les clusters du même type (activés au début de la cellularisation) établis sur base des données de Lu et al., ainsi qu'avec les clusters définis par DeRenzis comme regroupant les gènes purement zygotiques, zygotiques précoces et les cibles secondaires (indiqués en jaune dans la figure 2.7). Nous obtenons des résultats similaires pour les gènes activés plus tardivement ("$suxx$", "$ssux$", "$sssu$", "sx_Dsx_H" etc.), pour

les gènes dont les transcrits sont dégradés précocement ("$dxxx$", "dx_Ddx_H " etc), ou encore tardivement ("$sdxx$"," sd_Ddx_H" etc), indiqués respectivement en orange, bleu et vert dans la figure 2.7 . En fait, cette matrice met en évidence la cohérence des clusters obtenus à partir des trois expériences. Le cluster de gènes défini par Lu et al. montre un chevauchement significatif avec les clusters "uu_Duu_H" et "su_Dsu_H" identifiés durant mes analyses. Ceci indique que la méthode basée sur les profils de transition discrets permet une meilleure définition temporelle. Les clusters de gènes activés durant la cellularisation seront appelés par la suite "clusters zygotiques", les clusters regroupant les gènes dont les transcrits sont maternels et sont significativement dégradés pendant la cellularisation seront nommés "clusters maternels" et enfin les clusters regroupant des gènes activés au niveau maternel et zygotique seront nommés "clusters mixtes".

2.5 Conclusion du chapitre

L'analyse approfondie de séries temporelles provenant de données de puces m'a menée à établir une méthode simple adaptée à ce type de données puisqu'elle permet d'apprécier directement les variations de signal entre points temporels consécutifs. De plus, elle permet de regrouper des gènes suivant des profils de transition facilement interprétables. Ainsi 42 clusters contenant plus de 10 gènes ont été définis par cette méthode à partir des données de Pilot et al. (2006) et de Lu et al. (2009), auxquels j'ai ajouté le cluster publié par Lu et al. et cin clusters définis pas De Renzis et al. La comparaison de la composition de ces 48 clusters a montré que des clusters obtenus à partir de données différentes et présentant des profils de transition discrets comparables (par exemple "$dsss$" et "ds_Dss_H" ou "$uuss$" et "uu_Duu_H" définis à partir des données de Pilot et Lu respectivement) présentent un nombre significatif de gènes en commun. Ceci suggère que la méthode de groupement basée sur les profils discrets est relativement fiable. Cependant, cette méthode n'est applicable que pour des séries temporelles relativement courtes. En effet le nombre de combinaisons possibles est exponentiel en fonction du nombre de transition (3^n).

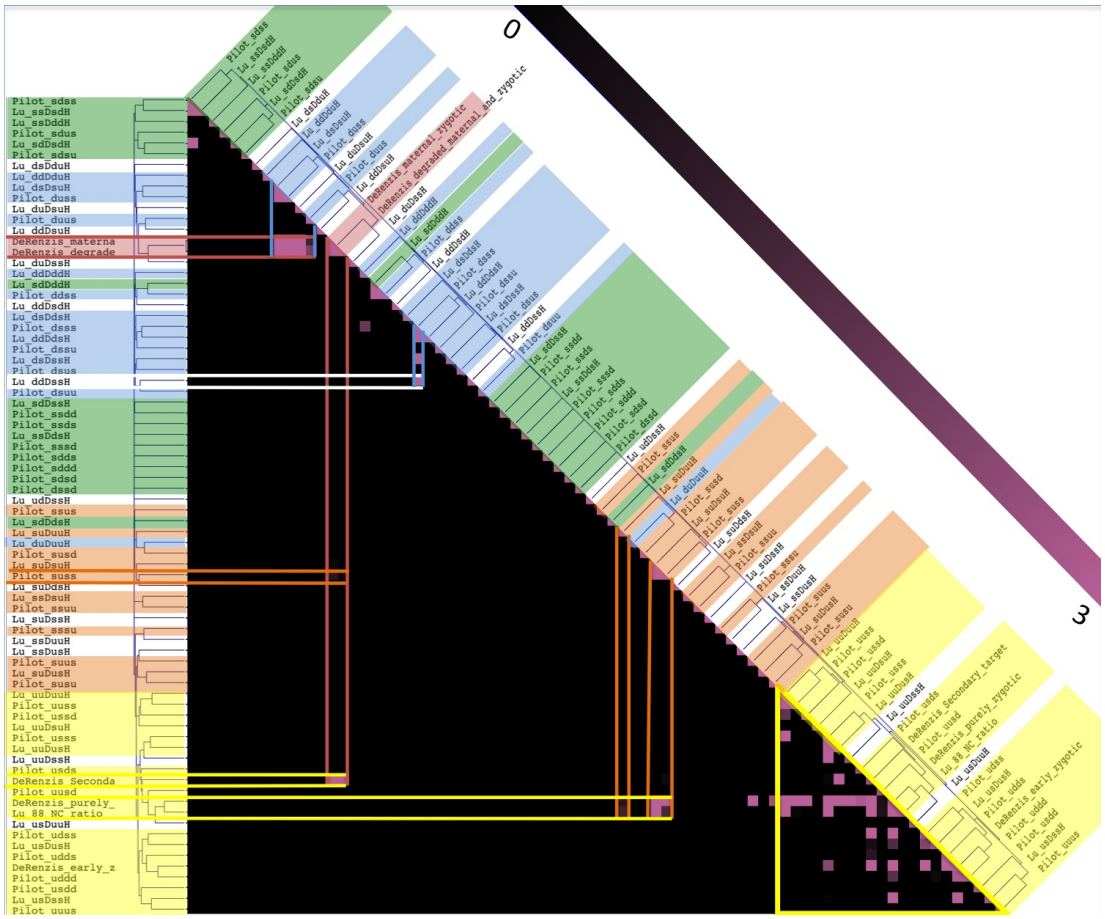


FIGURE 2.7: Comparaison des groupes de gènes co-exprimés en fonction de la significativité (binomiale corrigée pour les multi-tests) de leur chevauchement. - Les clusters de gènes induits avant ou au tout début de la cellularisation sont soulignés en jaune, ceux induit plus tardivement en orange, les gènes dont les transcrits sont dégradés précocement en bleu, enfin ceux dégradés plus tardivement en vert. Les clusters de De Renzis et al. (2007) comportant des gènes dont les transcrits sont maternels et zygotiques sont indiqués en rouge. Les traits à l'intérieur de la matrice mettent en évidence les correspondances entre groupes. Les clusters obtenus avec les données de Lu pour lesquels nous n'avons pas identifié le mécanisme de régulation ne sont pas surlignés. L'échelle des couleurs est indiquée par le gradient noir/violet diagonal correspondant aux significativités allant de 0 à 3 et plus.

3

Analyse fonctionelle des gènes et des régions non-codantes associées

Afin d'identifier des éléments cis-régulateurs impliqués dans la régulation de l'AGZ, j'ai procédé à une analyse fonctionnelle des gènes co-exprimés et des séquences non-codantes associées (5kb en amont du TSS, premier intron, UTR 5' et UTR 3'). Le schéma général d'analyse est présenté dans la figure 3.1. Le coeur du protocole (encadré gris de la figure 3.1) comprend l'analyse de l'enrichissement fonctionnel et la détection de motifs surreprésentés dans les clusters de gènes co-exprimés et a été appliqué à deux reprises. J'ai tout d'abord analysé séparément les clusters définis précédemment ("clusters primaires") (figure 3.1A). En intégrant les résultats de ces analyses, j'ai décidé de regrouper les gènes activés durant l'AGZ possédant des éléments cis-régulateurs potentiels communs ("cluster AGZ"). Dans un second temps, j'ai donc réappliqué le protocole sur le cluster AGZ (figure 3.1B). J'ai ensuite complété cette analyse en testant l'enrichissement des séquences pour une collection de motifs connus. Afin de prédire des éléments et modules cis-régulateurs potentiellement impliqués dans l'AGZ, j'ai scanné les séquences non codantes des gènes du cluster AGZ avec les motifs ainsi identifiés (figure 3.1C).

3.1 Analyse des clusters primaires

3.1.1 Enrichissement fonctionnel

Afin d'analyser la fonction des gènes présents dans chaque cluster et d'identifier des enrichissements particuliers, j'ai utilisé les annotations présentes dans la Gene Ontology (GO) et

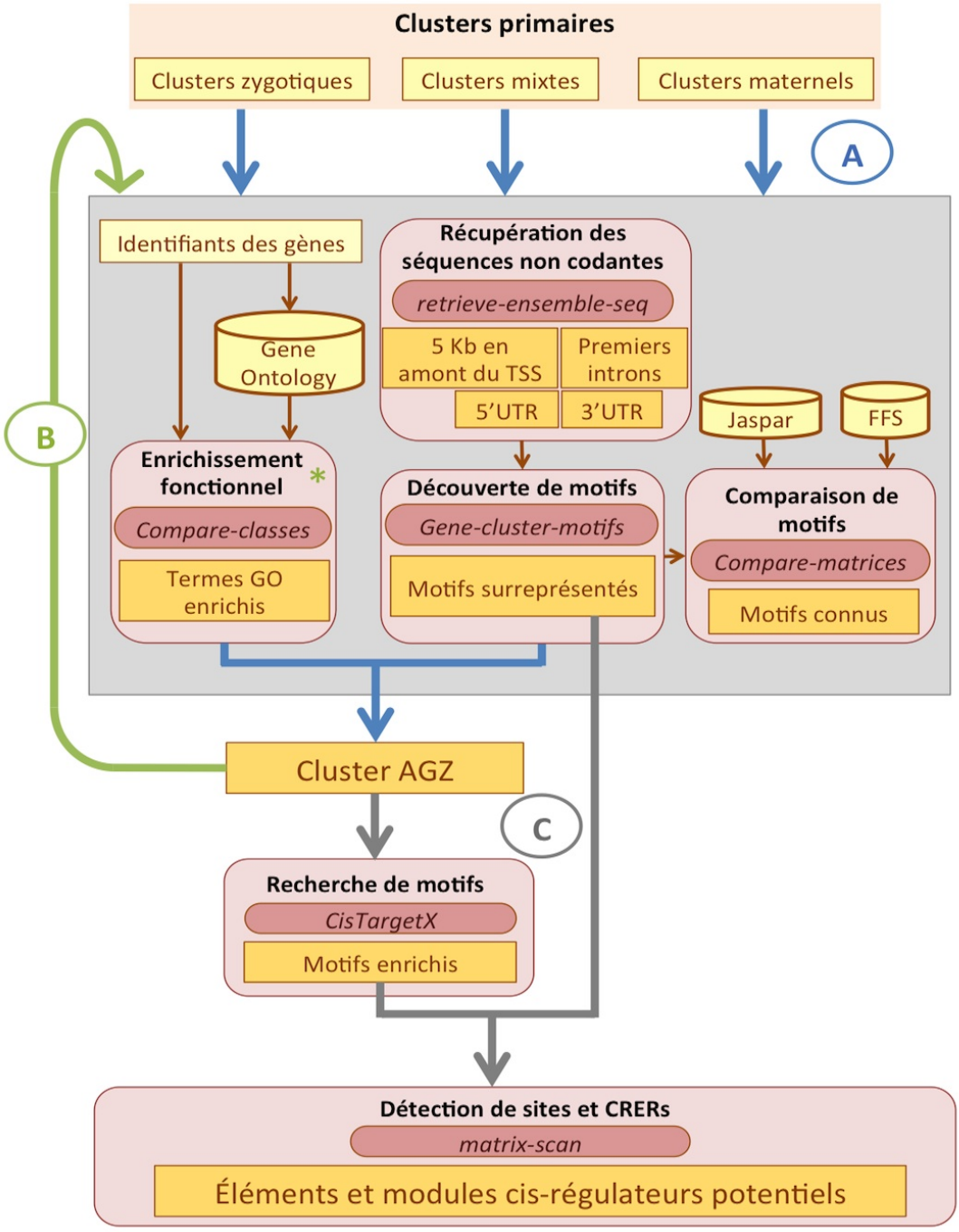


FIGURE 3.1: Organigramme des analyses fonctionnelles pour la prédiction d'éléments et modules cis-régulateurs potentiellement impliqués dans la régulation de l'AGZ. - A. Analyses fonctionnelles des clusters primaires. B. Analyses fonctionnelles du cluster AGZ (qui regroupe l'ensemble des gènes activés pendant l'AGZ). C. Recherche d'éléments cis-régulateurs pour le cluster AGZ. L'encadré gris met en évidence les analyses qui ont été appliquées de façon itérative. L'étoile verte indique des outils qui n'ont été utilisés que lors de l'itération B. FFS : Fly Factor Survey.

l'outil *compare-classes* (94).

GO est une base de données contenant des annotations fonctionnelles pour les gènes et leurs produits, qui utilise un vocabulaire contrôlé commun à toutes les espèces. Les termes utilisés sont organisés en trois classes : les processus biologiques dans lesquels les gènes sont impliqués, les fonctions moléculaires des produits géniques, et les composants cellulaires où ils sont actifs. A chaque terme sont associés un identifiant unique, un nom correspondant à la description, la classe, une définition avec ses sources bibliographiques, ainsi que les termes avec lesquels il est en relation directe. En effet, la base GO est structurée sous forme de graphe orienté acyclique, où chaque terme est en relation avec un ou plusieurs autres termes. Les termes présents aux classes les plus profondes de la hiérarchie correspondent aux descriptions les plus précises. Les termes sont de plus en plus généraux lorsque l'on remonte vers la "racine", qui correspond au nom de la classe. Les gènes sont annotés à partir de différentes sources. Un code ("*evidences*" en anglais) est utilisé pour différencier les annotations reposant sur des indications expérimentales, informatiques (inférences diverses), ou bibliographiques (annexe 6.1). L'explication détaillée de ce code est disponible sur le site web du consortium Gene Ontology (http ://www.geneontology.org/GO.evidence.shtml). J'ai décidé d'écarter les annotations dont la source n'est pas fiable ou peu informative, c'est à dire lorsque :

- l'inférence n'a pas été vérifiée dans le cas d'inférences informatiques automatiques (IEA) ;
- l'inférence a été faite par bibliographie mais aucun support biologique ou informatique n'est disponible (NAS) ;
- le gène a été annoté avec un des termes racine d'une des classes GO (BP : "biological process", MF : "molecular function" et CP : "cellular componant") signifiant qu'on ne sait rien sur ce gène dans la classe donnée (ND) ;
- aucune indication n'est disponible pour l'annotation (NR).

Au total, dans la version la plus récente disponible sur le site Flybase (http ://flybase.org/), 13.090 gènes sont annotés avec 5.513 termes GO. Après filtrage sur base des 4 critères ci-dessus, il reste 9.732 gènes associés à 4.892 termes distincts. Un gène pouvant être associé à plusieurs termes, le nombre total d'associations gène-terme s'élève à 47921. Les détails de la composition des classes en fonction des évidences sont donnés dans l'annexe 6.1 (pour la version fb_2011_08). Je me suis servie de cette version pour l'analyse avec *compare-classes* en opérant une extension des annotations vers la racine de l'ontologie.

Sur les 48 clusters analysés, seuls 18 présentent un enrichissement significatif (e-valeur < 0.1) dans au moins une classe GO ("biological process", "molecular function" ou "cellular component"). Les clusters enrichis en termes GO contiennent un nombre médian de 122 gènes, alors que les 30 autres contiennent un nombre médian de 6 gènes. Le cluster contenant les gènes purement zygotiques est enrichi en un très grand nombre de termes de la classe "biological process". Les termes les plus courants concernent l'engagement dans le destin cellulaire, la différenciation, la morphogenèse, l'organisation cellulaire (en particulier les neuroblastes, neurones, axones). Les termes concernant le développement des tissus (mésoderme, ectoderme, épithélium), des organes (sensoriels, disques imaginaux, coeur, glandes, post-embryonnaires), des systèmes (nerveux, digestif, circulatoire) sont également très significativement enrichis. Enfin, cette analyse a révélé les termes en relation avec la segmentation (détermination zygotique de l'axe AP, partitionnement périodique par les gènes pair-rule), la gastrulation et la régulation de la transcription dépendante de l'ARN polymérase II. En résumé, cette analyse de l'enrichissement reflète la mise en place des éléments essentiels au développement embryonnaire précoce. Les fonctions moléculaires représentées sont toutes en relation avec la liaison à l'ADN et la transcription. Le cluster "Pilot usss" est également enrichi en termes en relation avec le développement et la morphogenèse de l'embryon, la différenciation cellulaire et la neurogenèse. L'ensemble des résultats est fourni en matériel supplémentaire (fichier *GO_enrichment_compare_classes_all_clusters.xlsx* téléchargeable à l'adresse http ://rsat.bigre.ulb.ac.be/ elodie/these/resultats_supplementaires/functional_analysis/).

3.1.2 Éléments cis-régulateurs

J'ai procédé à la découverte de motifs surreprésentés dans chacun des clusters et pour chaque type de séquences non-codantes (5kb en amont, premier intron, 5'UTR et 3'UTR, cf. figure 3.7). Pour cela, j'ai utilisé le pipeline gene-cluster-motifs combinant différents programmes d'analyse de séquences régulatrices RSAT (95). Ce pipeline comprend une étape de découverte de motifs avec plusieurs outils, dont *oligo-* et *dyad-analysis*, ainsi que la construction des matrices poids-positions à partir des mots ou dyades significativement surreprésentés à l'aide des algorithmes *pattern-assembly* et *matrix-from-patterns* (les paramètres utilisés sont indiqués dans la partie 3.3.2). J'ai ensuite comparé tous les motifs découverts à partir des séquences correspondant aux différents clusters. Parmi 48 clusters analysés (ayant plus de 10 membres), 33 possèdent au moins un motif surreprésenté de manière significative ($sig \geq 2$). Certains motifs sont surreprésentés de façon récurrente dans plusieurs clusters.

3.1.2.1 Choix du modèle de background pour la détection de mots sur-représentés

oligo-analysis offre la possibilité de calculer soi-même les modèle de background. J'ai donc testé plusieurs modèles calibrés sur base de séquences de drosophiles :

- régions en amont ;
- introns ;
- 3'UTR ;
- 5'UTR.

J'ai également utilisé des modèles de background calculés à partir de séquences d'un vertébré (souris) et ou d'une bactérie (spirochaete) pour élargir la comparaison.

La figure 3.2 montre la composition en nucléotides et di-nucléotides de chaque jeu de séquences de référence. Il n'y a pas de différence entre les séquences non-codantes en amont ("upstream") du TSS (génome entier), des premiers introns, des 5'UTR, ou encore pour un mélange de ces trois types avec les séquences 3'UTR chez *Drosophila melanogaster* (figure 3.2A). Les séquences en amont du TSS des gènes AGZ montre une légère différence par rapport aux séquences citées précédemment. En revanche, la composition des séquences 3'UTR de Drosophila melanogaster diffère des autres types. La différence est d'autant plus importante dans *Mus musculus* et *Spirochaeta thermophila*.

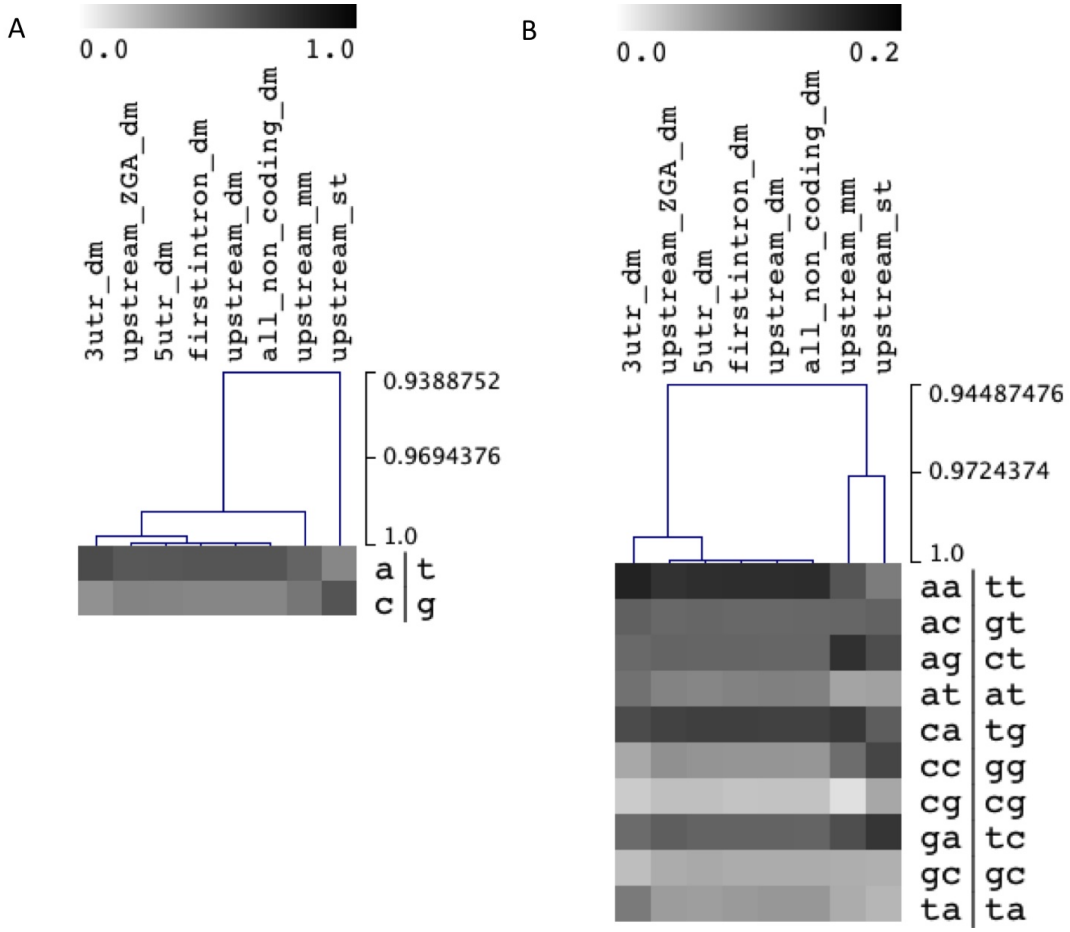


FIGURE 3.2: Composition en nucléotides de différents types de séquences. - dm : *Drosophila melanogaster*, mm : *Mus musculus*, st : *Spirochaeta thermophila*. Le clustering hiérarchique a été calculé par la méthode de Pearson non-centrée.

La figure 3.3 montre les oligomères de 7 nucléotides surreprésentés dans les séquences en amont du TSS en fonction du modèle de background. L'impact du background est bien mis en évidence en comparant l'utilisation de fréquences calculées à partir de séquences provenant d'une espèce différente, à partir des séquences d'entrée (modèle de Markov d'ordre 2), ou encore à partir d'un jeu de référence plus large (ensemble des 3'UTR du génome). Pour les autres tests, même si l'ordre diffère légèrement, les significativités des mots sont du même ordre de grandeur (annexe 6.2). J'ai ainsi choisi d'utiliser le modèle spécifiquement calculé pour chaque type de séquences.

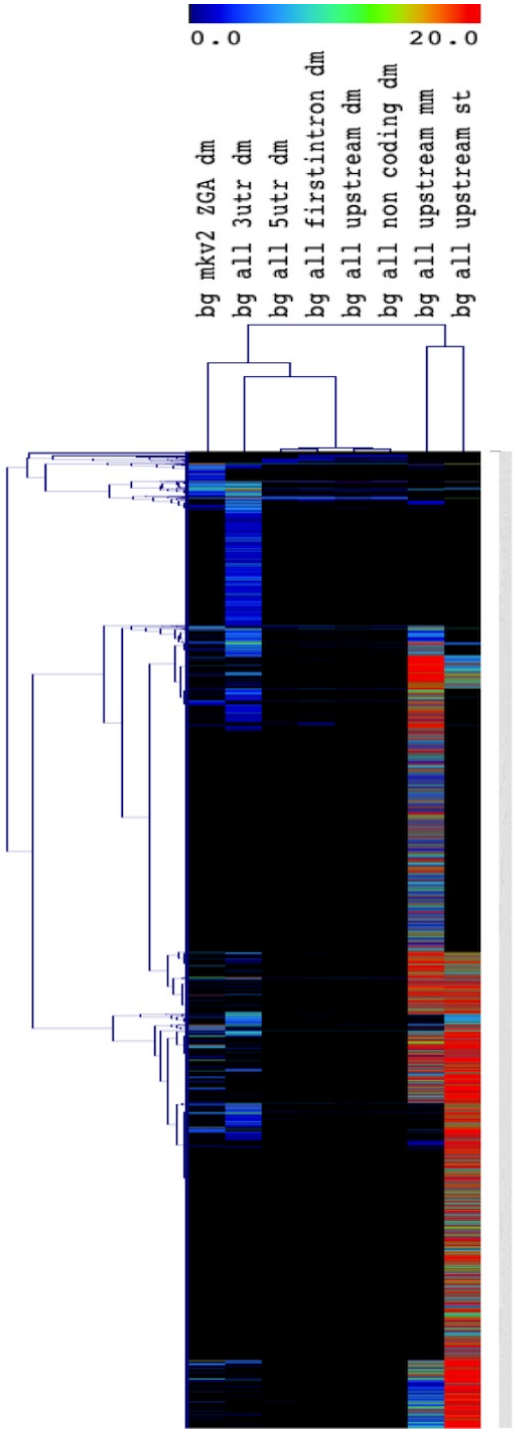


FIGURE 3.3: Heatmap représentant les résultats d'oligo-analysis sur les séquences en amont des TSS des gènes AGZ en utilisant divers modèles de background. - Les couleurs représentent la significativité des oligos (lignes). Chaque colonne correspond à un modèle différent. dm : *Drosophila melanogaster*, mm : *Mus musculus*, st : *Spirochaeta thermophila*.

3.1.2.2 Classification des clusters transcriptionnels en fonction des motifs cis-régulateurs découverts

La classification des clusters transcriptionnels sur la base des motifs surreprésentés (*oligo-analysis* : figure 3.4A ; *dyad-analysis* : annexe 6.6) nous apprend plusieurs choses. D'une part, le regroupement des clusters est semblable à celui observé précédemment, sur base de la composition en gènes (figure 2.7). En effet, les clusters sont regroupés de façon cohérente entre les différentes expériences (c'rst le cas pour les clusters "$usss$" et "us_Dss_H", provenant des données de Pilot et Lu, respectivement). Cette analyse montre en plus que les gènes activés avant ou au tout début de la cellularization présentent les mêmes motifs sur- ou sous-représentés dans leurs séquences non codantes, indépendamment du profil de transition entre les points temporels suivants (par exemple pour les clusters "$uxxx$" et "ux_Dux_H", où x représente n'importe quelle transition u,s,d). La même chose est observée pour les clusters maternels ("$dxxx$", "ds_Dss_H", "ds_Dsu_H"). D'autre part, les deux clusters "mixtes" définis par De Renzis, contenant à la fois des gènes à contribution maternelle et zygotique, sont classés différemment selon le type de séquences non-codantes que l'on analyse. En effet, les régions en amont regroupent ces deux clusters avec les clusters maternels, alors que les séquences 5'UTR et premiers introns les regroupent avec les clusters zygotiques. Ceci suggère que les éléments cis-régulateurs activant ces gènes durant l'oogenèse seraient plutôt présents dans les séquences en amont du TSS, alors que les éléments cis-régulateurs impliqués dans l'activation zygotique de ces gènes se positionneraient plutôt en aval du TSS (5'UTR et premiers introns). En ce qui concerne les régions 3'UTR, seuls trois clusters sont représentés dans l'analyse de surreprésentation des oligos (les clusters zygotiques "zygotique" et "Pilot $uuss$", et le cluster maternel "Pilot $dsss$") et dans l'analyse de surreprésentation des dyades (trois clusters maternels : "Pilot $dsss$", "Pilot $sdss$" et "Lu sd_Dds_H").

3.1.2.3 Classification des mots et dyades

Un intérêt de la classification des mots est qu'elle permet de regrouper les mots chevauchants, suggérant l'implication de motifs plus longs. La figure 3.4B présentent les matrices obtenues à partir de l'assemblage de mots chevauchants, ainsi que le nom des facteurs associés à des motifs significativement similaires (identifiés avec *compare-matrices*). Cette classification révèle également les particularités des clusters maternels, zygotiques et mixtes. En effet, plusieurs groupes de motifs surreprésentés sont associés à un type particulier de cluster. En

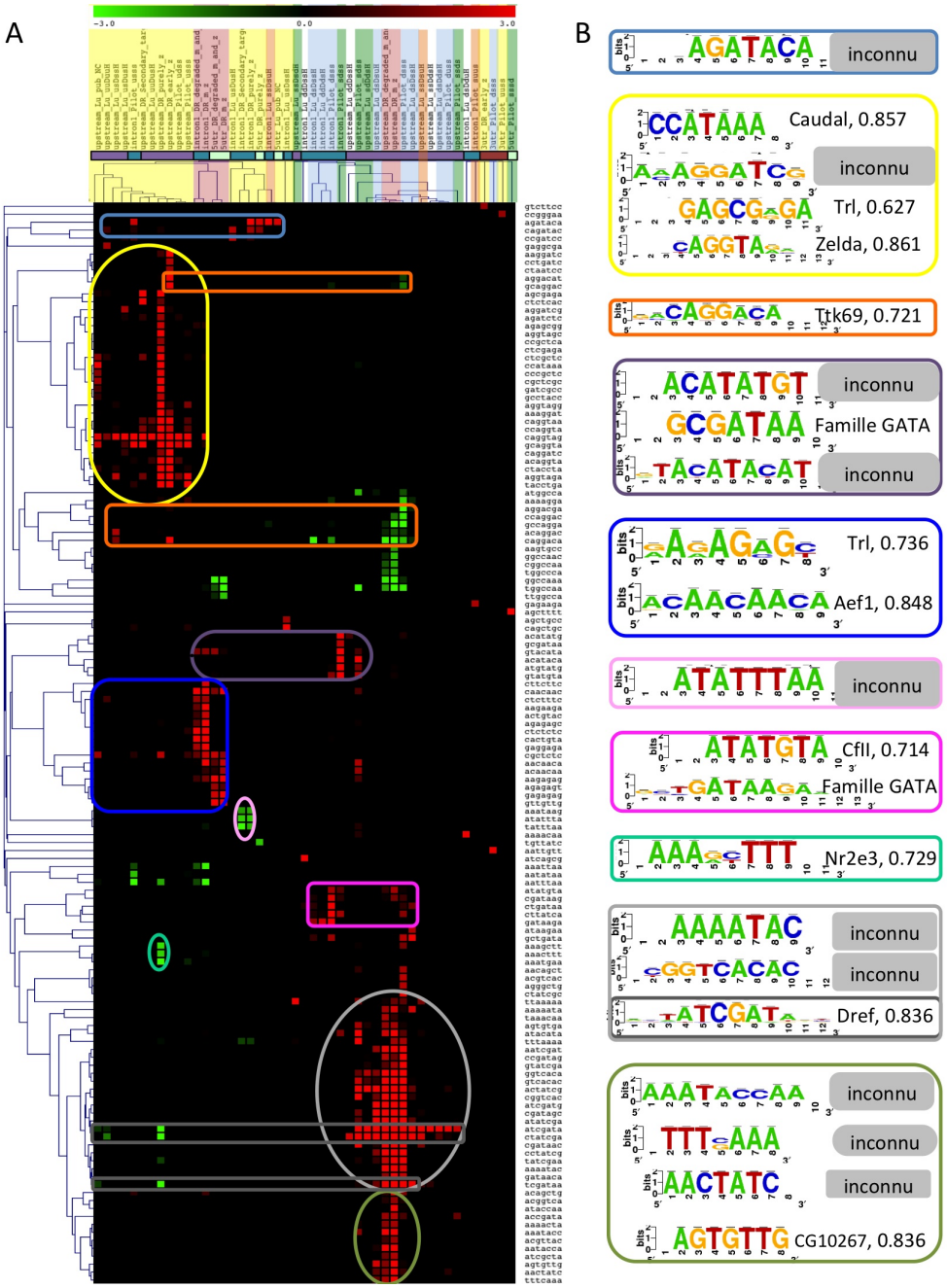


FIGURE 3.4: Double classification hiérarchique des oligonucléotides et des groupes de gènes en fonction de la significativité de la surreprésentation des oligonucléotides. - A. Les colonnes correspondent aux clusters de co-expression, plus exactement des différents ensembles de régions non codantes associées aux gènes de ces clusters en fonction de la significativité de la sous- (vert) ou surreprésentation (rouge) des oligos (lignes). La nature des séquences non codantes est indiquée par la barre sous les noms (premier intron : vert, 5'UTR : jaune pale, 3'UTR : rouge foncé et région 5 Kb en amont du TSS : violet). Les couleurs surlignant les noms des clusters correspondent à la légende de la figure 2.7. Clusterinring hièrarchique : corrélation de Pearson et règle d'agglomération moyenne B. Motifs résultant de l'assemblage des oligos surreprésentés dans chaque cadre de la couleur correspondante dans la heatmap. Le résultat de la comparaison avec des motifs connus est indiqué à côté de chaque logo.

référence à la coloration des clusters dans la figure 3.4A, les groupes jaune, bleu clair et orange correspondent aux clusters zygotiques, pour lesquels certains motifs issus de l'assemblage des mots chevauchants correspondent à des facteurs pertinents connus pour être actifs dans le blastoderme. En effet, nous retrouvons le motif de liaison de Zelda, un facteur impliqué dans l'activation zygotique mineure des gènes (68, 82, 83). En accord avec une étude qui vient d'être publiée (96), le motif de liaison de Zelda apparaît de façon générale dans les régions en amont des gènes de tous les clusters zygotiques (et dans les premiers introns du cluster "$usss$"), sans restriction aux gènes induits durant le stade pré-blastoderme cellulaire. Le mot 'CAGGTAG' (correspondant à une la variante la plus utilisée du motif de liaison de Zelda) est également trouvé dans les premiers introns du cluster mixte maternel-zygotique. Dans le cluster de gènes purement zygotiques, nous retrouvons également un motif proche de celui associé à Caudal, un facteur de transcription maternel connu pour son rôle dans la segmentation de l'embryon.

Deux autres motifs également surreprésentés dans ces clusters ne ressemblent à aucun motif répertorié dans les bases de données de référence (FlyFactorSurvey (39), JASPAR5 (38)). Le motif 'AGATACA' est particulièrement surreprésenté dans les premiers introns, ainsi que dans les régions en amont du le cluster "uu_Duu_H". En fait, ce motif pourrait être impliqué dans le phénomène de transvection ("chromosome pairing") (REF : Lewis et al, 1995) et ainsi participer au rapprochement physique de différentes régions de régulation. Le second motif ('AmAAGGATCG') est surreprésenté dans les clusters de gènes purement zygotiques et dans le cluster "us_Dus_H" suggérant une implication dans le contrôle de l'activation de la transcription dépendante du ratio NC.

Le motif de liaison de Trithorax-like (Trl ou facteur GAGA) est également surreprésenté (de façon moins significative) dans les régions en amont des gènes purement zygotiques, cibles secondaires soumises au ratio NC, selon Lu et collaborateurs. Le motif de liaison de Trl apparait aussi dans un groupe de mots associés aux 5'UTR et aux premiers introns des gènes des clusters mixtes (figure 3.4, groupe bleu foncé). Trl est un facteur maternel général agissant à divers niveaux de la régulation transcriptionnelle ((co-)facteur de transcription (97), remodelage de la chromatine (98, 99, 100)). L'implication de ce facteur semble être générale au niveau de l'AGZ. Son rôle sera discuté plus loin dans le manuscrit.

Toujours au sein des 5'UTR et premiers introns des clusters mixtes, nous avons mis en évidence un motif ressemblant au motif de liaison d'Aef1. Ce facteur est un répresseur maternel de la transcription dont la quantité baisse entre 2 et 4h après la fécondation (Flybase, données de RNA-seq (101)). Ceci suggère que les ARNm synthétisés à partir du gène aef1 sont dégradés.

La protéine Aef1 pourrait être titrée par la quantité croissante d'ADN au fur et à mesure des divisions mitotiques et, tant que son niveau est important dans l'embryon, Aef1 pourrait donc contrecarrer l'action activatrice de Trl.

Un dernier motif correspondant au motif de liaison de Tramtrack (TTK) apparaît dans les séquences en amont de gènes du cluster "Lu us_Dss_H" et du cluster regroupant les gènes précoces définis par De Renzis et al. (68) (groupe orange). TTK est connu pour être un répresseur maternel dont la titration au cours des cycles mitotique permet l'activation des gènes zygotiques (78). De plus, il a été montré que TTK est capable d'inhiber l'activation médiée par Trl (dont le motif est également surreprésenté dans ce cluster) au niveau du promoteur de even-skipped (eve) par une interaction protéine-protéine directe (102). Ce motif est sous-représenté de façon significative (significativité de 5.01) dans les séquences en amont des gènes des clusters maternels et mixtes. Comme TTK est un facteur maternel, si les gènes exprimés au niveau maternels contenaient des sites de liaison de ce facteur, ils pourraient être réprimés et du coup ne pourraient pas assurer leur(s) fonction(s).

Dans les gènes des clusters mixtes, nous trouvons également un motif sous-représenté (significativité 3.98) dans les régions en amont et dans les 5'UTR. Ce motif (ctttGGCCaaa) ne correspond à aucun des motifs annotés dans les bases de données. En outre, ces séquences comportent trois motifs surreprésentés (groupe violet), dont deux sont inconnus ("ACATATGT", "TACATACAT"), alors que le troisième ressemble au motif de liaison de la famille de facteurs GATA. Quatre facteurs de cette famille existent chez Drosophila melanogaster : Serpent, GATAe, Grain et Pannier. Aucun des transcrits synthétisés à partir de ces gènes n'est présent au niveau maternel : les facteurs Serpent, Grain et Pannier sont purement zygotiques (Serpent est en fait activé de façon secondaire par des facteurs zygotiques), alors que GATAe n'appartient à aucun cluster et aucun transcrit n'a été détecté dans les expériences de Gelbart et collaborateurs (101). La surreprésentation de ce motif dans les 5'UTR suggère une implication potentielle dans une activation zygotique plutôt que maternelle de ces gènes. Grain et/ou Pannier pourraient donc être impliqués dans l'activation des cibles secondaires appartenant aux clusters mixtes.

Enfin, les groupes rose, gris et kaki de la figure 3.4 sont associés aux séquences non codantes des clusters maternels et aux séquences en amont des gènes des clusters mixtes. Le groupe gris comporte trois motifs surreprésentés dans la majorité des clusters. Deux d'entre eux sont inconnus ("GGTCACAC" et "AAAATAC"). Le troisième correspond au motif de liaison de Dref appelé DRE ("DNA replication-related element") et ayant comme consensus

"TATCGATA". Ce motif est par ailleurs sous-représenté dans les séquences en amont des gènes purement zygotiques et des gènes appartenant au cluster "uu_Duu_H" dépendant du ratio NC. Dref est un facteur maternel impliqué dans l'activation de gènes impliqués dans la réplication. Hirose et collaborateurs (103) ont montré que la protéine Dref s'accumule dans les noyaux à partir du cycle mitotique 8 (début de la formation du blastoderme syncytial) et ont suggéré que Dref permettrait la bonne coordination de l'activation zygotique de ces gènes. D'autre part, BEAF32 (insulateur) reconnait également sur le motif "TATCGATA" (BEAF32 reconnait seulement le mot "tatcgata" et aucun autre variant). Hart et collaborateurs (104) ont montré que Dref et BEAF32 pouvaient entrer en compétition pour la liaison à ce motif dans les cellules en prolifération rapide. Ces résultats suggèrent que l'action de Dref et/ou BEAF32 n'est pas restreinte aux gènes reliés à la réplication. La déplétion de ce motif dans les gènes du cluster "uu_Duu_H" et dans les gènes purement zygotique pourrait empêcher l'activation aberrante de ces gènes.

Le groupe rose associe des motifs présents dans les régions en amont, dans les premiers introns et 5'UTR des clusters maternels "Pilot $dsss$" et "Lu ds_Dss_H", ainsi que dans les séquences en amont (uniquement) des gènes des clusters mixtes. Le premier motif correspond au motif de liaison de CF2, qui est connu pour être actif durant l'oogenèse et pour participer à la polarisation dorso-ventrale. Le second motif correspond au motif de liaison des facteurs de transcription de la famille GATA. Comme vu précédemment, les facteurs de cette famille ne sont pas actifs dans l'oeuf. Ce motif pourrait être en fait une partie du motif DRE.

Le groupe vert kaki contient en fait deux sous-groupes. Le premier est associé aux séquences en amont des gènes des clusters mixtes, où un motif est mis en évidence, mais qui ne correspond à aucun motif connu (premier motif de la boite de même couleur dans la figure 3.4B). Le deuxième sous-groupe, composé seulement de trois motifs isolés, est aussi associé au cluster "Pilot $dsss$" (en plus des clusters mixtes). Deux des motifs découlant des ces mots ne sont pas connus, alors que le troisième correspond au motif de liaison de CG10267 (Zif). Zif semble être exprimé au niveau maternel et est impliqué dans la polarité et la prolifération neuronale. Son rôle ici n'est pas évident.

L'analyse de la surreprésentation des dyades (annexe 6.6) met en évidence les mêmes motifs que précédemment. Particulièrement pour les clusters zygotiques. Pour les clusters maternels, les motifs les plus significatifs (DRE et "GGTCACAC") sont retrouvés, la plupart des autres motifs sont riches en A/T.

3.1.2.4 Sélection des gènes pour l'étude de l'AGZ

L'analyse précédente m'a permis de mettre en évidence le fait que les clusters zygotiques partagent plusieurs motifs, j'ai donc décidé de les regrouper pour les analyses suivantes. Quant aux gènes maternel-zygotiques, ils sont régulés d'une façon particulière. En effet, les motifs surreprésentés dans leurs séquences en amont sont partagés avec les clusters maternels du type "$dxxx$", alors que les motifs surreprésentés dans les régions 5'UTR et les premiers introns sont plutôt partagés avec les clusters zygotiques (motif de liaison de Trl et de Zelda). À cause de cette dualité, j'ai décidé de ne pas étudier les gènes maternel-zygotiques pour l'analyse de la régulation de l'AGZ.

Je me suis demandé si les gènes induits plus tardivement étaient régulés de la même façon que les gènes des clusters zygotiques. Dans l'analyse de surreprésentation de motifs, seul les gènes du cluster "Lu $ss_D su_H$" ont des motifs surreprésentés. Ce dernier regroupe les gènes activés durant la phase tardive de cellularisation. En ce qui concerne les dyades, quatre clusters contiennent au moins un motif dont la significativité est ≥ 2, mais leur regroupement est mal défini (ils font partie des branches non-résolues de l'arbre obtenu par la classification hiérarchique, voir annexe 6.6A). Ces clusters n'ont pas de motifs en commun avec les gènes des clusters zygotiques. Ils n'ont en fait que très peu de motifs surreprésentés. Ces gènes ne seront donc pas considérés dans les analyses suivantes.

Pour résumer, je retiens dix clusters : "Pilot $usss$" (87 gènes), "Pilot $udss$" (23 gènes), "Pilot $uuss$" (23 gènes), "Pilot $ussd$" (16 gènes), "Lu $us_D us_H$" (14 gènes), "Lu $us_D ss_H$" (24 gènes), "Lu $uu_D us_H$" (27 gènes), "Lu $uu_D uu_H$" (60 gènes), "DeRenzis purement zygotiques" (325 gènes), "DeRenzis zygotiques précoces" (60 gènes). Les cibles secondaires purement zygotiques sont incluses dans le cluster de gènes purement zygotiques. J'ai obtenu ainsi un cluster de 417 membres que je nommerai le "cluster AGZ".

3.2 Analyse du cluster AGZ

3.2.1 Analyse fonctionnelle avec Gene Ontology

Afin d'étudier les fonctions des gènes présents dans le cluster AGZ, j'ai analysé l'enrichissement du cluster en termes Gene Ontology (GO) avec l'outil *compare-classes* (94) et les outils en ligne *DAVID* (http ://david.abcc.ncifcrf.gov/) (105, 106) et *GOToolBox* (http ://genome.crg.es/GOToolBox/) (107).

J'ai paramétré les programmes *compare-classes*, textitDAVID et *GOToolBox* afin d'obtenir des résultats comparables. Cependant, chaque programme a des particularités au niveau des calculs statistiques de significativité. Les trois algorithmes basent leur calcul de significativité sur la probabilité hypergéométrique mais présentent certaines différences dans le calcul des scores de significativité. Par exemple, *DAVID* utilise le score EASE (108), qui consiste à modifier le test exact de Fischer (lui même basé sur la loi hypergéométrique) en procédant à l'extraction d'un élément et en recalculant la probabilité exacte de Fisher. Cela permet de pénaliser la significativité des catégories contenant peu de membres. En effet, si une catégorie contient peu de membres et qu'on les retrouve tous dans un jeu d'une centaine de gènes, ceci peut-être du au hasard. Cependant, la présence des quatre membres peut avoir une pertinence biologique. Une différence majeure existe entre *GOToolBox* et les deux autres méthodes, en ce qui concerne le choix du groupe de gènes de référence. En effet, *DAVID* (par défaut) va calculer les probabilités en fonction du nombre total de gènes annotés dans au moins une des classes analysées, choix que j'ai également fait pour mes analyses avec compare-matrices. Par contre, *GOToolBox* utilise une référence globale (nombre total des gènes de la drosophile). Ceci entraîne une significativité va être plus importante, comme on peut l'observer dans le tableau 3.1. Le fait de prendre en compte tous les gènes est cependant contestable, car ceux qui n'ont aucune annotation GO ne pourront jamais apparaître à l'intersection d'un cluster et d'une classe GO, et ne font donc pas partie de "l'univers" accessible à la distribution hypergéométrique. Pour les trois méthodes, la correction de Bonferroni pour les tests multiples a été appliquée.

GOToolBox, *DAVID* et *compare-classes* revoient 226, 207 et 184 termes significativement enrichis (e-value < 0.01) respectivement. J'ai comparé les résultats grâce à compare-scores en utilisant les noms de terme GO afin d'éviter les incompatibilités dues aux différences de versions (identifiants obsolètes dans la base de *GOToolBox*). J'ai ainsi identifié 149 termes significativement enrichis dans les trois analyses (les différences s'expliquent vraisemblablement par l'absence de filtrage des annotations pour *DAVID*, par le choix de la référence et la version ancienne de l'ontologie et des annotations pour GOToolBox, alors que *compare-classes* renvoie le moins de termes car c'est la méthode la plus restrictive par le choix de la référence et par le filtrage). Les 50 premiers termes classés selon la significativité de compare-classes sont montrés dans le tableau 3.1 (la totalité des résultats sont disponibles à l'adresse http ://rsat.bigre.ulb.ac.be/ elodie/these/resultats_supplementaires/functional_analysis/).

Le cluster ZGA est donc enrichi pour beaucoup de termes. L'analyse séparée des clusters primaires ne donne pas autant de termes. Le cluster de gènes purement zygotiques défini par

De Renzis et collaborateurs contient le plus grand nombre termes enrichis (90 termes). Nous retrouvons ces termes enrichis dans le cluster ZGA. Plusieurs termes enrichis sont relatifs à la morphogénèse, à la segmentation, à la gastrulation, à la régulation des gènes etc. L'ensemble de ces termes sont cohérents avec les processus développementaux du stade étudié, pendant lequel l'embryon subit énormément de changements, tant au niveau morphologique que transcriptionnel. C'est à ce moment qu'à lieu la mise en place et la différenciation des principaux tissus. J'ai procédé à un contrôle négatif avec 10 clusters composés de gènes sélectionné au hasard. Aucun terme n'est ressorti, ni avec *compare-classes*, ni avec *DAVID*. En revanche, *GO-ToolBox* renvoie quelques termes (très généraux) enrichis, ce qui confirme l'importance de la prise en compte du nombre de gènes annotés plutôt que totaux.

TABLE 3.1: Comparaison de la composition de gènes entre le cluster ZGA et les classes de la Gene Ontology. La table rapporte les e-valeurs de trois outils (*compare-classes*, *GOToolBox* et *DAVID*) pour les 50 termes les plus significatifs (selon *compare-classes*).

Terme GO	*compare-classes*	GOToolBox	DAVID
BP : anatomical structure morphogenesis	8.00^{-31}	1.55^{-52}	7.15^{-36}
BP : multicellular organismal development	9.50^{-30}	6.89^{-60}	2.00^{-43}
BP : organ development	4.10^{-29}	3.84^{-60}	1.91^{-45}
BP : biological regulation	1.20^{-28}	3.84^{-39}	1.69^{-22}
BP : anatomical structure development	8.40^{-28}	4.51^{-63}	8.26^{-45}
BP : developmental process	7.60^{-27}	3.04^{-58}	4.16^{-39}
BP : system development	1.60^{-25}	1.83^{-59}	1.22^{-47}
BP : regulation of biological process	2.70^{-25}	1.70^{-36}	2.00^{-22}
BP : cell fate commitment	8.80^{-23}	1.95^{-34}	5.09^{-30}
BP : pattern specification process	8.90^{-22}	2.69^{-32}	3.84^{-30}
BP : generation of neurons	1.00^{-21}	3.76^{-27}	4.57^{-22}
BP : regulation of cellular process	3.50^{-21}	9.42^{-31}	2.08^{-19}
BP : regionalization	1.40^{-20}	1.35^{-31}	3.33^{-29}
BP : embryonic morphogenesis	5.50^{-20}	5.62^{-23}	1.57^{-21}
BP : cell fate determination	1.10^{-19}	9.72^{-29}	9.49^{-25}
BP : regulation of transcription. DNA-dependent	7.00^{-18}	1.94^{-10}	1.10^{-18}
BP : cellular developmental process	2.70^{-17}	1.22^{-47}	2.57^{-28}
BP : cell differentiation	4.00^{-17}	1.02^{-43}	1.51^{-29}
BP : regulation of RNA metabolic process	9.10^{-17}	8.40^{-11}	4.34^{-18}
BP : organ morphogenesis	9.90^{-17}	1.66^{-40}	4.68^{-34}
BP : gastrulation	1.30^{-16}	1.36^{-12}	6.36^{-12}
BP : tissue development	1.70^{-16}	7.51^{-28}	1.03^{-20}
BP : regulation of macromolecule biosynthetic process	5.00^{-16}	9.45^{-15}	2.79^{-16}
BP : nucleotide and nucleic acid metabolic process	8.50^{-16}	8.36^{-16}	1.14^{-17}
BP : multicellular organismal process	1.50^{-15}	9.32^{-51}	3.93^{-28}
BP : regulation of gene expression	4.60^{-15}	3.22^{-17}	1.71^{-16}
BP : tissue morphogenesis	1.10^{-14}	3.61^{-21}	3.69^{-17}
BP : regulation of cellular biosynthetic process	3.20^{-14}	7.52^{-14}	9.52^{-16}
BP : regulation of biosynthetic process	3.80^{-14}	7.52^{-14}	9.52^{-16}
BP : neuron differentiation	5.60^{-14}	2.51^{-21}	1.86^{-16}
BP : regulation of macromolecule metabolic process	1.40^{-13}	1.08^{-15}	4.90^{-15}
BP : epithelium development	2.90^{-13}	1.02^{-16}	1.40^{-14}
BP : embryonic pattern specification	3.40^{-13}	7.48^{-23}	3.03^{-19}
BP : regulation of cellular metabolic process	4.70^{-13}	8.05^{-15}	1.14^{-15}
BP : morphogenesis of an epithelium	6.30^{-13}	1.02^{-16}	2.08^{-14}
BP : nervous system development	1.10^{-12}	7.18^{-31}	3.01^{-30}
BP : regulation of primary metabolic process	1.70^{-12}	1.08^{-13}	1.55^{-14}
BP : cell development	2.00^{-12}	1.67^{-29}	1.15^{-14}
BP : segmentation	2.90^{-12}	1.37^{-22}	7.56^{-19}
BP : post-embryonic development	3.40^{-12}	5.26^{-19}	2.85^{-15}
BP : regulation of transcription from RNA polymerase II promoter	3.70^{-12}	8.56^{-14}	5.88^{-16}
BP : locomotion	3.80^{-12}	4.49^{-10}	2.10^{-08}
BP : regulation of metabolic process	6.70^{-12}	1.40^{-15}	1.08^{-14}
BP : neuron development	7.80^{-12}	1.55^{-19}	1.67^{-13}
BP : imaginal disc development	9.90^{-12}	3.48^{-19}	8.50^{-18}
BP : cell morphogenesis involved in differentiation	1.40^{-11}	1.87^{-16}	3.85^{-12}
BP : cell fate specification	1.80^{-11}	1.86^{-14}	4.84^{-11}
BP : neurogenesis	2.20^{-11}	2.77^{-28}	1.83^{-22}
BP : sensory organ development	3.10^{-11}	8.94^{-20}	2.08^{-16}
BP : blastoderm segmentation	4.20^{-11}	3.83^{-20}	1.90^{-18}

3.2.2 Analyse des éléments cis-régulateurs

Afin de découvrir des éléments et modules cis-régulateurs (CRM pour "*Cis-Regulatory Element*") potentiellement impliqués dans l'AGZ, j'ai procédé à diverses analyses des régions non-codantes des gènes du cluster AGZ. J'ai tout d'abord recherché des motifs significativement surreprésentés dans les régions 5 kb en amont du TSS, dans les premiers introns et dans les 5'UTR.

Afin d'analyser l'enrichissement en motifs de liaison connus, j'ai utilisé l'outil *CisTargetX*, qui inclu une collection très étendue de matrices poids-positions pour la drosophile (109) (http ://med.kuleuven.be/cme-mg/lng/cisTargetX/). En intégrant ces résultats et ceux de l'analyse des clusters primaires, j'ai sélectionné un jeu de matrices qui seront ultérieurement utilisées pour scanner les régions non codantes et détecter des régions significativement enrichies en sites (CRER pour "*Cis-Regulatory element Enriched Region*"), considérées comme des CRM potentiels.

3.2.2.1 Découverte de motifs

La première étape consiste à récupérer les séquences non codantes en amont du TSS, ainsi que celles des premiers introns et des UTR. J'ai choisi de récupérer les 5kb en amont du TSS, car on retrouve fréquemment des modules cis-régulateurs sur plusieurs Kb en amont des gènes,. En effet, il a été montré que les CRM pouvaient se trouver tant dans les régions inter-géniques que dans les introns (particulièrement dans le premier) et les UTR (110). Pour de cela j'ai utilisé *retrieve-ensembl-seq* avec les options me permettant d'optimiser l'analyse de découverte de motifs (cf. section 3.3.1 et figure 3.7). Afin de découvrir les motifs les plus significativement surreprésentés dans les séquences non codantes des gènes du cluster AGZ, j'ai utilisé le pipeline gene-cluster-motifs. J'ai testé par ailleurs l'algorithme MEME (21), mais à cause du grand nombre de séquences à analyser, ce logiciel n'a jamais donné de résultats (le temps de calcul augmente en fonction du carré de la taille du jeux de séquences). En revanche, *gene-cluster-motifs* permet de tester facilement et rapidement différents paramètres à différentes étapes du pipeline (voir section 3.3.2). La recherche de motifs connus avec *CisTargetX* a été opérée avec les paramètres par défaut.

3.2.2.2 Sélection des motifs parmis les matrices découvertes

108 matrices ont été découvertes (19 à partir d'*oligo-analysis* et 79 à partir de *dyad-analysis*). Ce grand nombre de matrices a vraisemblablement deux origines. Premièrement, plusieurs assemblages alternatifs peuvent être formés à partir de groupes de mots partiellement chevauchants. Ensuite, l'assemblage des dyades renvoie certains motifs très longs (environ 25 nucléotides) composés de plus de 10 dyades, mais ces chevauchements n'ont pas forcément de sens. En effet la plupart des matrices résultantes sont construites à partir de très peu de sites (entre 1 et 50 sites). J'ai tout d'abord écarté les matrices construites à partir de moins de 100 sites. Ensuite, j'ai procédé à un réalignement des matrices restantes en utilisant *compare-matrices*. Lorsque les matrices s'alignent avec une corrélation de Pearson normalisée supérieure à 0.8, j'ai gardé la matrice la moins dégénérée. Un avantage de cette procédure est qu'elle indique que certains motifs ont été découverts dans plusieurs types de séquences. J'ai ainsi sélectionné 9 matrices que j'ai ensuite comparées à différentes bases de motifs connus (corrélation normalisée > 0,7).

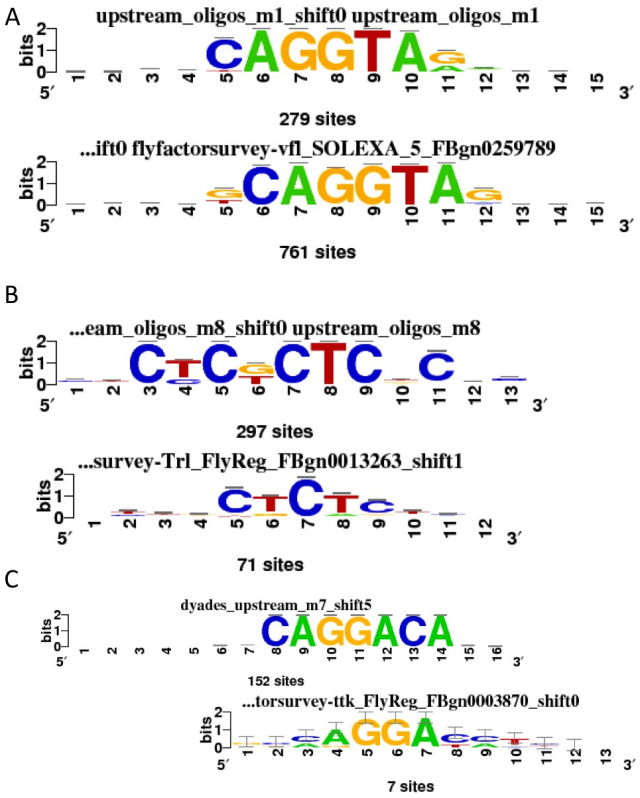
3.2.2.3 Motifs connus retrouvés par l'approche *de novo* et par *CisTargetX*

Le motif de liaison de Zelda est le motif le plus significatif dans les deux approches (figure 3.5A et annexe 6.7). Ce facteur étant connu pour être impliqué de façon générale dans l'AGZ (96), c'est un bon indicateur de la pertinence du choix des gènes AGZ.

L'Annexe 8 montre que le motif le plus enrichi détecté par CisTargetX (GAGAGA) correspond au motif de liaison de Trl. En ce qui concerne les résultats de la découverte de motifs (figure 3.5B), Trl arrive en troisième position.

3.2.2.4 Autres motifs connus

Le motif de liaison de TTK apparaît comme surreprésenté en découverte de motifs, mais n'est pas détecté par CisTargetX. Lors de l'analyse des clusters primaires dans le chapitre, ce motif a été détecté dans les séquences du cluster regroupant les gènes zygotiques précoces de De Renzis, ainsi que dans celles du cluster regroupant les gènes activés durant la phase lente de cellularisation ("Lu $us_{D}ss_{H}$"). En particulier, le mot "caggac" est significativement surreprésenté pour le cluster AGZ (avec une significativité de 3.61, un peu plus faible que dans les clusters plus restreints) (figure 3.5C). De manière intéressante, un antagonisme entre Trl et



oligo-analysis	40.71
Base de données	FlyFactorSurvey
Nom du facteur	**Zelda**
Ncor	0.887
CistargetX	12.76

oligo-analysis	4.83
Base de données	FlyFactorSurvey
Nom du facteur	**Trl**
Ncor	0.710
CistargetX	3.466

dyad-analysis	3.61
Base de données	FlyFactorSurvey
Nom du facteur	**Ttk**
Ncor	0.751

FIGURE 3.5: Correspondances entre motifs découverts et motifs connus. - Pour trois motifs découverts, un alignement avec un motif connu est présenté à gauche, en face d'un tableau donnant la significativité du motif découvert, le facteur se liant au motif connu correspondant, la corrélation normalisée (Ncor) de l'alignement, et le score d'enrichissement calculé par CisTargetX pour le même motif connu ou pour un motif similaire.

TTK a été récemment mis en évidence (102) dans la régulation du gène *eve*. J'ai donc décidé de garder ce motif pour les analyses suivantes.

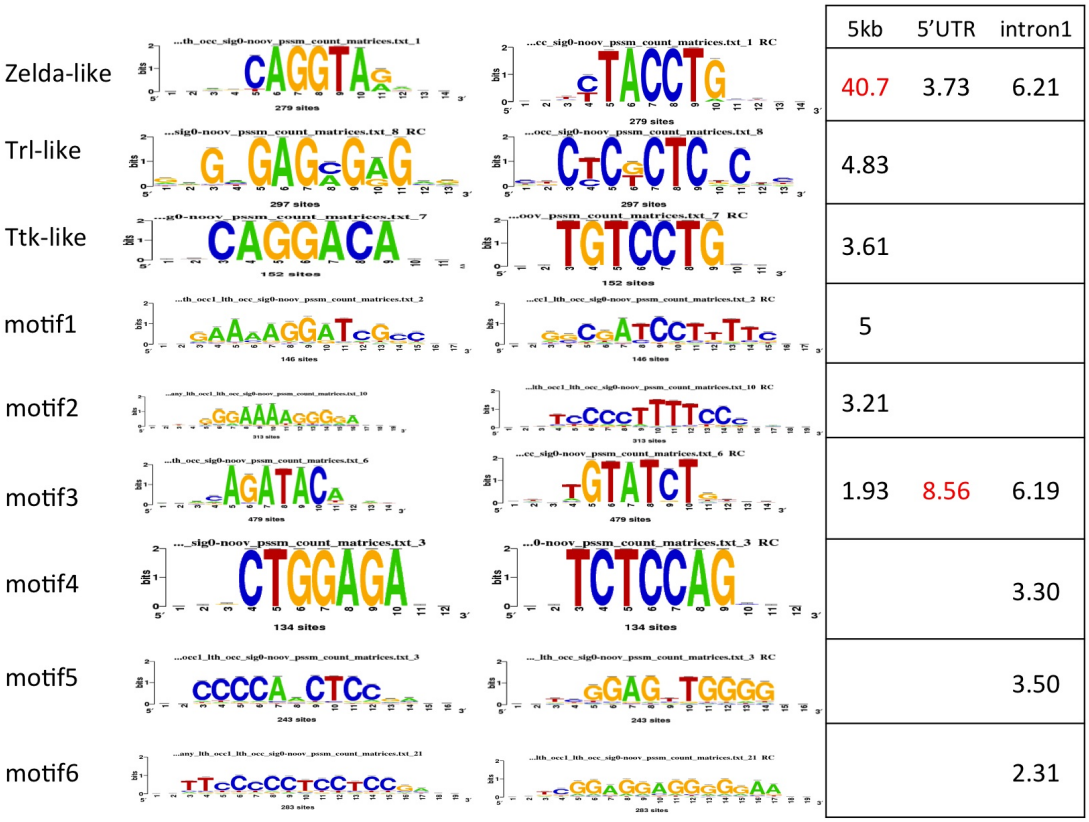
Comme le montre l'annexe 6.7, plusieurs motifs de liaison de FT impliqués dans la segmentation (Dorsal, Krüpple, Bicoid) sont également enrichis dans les séquences non codantes des gènes AGZ. Cependant, ces FT sont impliqués dans la différenciation antero-posterieur précoce, or je recherche des facteurs impliqués dans le mécanisme ubiquiste de l'activation zygotique. Donc, j'ai écarté les motifs reconnus par les gènes de segmentation pour les analyses subséquentes. Le motif de liaison de Su(H) ressort également. Durant la cellularisation, ce facteur est impliqué dans la définition du mésectoderme (cellules entourant le neuroectoderme) en inhibant l'action de Twist et Dorsal dans le mésoderme et permettant le recrutement du fragment intracellulaire de Notch dans le mésectoderme (annexe 6.8) (111). Son action localisée n'en fait pas un bon candidat pour la régulation globale de l'AGZ.

3.2.2.5 Motifs inconnus

Les motifs inconnus "motif1" et "motif3" (figure 3.6) ont été identifiés lors de l'analyse des clusters primaires, et le regroupement des clusters a renforcé leur significativité. Les quatre autres motifs sont spécifiques au cluster AZG. Les motifs "motif5" ("CCCCANCTCC") et "motif6" ("CCCCTCCTCC") se ressemblent par la présence de nombreux "C", mais diffèrent par les lettres (A ou T) qui les séparent. La présence d'une seule position différente dans les sites de liaison peut entrainer une reconnaissance par des facteurs différents (par exemple Zelda et Snail reconnaissent respectivement et exclusivement "CAGGTAG" et "CAGGTG"). Nous avons donc retenu ces deux motifs pour la prédiction de CRM.

3.2.2.6 Détection de modules cis-régulateurs (CRM) potentiels

Afin d'identifier des CRM potentiels dans les région non codantes des gènes AGZ (5kb en amont du TSS, premier intron, 5'UTR), j'ai utilisé *matrix-scan* (25) pour détecter les régions enrichies en sites (CRER : *Cis- Regulatory element Enriched Region*). Lors de cette analyse, il s'est avéré que les motifs 5 et 6 reconnaissent partiellement les mêmes sites, en dépit des différences pour les résidus A ou T centraux. Si les sites reconnus par ces matrices sont chevauchants, le nombre de sites dans une région est alors sur-estimé, ce qui fausse les résultats du test binomial. Cependant, lors du scan des régions non codantes pour la recherche de CRER, j'ai utilisé l'option qui permet d'écarter les sites chevauchants pour le calcul de la significativité



	5kb	5'UTR	intron1
Zelda-like	40.7	3.73	6.21
Trl-like	4.83		
Ttk-like	3.61		
motif1	5		
motif2	3.21		
motif3	1.93	8.56	6.19
motif4			3.30
motif5			3.50
motif6			2.31

FIGURE 3.6: Liste des motifs sélectionnés pour la prédiction de CRM. - Les chiffres indiquent la significativité de la surreprésentation des motifs selon le type de séquences non-codantes.

de l'enrichissement en site de la région (cf. 3.3.3). Le maintien des deux matrices partiellement similaires ne pose donc pas de problème pratique pour la prédiction de CRM.

J'ai ainsi détecté 932 CRM potentiels non chevauchants associés à 280 gènes : 538 CRM dans les 5kb en amont du TSS, 172 dans les premiers intron et 528 dans les 5'UTR. L'analyse de ces CRM sera approfondie dans le chapitre suivant.

3.3 Matériel et Méthodes

3.3.1 Récupération des séquences non-codantes

Pour récupérer les séquences non codantes, j'ai utilisé l'outil *retrieve-ensembl-seq* de la suite logicielle RSAT (112) avec les options suivantes :

- *-alltranscripts* : permet de prendre en compte tous les transcrits, en fusionnant les régions sélectionnées (par exemple en amont) relatives à ces transcrits. En effet, quand un gène présente des transcrits multiples du fait de la présence de promoteurs alternatifs, choisir l'un d'eux (le plus long en général) n'est pas justifié. Idéalement, les données de puces ou les profils d'occupation de l'ARN polymérase II doivent pouvoir être utilisées pour identifier lequel des transcrits est le plus probablement exprimé. Cependant, le fait de les considérer tous n'introduit qu'un faible biais lors de la découverte de motifs. De plus, les gènes du cluster ZGA expriment en moyenne 1,8 transcrits alternatifs.
- *-nogene* (figure 3.7, puce 1) : permet de tronquer la séquence en amont du TTS si la longueur entre le gène en amont et le TTS du gène d'intérêt est plus courte que la valeur demandée.
- *-maskcoding* (figure 3.7, puce 2) : permet de masquer des séquences codantes (les séquences codantes et non codantes ont des compositions en nucléotides différentes).
- *-uniqseq* (figure 3.7, puce 3) permet de fusionner les régions chevauchantes. Cette option est essentielle pour éviter les redondances qui introduiraient un biais lors de la découverte de motifs surreprésentés.
- *-rm* : permet de masquer les régions répétées.

La figure 3.7 (puce 4) montre que les différents types de séquences non codantes peuvent se chevaucher. Ceci peut poser problème lors de la découverte de motifs dans la mesure où j'ai calculé le background (cf. 3.1.2.1) à partir des séquences non codantes de même type que les séquences analysées.

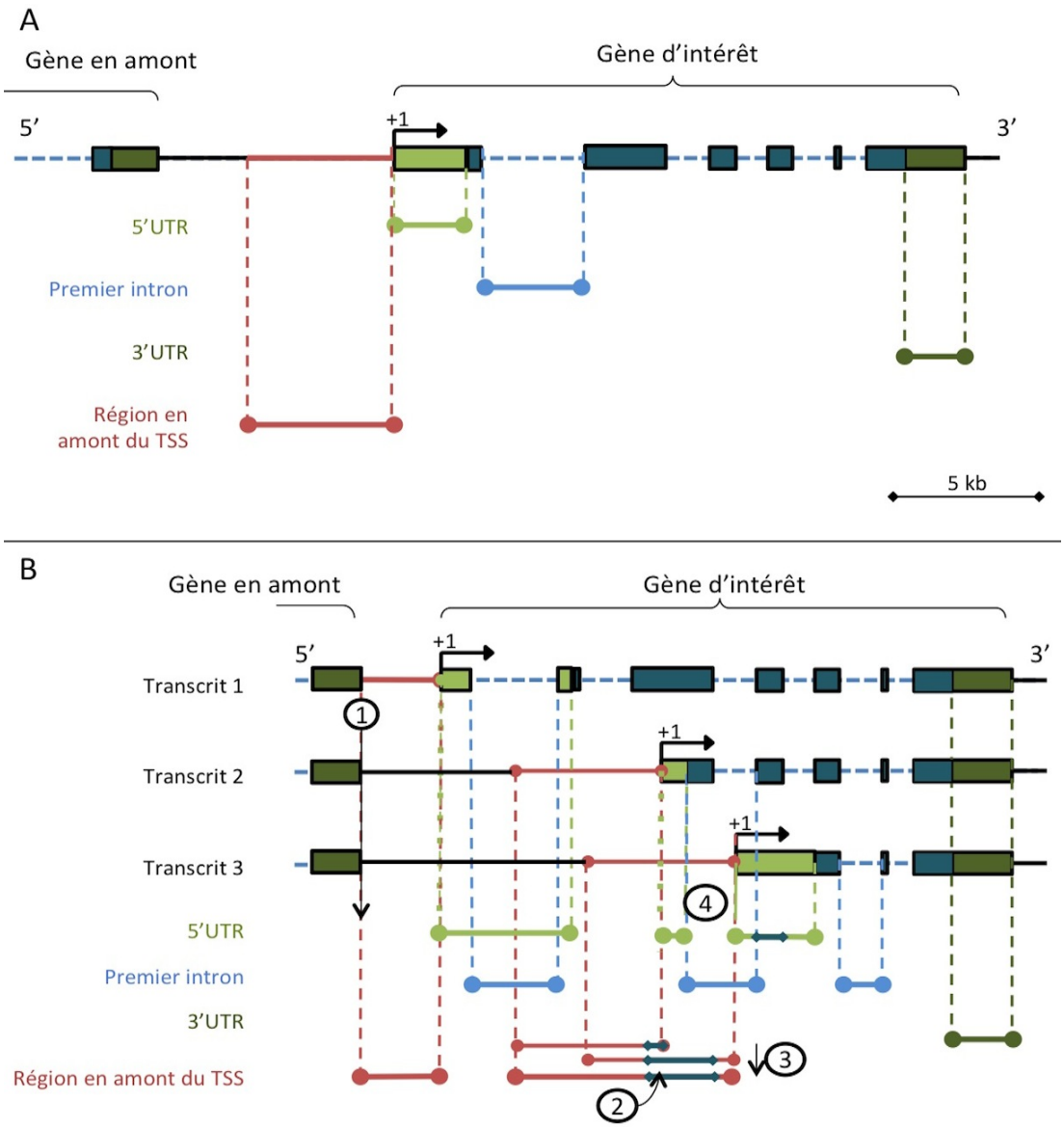


FIGURE 3.7: Représentation schématique des options utilisées dans *retrieve-ensembl-seq*. - A. Récupération des séquences non-codantes pour un gène d'intérêt. B. Le gène d'intérêt exprime trois transcrits alternatifs et tous sont pris en compte (-alltranscripts). Les introns sont représentés en bleu clair, les UTR en vert pâle et foncé, et les régions codantes en bleu-vert. (1) limitation de la séquence en fonction du premier gène en amont ; (2) masquage des séquences codantes ; (3) fusion des séquences chevauchantes provenant des transcrits alternatifs ; (4) Chevauchements des différents types de séquences (codante, UTR, promoteur, ...) générés par la présence des transcrits alternatifs.

3.3.2 Paramètres pour la découverte de motifs

Les algorithmes implémentés dans les outils *oligo-analysis* et *dyad-analysis* (113) sont basés sur le comptage de mots simples (*oligo-analysis*) ou de paires de mots espacés (*dyad-analysis*). Le nombre d'occurrences de chaque mot (ou dyade) d'une taille donnée est calculé dans le jeu de séquences à tester et comparé au nombre attendu au hasard, d'après un un jeu de séquences de référence (modèle de background). Le choix de ce modèle va donc être capital pour identifier les mots (ou dyades) les plus pertinent(e)s. La composition en nucléotides varie entre les espèces mais aussi selon le type de séquence (codantes , intronique, intergénique, ...). La significativité du nombre d'occurrences observée est estimée sur base de la distribution binomiale. Le programme retourne une P-valeur nominale (risque de faux-positifs pour un mot donné), une E-valeur (nombre de faux-positifs attendus pour un seuil de P-valeur donné) et un indice de significativité ($sig = -log10(E - value)$). J'ai gardé les mots (et dyades) dont la significativité est supérieure à 0.

3.3.2.1 *oligos-analysis*

J'ai recherché la surreprésentation d'oligomères de 7 nucléotides sur les deux brins en ne comptant pas les occurrences chevauchantes. J'ai utilisé un modèle de Markov de l'ordre le plus élevé possible (k-1) calibré sur un jeu de données externe, composé de toutes les séquences du même type que celles analysées. Par exemple, pour estimer la sur-représentation des heptamères (7nt) dans les introns du cluster AZG, j'ai considéré comme fréquence attendue au hasard, la fréquence des mêmes heptamères dans l'ensemble des introns de drosophile.

3.3.2.2 *dyad-analysis*

J'ai recherché les dyades formées de monades de 3 nucléotides espacées de 0 à 20 nucléotides en ne comptant pas les occurrences chevauchantes.Le modèle de background correspond à la fréquence de toutes les dyades possibles (xxxnxxx où x représente n'importe quelle base et n l'espacement entre les monades, avec $0 \leq n \leq 20$) dans l'ensemble des séquences de drosophile de même type que les séquences analysées.

3.3.2.3 *matrix-from-pattern*

Cet outil génère une série de matrices position-poids à partir d'une liste de mots surreprésentés produite par *oligo-analysis* et *dyad-analysis*. Il enchaîne les étapes suivantes :

- assemblage des mots (et dyades) chevauchants (*pattern-assembly*) ;
- construction de matrice de significativité (*convert-matrix*) ;
- scanning des séquences analysées (*matrix-scan-quick*) avec ces matrices afin d'obtenir des sites, qui sont ensuite alignés pour produire des matrices de comptage et les logos résultants (*convert-matrix*).

En outre, *oligo-analysis* permet de filtrer les motifs résultant des assemblages de mots les moins significatifs grâce au seuil appliqué sur le poids des sites à utiliser pour la construction des matrices de comptage. Le nombre de matrices en sortie ne correspond donc pas toujours au nombre d'assemblages de mots.

3.3.3 Paramètres pour la prédiction de sites et de modules cis-régulateurs (CRM)

La détection des sites et modules cis-régulateurs a été effectuée avec le programme *matrix-scan* (25). Ce programme supporte des modèles Markoviens, mais le temps de calcul augmente très rapidement avec l'ordre du modèle. Nous avonc donc choisi un modèle de Markov d'ordre m=2 plus faible que pour la découverte de motifs.

Les options utilisées sont :

- *-2str* : Scan sur les deux brins ;
- *-bginput -markov 2* : modèle de background calculé à partir des séquences avec un modèle de Markov d'ordre 2 ;
- *-uth pval 0.0001* : p-value maximale fixée à 10-4 pour la détection de sites ;
- *-lth crer_site_distance 6* : espace minimal de 6 pb entre les positions de début des sites pris en compte pour la détection de régions enrichies (prédiction de CRM) (pour éviter de prendre en compte des occurrences chevauchantes) ;
- *-lth crer_size 30 –uth crer_size 800* : taille des CRER entre 30 et 800 pb ;
- *-lth crer_sig 2* : significativité minimale de 2 pour l'enrichissement des régions.

3.3.4 Bases de données de PSSMs

3.3.4.1 FlyFactorSurvey

FlyFactorSurvey (39) (http ://pgfe.umassmed.edu/TFDBS/) fournit une base de motifs de liaison pour des facteurs de drosophile, en provenance de différentes sources :

- données de "bacterial one-hybrid" (B1H) (47, 48) ;
- données de "DNAse1 footprinting" (44) (disponibles à l'adresse http ://www.flyreg.org/) ;

– base de données TIFFIN (45), comprenant des motifs découverts à partir de promoteurs et validés par comparaison avec des motifs obtenus sur la base d'expériences de SELEX (JASPAR55) et de DNAse1 footprinting (http ://www.sanger.ac.uk/resources/databases/tiffin/index.jsp).

3.3.4.2 Jaspar core insect

Jaspar core insect (38) est disponible sur le site web de Jaspar (http ://jaspar.genereg.net/). Cette base regroupe des matrices construites en grande partie à partir d'expérience de SELEX, auxquelles s'ajoutent des matrices générées à partir de sites publiés.

3.4 Conclusion du chapitre

L'analyse des motifs surreprésentés dans les clusters primaires a révélé que les clusters maternels et zygotiques possèdent des éléments cis-régulateurs différents et exclusifs, alors que les clusters mixtes partagent des caractéristiques avec chaque groupe. En effet, les motifs surreprésentés dans les régions en amont du TSS des gènes maternel-zygotiques sont également surreprésentés dans les régions non codantes des gènes des clusters maternels. D'un autre côté, les motifs détecté dans les premiers introns des gènes maternel-zygotique sont également surreprésenté dans les régions non codantes des gènes des clusters zygotiques. Les motifs découverts dans les clusters primaires se sont avérés cohérents pour les clusters ayant des profils de transition comparables prédits à partir de différentes sources ("Pilot $usss$" et "Lu us_Dss_H" par exemple). Cette analyse m'a permis de définir plusieurs motifs cis-régulateurs potentiellement impliqués dans la régulation des gènes exprimés au niveau maternel et zygotique, répartis de manière différente dans les régions non codantes associées (promoteurs, introns, UTR). Ces motifs ressemblent pour une bonne partie à des consensus bien établis pour des facteurs tels que Zelda, Trl ou Caudal, impliqués dans le développement embryonnaire précoce de la drosophile, alors que d'autres motifs ne ressemblent à rien de connu et pourraient révéler l'intervention de facteurs de transcription en attente de caractérisation. J'ai enfin regroupé des clusters en combinant des critères d'expression, de composition en gènes et de motifs découverts, afin d'approfondir les étapes ultérieures de la recherche sur base d'un cluster cohérent correspondant précisément à l'AGZ. Par ailleurs, l'analyse statistique des annotations GO associées à ces gènes met en évidence les différents processus mis en place à ce stade du développement. Le regroupement des clusters zygotiques a renforcé la significativité des motifs détectés précédemment, à part dans le cas de Caudal dont le motif de liaison n'est plus détecté. Ainsi, cinq

motifs ont été redécouverts dans le cluster AGZ : trois motifs de liaison de facteurs connus (Zelda, Trl et TTK) et deux motifs inconnus.

4

Analyse génomique des profils de marques épigénétiques, d'occupation par des facteurs de transcription, et d'accessibilité de la chromatine

La découverte du motif de liaison de Trl m'a conduite à m'intéresser à la régulation de la transcription par le modelage de la structure de la chromatine. Comme nous l'avons vu dans l'introduction, la présence des FT et de leurs sites de liaison n'est pas suffisante pour expliquer la régulation de l'expression des gènes. En effet, l'accessibilité de ces régions joue un rôle important dans cette régulation.

Mes recherches bibliographiques autour de Trl m'ont conduit à suspecter une implication de la protéine CBP (CREB-Binding Protein), codé par le gène *nejire*. CBP fait partie d'un complexe de remodelage de la chromatine (TAC1) incluant des membres du groupe Trithorax, dont le recrutement au promoteur du gène hsp70 est facilité par Trl et HSP (114).Synthétisé au niveau maternel, CBP est un co-régulateur général de la transcription recruté à la chromatine par des centaines de facteurs. CBP joue un rôle décisif dans la différenciation cellulaire lors du développement (115) en participant à la régulation de plusieurs voies de signalisation (co-activateur de Mad dans la voie Dpp/Scw (116), de CI dans la voie Hedgehog (117), de Dorsal dans la voie Toll (118) ; répresseur de dTCF dans la voie Wnt/Wingless (119)). CBP agit au moins à deux niveaux de régulation. En effet, CBP est capable d'interagir directement ou indirectement (par l'intermédiaire des complexes médiateurs (120)) avec les facteurs gé-

néraux de transcription et l'ARN polymérase II (121) et fait ainsi le lien entre les facteurs de transcription et la machinerie basale de la transcription. CBP a également une activité acétyltransférase. En effet, CBP peut s'associer avec des membres du groupe Trithorax (TRX et UTX) pour contrer les effets répresseurs des complexes Polycomb PRC1 et PRC2, qui déacétylent et triméthylent la lysine K27 de l'histone H3 (H3K27) (122). En fait, UTX déméthyle H3K27me3 et TRX tri-méthyle la lysine K4 de H3, CBP cible particulièrement les H3K4me3 (123) et acétyle H3K27. A cet égard, dans le cadre du projet modENCODE, qui regroupe un grand nombre d'expériences de ChIP-seq à différents stades du développement de la drosophile (http ://www.genome.gov/modencode/), Nègre et collaborateurs ont récemment mis en évidence que CBP était associé aux CRM activateurs à tous les stades du développement, et que la présence de CBP était suffisante pour prédire de vrais CRM (124). Cette étude couvre également quatre marques de méthylation, deux marques d'acétylation de l'histone H3, l'ARN polymérase II (GSE16013), Polycomb-like (Pcl) (GSE23127), qui est un membre du groupe Polycomb et est associé au complexe PRC2 pour la tri-méthylation de H3K27 (125), et enfin Trl (GSM614652). Ils ont également trouvé une corrélation entre la présence de la marque H3K4me1 et des pics CBP avec l'activité des CRM. Enfin, leurs résultats suggèrent que les marques H3K4me3 et d'acétylation sont spécifiques des promoteurs actifs. Toutes les données sont disponibles en matériel supplémentaire de l'article de Nègre et collaborateurs (124).

Par ailleurs, j'ai récupéré des données d'accessibilité à la DNAse1 générée dans le cadre du projet Berkeley Drosophila Transcription Network Project (BDTNP) (126, 127). Les auteurs fournissent leurs données sous la forme de reads bruts (dans un format appelé SRA, pour"*sequence read archive*"), ainsi que sous la forme d'un fichier contenant les densités normalisées des reads, ou encore d'un fichier contenant les pics identifiés. Ces données sont disponibles sur le site http ://bdtnp.lbl.gov/Fly- Net/browseAccess.jsp.

Enfin, lors de la rédaction de ce manuscrit, des expériences de ChIP-seq avec le facteur Zelda ont été publiées (96) et j'ai donc pris en compte in extremis ces données.

Dans un premier temps, j'ai utilisé les pics provenant des expériences de Trl, de CBP et de la DNAse1 afin de calculer la significativité de leurs chevauchements avec les séquences non-codantes des gènes AGZ. Afin d'identifier des facteurs potentiellement impliqués dans le recrutement CBP, ou des co-facteurs de Trl et Zelda, ou plus généralement des facteurs présents dans les régions actives (H3K4me1) ou accessibles (DNAse1) de la chromatine, j'ai procédé à la découverte de motifs surreprésentés dans l'ensemble des pics avec le programme *peak-motifs* de la suite RSAT (cf. annexe 6.2.1). J'ai également analysé les sous-ensembles de pics

chevauchant les régions non codantes des gènes AGZ pour identifier des motifs spécifiquement associés à l'AGZ. J'ai enfin analysé les profils de liaison et d'occupation des différentes protéines, ainsi que l'accessibilité de la chromatine dans les CRM AGZ, afin de vérifier si ces CRM sont spécifiquement associés à un environnement chromatinien particulier.

4.1 Analyses des jeux de pics

J'ai tout d'abord utilisé les pics mis à disposition pour les expériences de ChIP-seq pour CBP (E0-4h) et Trl (E0-8h), d'accessibilité à la DNAse1 (stade 5), et de ChIP-seq pour Zelda (cycles mitotiques 8-9, ce qui correspond grosso modo à 1h après la fécondation ; cycles mitotiques 13-14, 2h ; cycle mitotique 14, 3h). Les auteurs des différentes expériences ont utilisé diverses méthodes afin de détecter les pics de liaison de facteurs (MACS (128)) ou d'accessibilité à la chromatine (méthode décrite dans Thomas et al. (126)). Harisson et al. (2011) ont mis au point leur propre outil de détection de pics *Grizzly Peak* (http ://eisenlab.org/software/grizzly). L'intersection entre les régions non-codantes et les pics a été faite avec les outils fournis dans la suite BedTools (129), qui permet de manipuler des fichiers "BED" contenant la localisation génomique des régions. J'ai utilisé Galaxy (http ://main.g2.bx.psu.edu/) pour récupérer les séquences nucléotidiques des pics sur la base de leurs coordonnées.

4.1.1 Chevauchement entre les pics analysés et les régions non codantes associées aux gènes AGZ

J'ai croisé les données sur les pics avec les annotations génomiques de la drosophile afin d'obtenir un fichier associant à chaque gène la présence de pics dans les séquences non-codantes associées. J'ai également croisé les données sur les pics provenant de différentes expériences afin de rechercher des associations entre marques et facteurs. Pour calculer l'enrichissement des pics dans les régions non codantes de gènes AGZ, j'ai utilisé l'outil *compare-classes* de la suite RSAT. Le tableau 4.1 montre que les séquences non codantes des gènes AGZ sont très significativement enrichies en pics CBP. De manière générale, le recrutement de CBP, la liaison de Trl et l'accessibilité de la chromatine sont largement enrichis dans le cluster AGZ par rapport aux régions non codantes de l'ensemble du génome.

De plus, en visualisant les résultats sur le genome browser de USCS (http ://genome.ucsc.edu) (130), je me suis aperçue que ces pics pouvaient être mal définis. En effet, la Figure 26 montre deux cas où la prédiction des pics n'est pas pertinente au regard des profils de liaisons représentés

TABLE 4.1: Significativité du chevauchement entre pics et séquences non codantes des gènes AGZ. Le nombre total réfère au nombre total de gènes qui possèdent au moins un pic dans une des régions noncodantes.

CBP E0-4h	Trl E0-8h	DNAse stg5	nombre total	intersection	significativité
			1972	234	103,08
			1074	155	72,16
			567	117	69,67
			2234	186	49,82
			313	72	44,42
			1173	124	40,24
			6396	265	19,3

par la densité en reads normalisée. Dans le premier cas (figure 4.1A), le pic recouvre une région très étendue autour du pic réel de liaison, et sa longueur dépasse les 4,5 Kb. Le second cas (figure 4.1B) révèle que MACS, avec les paramètres utilisés par les auteurs, manque de résolution et ne prédit qu'un pic manifestement produit par la fusion de trois pics distincts. J'ai tout de même procédé à la découverte de motifs surreprésentés dans ces pics pour identifier des facteurs potentiellement impliqués dans le recrutement de CBP, de co-facteurs de Trl, ou encore de motifs associés à des régions accessibles et/ou actives de la chromatine. Ces résultats sont présentés dans la section suivante.

Mon but étant de définir un environnement chromatinien spécifique pour les CRM prédits, j'ai changé de stratégie pour étudier l'enrichissement des différentes marques dans les CRM prédits, en m'appuyant sur les données de densité de reads. La méthode et les résultats sont présentés dans la section 4.2.

4.1.2 Analyse des pics avec *peak-motifs*

J'ai procédé à deux analyses complémentaires. Premièrement, j'ai analysé l'ensemble des pics des jeux de données de CBP, Trl, Zelda, d'accessibilité à la DNAse1 et de H3K4me1. Ensuite, j'ai analysé les pics chevauchant les régions non codantes des gènes AGZ (que je nommerai "pics AGZ") de ces mêmes jeux. En effet, je voulais identifier les motifs reliés à ces différentes marques, mais surtout identifier ceux qui sont spécifiques à ces marques dans les séquences non-codantes des gènes AGZ. A cet égard, j'ai utilisé *peak-motifs*, qui combine plusieurs outils de découverte de motifs (*oligo-analysis*, *dyad-analysis*, *position-analysis* et *local-*

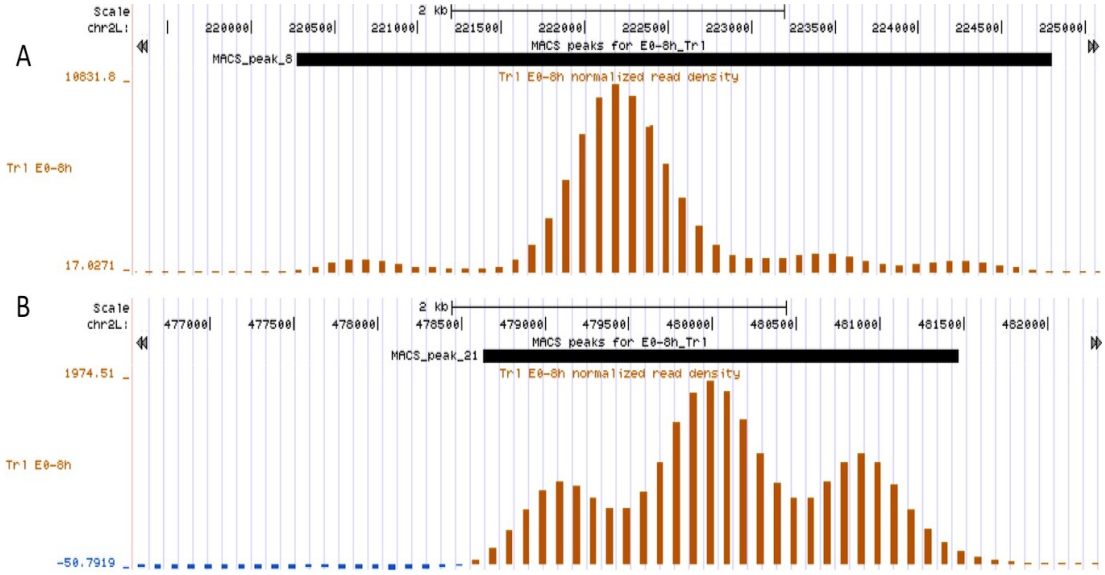


FIGURE 4.1: Visualisation de données provenant d'une expérience de liaison de Trl entre 0 et 8h après la fécondation (ChIP-seq). - A et B montrent les profils en densité de "reads" normalisée (valeurs positives en orange, négatives en bleu) ; la piste "MACS peaks for E0-8h_Trl" contient les pics prédits par MACS.

word-analysis), ainsi qu'un outil de comparaison de motifs (*compare-matrices*). J'ai participé à l'évaluation de ce logiciel en procédant au choix et à l'analyse de cas d'études pour un protocole d'utilisation de l'outil (annexe 6.2.1). Ces analyses m'ont permis d'apprécier l'impact des différents paramètres sur les résultats. J'ai ainsi utilisé des modèles de background appropriés à chaque cas. Pour les jeux de pics entiers contenant plus d'un Mb de séquences, j'ai utilisé le modèle de markov d'ordre $n-2$, où n est la longueur des oligo-nucléotides analysés. Pour les sous-ensembles de pics, en général inférieur à un Mb, j'ai utilisé un modèle de Markov d'ordre $n-3$. J'ai également décidé de ne pas couper les pics malgré la présence de pics très longs. En effet, comme nous l'avons vu plus haut, les pics ne correspondent pas forcément à une seule région de liaison. Si l'on considère l'occupation des histones, l'approche ChIP-seq met en évidence des régions éventuellement occupées par plusieurs nucléosomes marqués, et les pics longs ne sont donc pas forcément aberrants.

Les figures 4.2 et 4.3 montre les motifs découverts dans les pics analysés qui recoupent ceux découverts lors de l'analyse des séquences non codantes des gènes AGZ.

De manière générale, à part pour les pics d'occupation de H3K4me1, le motif de liaison de Zelda (consensus "CAGGTAg") est fortement sur-représenté dans les pics AGZ. C'est bien

entendu particulièrement le cas dans les pics provenant des expériences de liaison de Zelda. Les résultats montrés ont été calculés avec l'outil *position-analysis*, ce qui nous permet de préciser que ces motifs sont centrés au milieu des pics. Le motif résultant est construit à partir de 3530 sites et représente probablement un modèle de liaison plus proche du "vrai" consensus que ceux intégrés dans la base de données FlyFactorSurvey (les deux matrices présentes ont été construites sur la base de 18 sites documentés et de 761 sites SELEX, respectivement). Nous retrouvons également le motif de liaison de Zelda dans le jeu entier de pics de CBP, mais de façon moins significative que pour les gènes AGZ. Dans les autres jeux entiers de pics, nous ne détectons pas ce motif. Le motif de liaison de Trl et le motif "cAGATACa" (potentiellement impliqué dans le phénomène de transvection) apparaissent fortement surreprésentés dans l'ensemble des jeux et sous jeux de pics. Le motif de liaison de Trl est généralement plus significativement sur-représenté dans les jeux entiers de pics que dans les pics AGZ, contrairement au motif "cAGATACa" qui est généralement plus fortement sur-représenté dans les pics AGZ. Le motif "cAGATACa" est cependant plus significatif dans les jeux entiers des pics Zelda 1h et 2h. Ceci suggère que ce motif est plus spécifiquement associé à la liaison de Zelda dans les séquences non codantes AGZ durant la cellularisation (Zelda 3h) que durant les intervalles de temps précédents. Les données de DNAse1 nous confirment que les motifs identifiés se trouve dans les régions ouvertes de la chromatine et donc accessibles aux TF.

4.2 Analyse des densités de "reads"

J'ai utilisé les données de densité de reads représentant les profils de liaison et d'occupation de protéines ou d'accessibilité à la DNAse1. Ces données contiennent des densités de reads discrétisées. Le génome a été fractionné en segments de 20 ou 100 pb, pour les données d'accessibilité à la chromatine ou d'occupation et de liaison à la chromatine, respectivement. J'ai alors calculé une densité moyenne (pour les profils d'occupation d'histones) ou normalisée par le background (pour les profils de liaison de facteurs) pour chaque segment.

J'ai analysé les profils de liaison de Pcl, Trl et Zelda, les profils de recrutement de CBP, les profils d'occupation des histones H3 (tri-méthylation des lysines 4 (K4me3), 27 (K27me3) et 9 (K9me3) ; monométhylation des lysine 4 (K4me1) ; acétylation des lysines 9 (K9ac) et 27 (K27ac)), ainsi que les profils d'accessibilité de la chromatine pour différentes classes temporelles. Au total, j'ai analysé 35 profils pour établir des recoupements avec les CRM prédits. J'ai

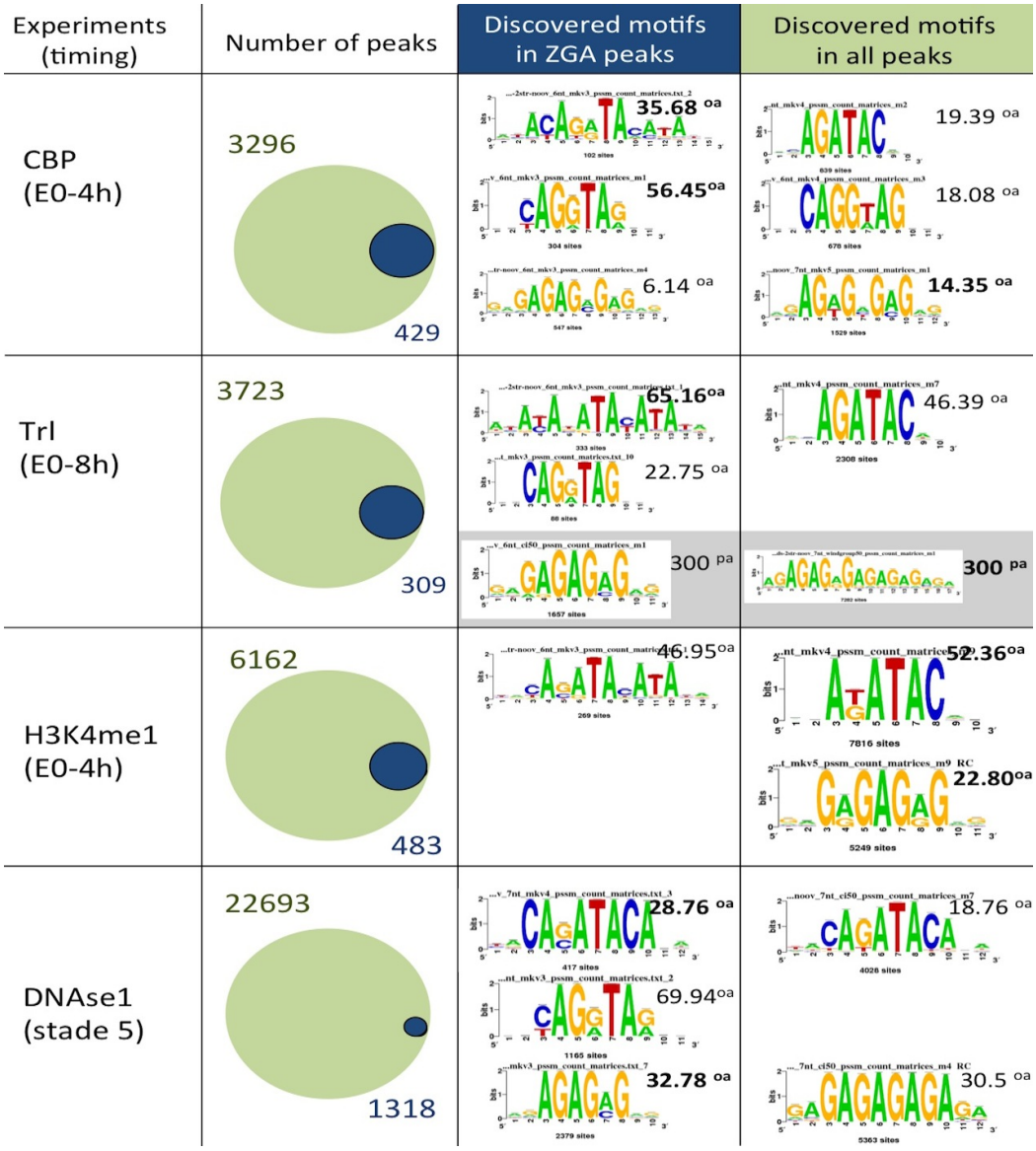


FIGURE 4.2: Découverte de motifs effectuée sur les pics CBP, Trl, H3K4me1 et accessibilité à la DNAse1. - Les cercles verts représentent l'ensemble de pics, les cercles bleus foncés représentent les pics chevauchant les séquences non-codantes des gènes AGZ. La significativité est indiquée à coté de chaque logo, l'exposant indique l'outil utilisé (oa : *oligo-analysis*, pa : *position-analysis*). Les motifs surlignés en gris correspondent aux motifs de liaison du facteur étudié. Lorsque le motif a été découvert dans les deux types de jeux de pics, la significativité la plus importante est mise en gras.

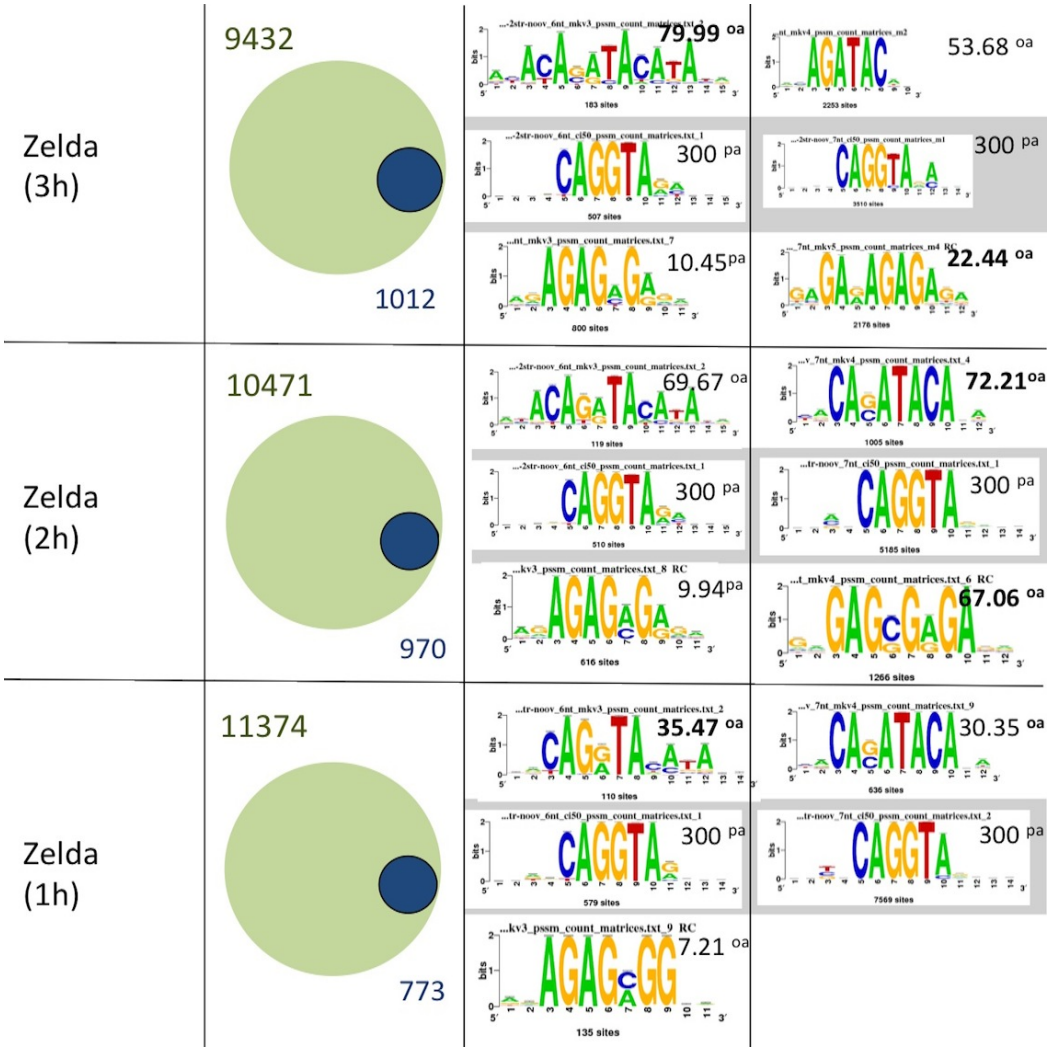


FIGURE 4.3: Découverte de motifs effectuée sur les pics Zelda. - Les cercles verts représentent l'ensemble de pics, les cercles bleus foncés représentent les pics chevauchant les séquences non-codantes des gènes AGZ. La significativité est indiquée à coté de chaque logo, l'exposant indique l'outil utilisé (oa : *oligo-analysis*, pa : *position-analysis*). Les motifs surlignés en gris correspondent aux motifs de liaison du facteur étudié.Lorsque le motif a été découvert dans les deux types de jeux de pics, la significativité la plus importante est mise en gras.

également considéré plusieurs contrôles négatifs et positifs pour évaluer la spécificité de ces recoupements.

4.2.1 Les contrôles positifs et négatifs

Afin de vérifier la spécificité des associations observées pour les CRM prédits avec les matrices résultant de la découverte de motifs sur-représentés dans les séquences non codantes des 417 gènes AGZ (que je nommerai dans cette partie "matrices AGZ"), j'ai utilisé trois contrôles négatifs et un contrôle positif. Le premier contrôle négatif a consisté à prédire des CRM à l'aide des matrices AGZ sur les régions non codantes de 417 gènes choisis au hasard. Le second contrôle négatif a consisté à permuter les matrices AGZ et à utiliser les matrices résultantes pour prédire des CRM sur les séquences non codantes des gènes AGZ. J'ai également utilisé les 317 CRM non actifs dans le blastoderme disponible dans REDFly (131) comme troisième contrôle négatif. Enfin, j'ai utilisé comme contrôle positif les 114 CRM répertoriés dans REDFly comme étant actifs dans le blastoderme.

4.2.2 Méthodologie

4.2.2.1 Calcul de densité intégrée

Afin de calculer la spécificité de liaison sous une région donnée, j'ai calculé l'interpolation linéaire entre les densités moyennes ou normalisées de reads assignées aux segments (ou "bins") présents sous et directement autour de la région. La figure 4.4A présente schématiquement les annotations utilisées dans les formules suivantes. La densité intégrée S_u entre les bins 1 à n présents sous la région r a été calculée comme suit :

$$S_u = \sum_{i=1}^{n-1} \Big(\frac{h_i + h_{i+1}}{2} \Big) (x_{i+1} - x_i), \tag{4.1}$$

où h_i est la valeur de la densité assignée au bin i et x_i est la position centrale du bin i. J'ai ensuite estimé les densités h_s et h_e à la position de début x_s et de fin x_e de la région, respectivement :

$$h_s = h_0 + \frac{x_s - x_0}{(x_1 - x_0)(h_1 - h_0)}, \tag{4.2}$$

$$h_e = h_{n+1} + \frac{x_e - x_n}{(x_{n+1} - x_n)(h_{n+1} - h_n)}, \tag{4.3}$$

où x_0 et x_{n+1} sont les positions centrales des bins entourant la région en amont et en aval, respectivement, et h_0 et h_{n+1} sont les densités qui leur sont associées. J'ai ensuite calculé la densité intégrée S_s entre le début de la région (x_s) et le centre du premier bin (x_1) sous la région :

$$S_s = \Big(\frac{h_1 + h_s}{2}\Big)(x_1 - x_s), \tag{4.4}$$

ainsi que la densité intégrée S_e entre le centre du bin le plus à droite (x_n) sous la région et la fin de la région (x_e) :

$$S_e = \Big(\frac{h_n + h_e}{2}\Big)(x_e - x_n). \tag{4.5}$$

La densité intégrée S_r sous la région entière est donc :

$$S_r = S_s + S_u + S_e. \tag{4.6}$$

Enfin, j'ai normalisé S_r afin d'obtenir des valeurs comparables entre les régions :

$$S_{norm} = \frac{S_r}{L_r}, \tag{4.7}$$

où L_r est la longueur de la région.

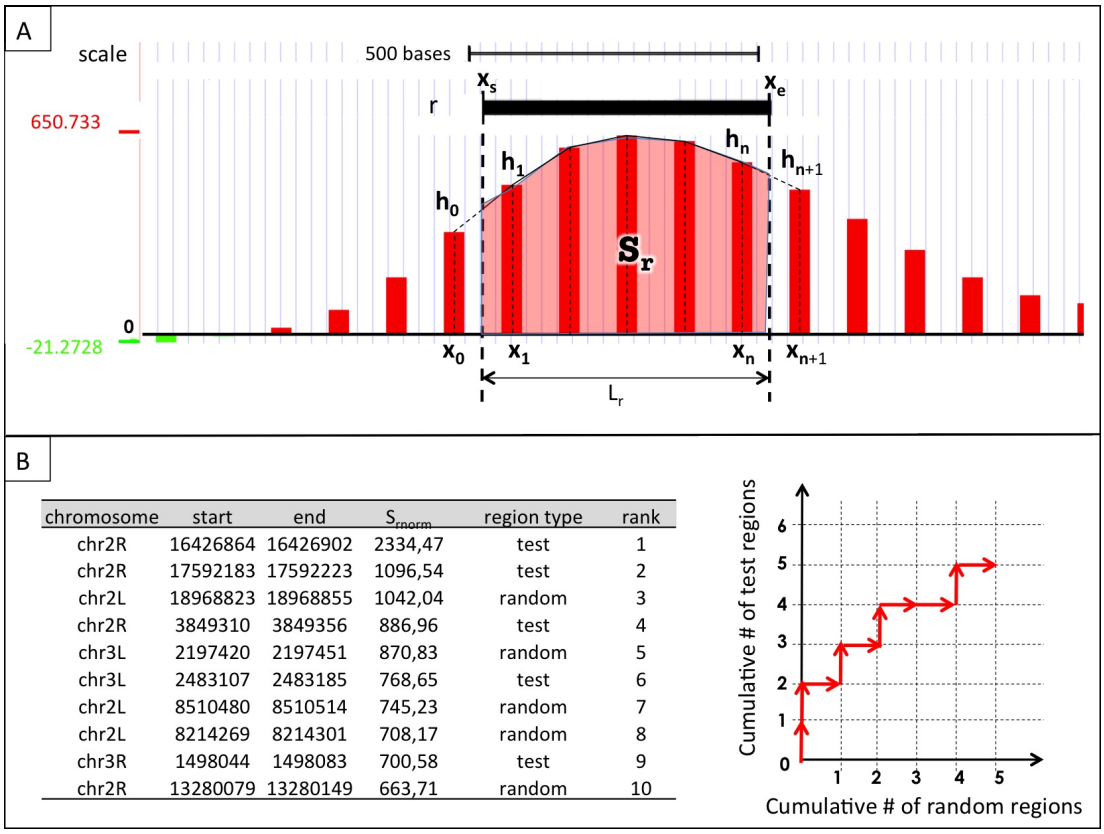


FIGURE 4.4: Calcul de la densité en reads intégrée sous une région données et representation des résultats sour la forme d'une courbe de ROC. - A. L'histogramme représente les densités en reads discrétisées en segment de 75 pb. La hauteur de chaque barre correspond à la densité en reads normalisée dans le segment. Les lignes pointillées verticales indiquent le centre de chaque segment, tandis que la surface rouge correspond à la densité en reads intégrée S_r obtenue par interpolation des densités normalisées (cf. texte). Les annotations x, h et L correspondent aux symboles utilisés pour le calcul de S_r. B. À gauche : classement des régions aléatoires et régions testées (ex. CRM) en fonction de la densité intégrée obtenue pour une expérience de localisation génomique donnée. À droite : l'axe des abscisses et l'axe des ordonnées correspondent au nombre de régions aléatoires et au nombre de régions tests rencontrées dans le classement, respectivement. La courbe est tracée en suivant l'un des deux axes selon le type de régions rencontrées dans le classement.

4.2.2.2 Construction des courbes ROC

J'ai généré dix fois plus de régions choisies au hasard que de régions à tester (CRM contrôles ou AGZ) en utilisant un vecteur regroupant les longueurs des régions à tester. J'ai ensuite calculé les intensités S_r sous toutes les régions et procédé au classement des régions en fonction de l'intensité (figure 4.4B).

J'ai ensuite normalisé les axes afin de pouvoir analyser la spécificité et la sensibilité du critère choisi (par exemple la densité en reads intégrée obtenue pour un facteur donné) pour la discrimination des CRM par rapport aux régions aléatoires. La figure 4.5 donne le résultat obtenu en classant les CRM prédits dans les régions non-codantes des gènes AGZ et les régions choisies au hasard en fonction de la densité intégrée S_r calculée à partir des densités en reads provenant de l'expérience de liaison du facteur Zelda. L'abscisse indique la fraction de régions choisies au hasard présentes dans le classement et représente donc le taux de faux positifs (FP) attendu si l'on cherche à détecter les CRM en utilisant le profil de liaison de Zelda. L'ordonnée représente la fraction de CRM présents dans le classement. La ligne diagonale représente donc une répartition homogène des CRM et des régions aléatoires dans le classement. Dans le cas présenté ici, nous pouvons voir qu'un seuil minimal de liaison de Zelda (S_r) fixé à 9.23 (les valeurs d'intensité allant de 4344.79 à 0) permet de détecter 80% des CRM et 20% des FP. Ceci indique que le facteur Zelda est recruté spécifiquement dans les CRM. De manière générale, plus une courbe s'éloigne de la diagonale en suivant l'axe des ordonnées, plus le facteur (ou la marque chromatinienne) est spécifiquement associé(e) aux CRM. Inversement, si la courbe suit la diagonale ou l'axe des abscisses, cela indique qu'il n'y a pas d'enrichissement particulier ou que la liaison du facteur est évitée dans les régions définies par les CRM, respectivement. Afin de mesurer plus précisément la spécificité des facteurs testés pour les CRM, j'ai calculé l'aire sous la courbe ($AUC_{0.2}$) pour un seuil de détection de 20% de FP.

L'$AUC_{0.2}$ a été calculée comme suit :

Si l'on considère la courbe découlant directement du classement des CRM et régions aléatoires (figure 4.4B), la valeur maximale de l'axe des abscisses, que nous nommons n, correspond au nombre total de régions aléatoires et la valeur maximale de l'axe des ordonnées, que nous nommons m, correspond au nombre total de CRM.

Au rang k, i_k régions aléatoires et $j_k = f(i_k)$ CRM seront présents dans le classement, avec i_k appartenant à $[0, n]$ et j_k appartenant à $[0, m]$. On définit x_k et y_k la fraction du nombre

total de régions aléatoires et la fraction du nombre total de CRM, présentes entre le rang 1 et le rank k, respectivement. Ainsi $x_k = \frac{i_k}{n}$ et $y_k = \frac{j_k}{m}$. x_k et y_k sont donc compris entre 0 et 1. L'aire totale AUC_{tot} sous la courbe tracée en fonction de $f(x_k)$ est donc :

$$AUC_{tot} = \sum_{x_k=0}^{1} (f(x_k)) \tag{4.8}$$

Étant donnée que la courbe $f(x_k) = x_k$ correspond à un enrichissement nul, j'ai retiré l'aire sous cette courbe à l'aire totale AUC_{tot} afin d'apprécier directement l'enrichissement du facteur étudié dans les CRM.

$$AUC = \sum_{x_k=0}^{1} (f(x_k) - x_k) \tag{4.9}$$

Ainsi les AUC positives, nulles ou négatives indiquent un enrichissement, aucun enrichissement ou un évitement des facteurs testés, respectivement. Le calcul de l'enrichissement en CRM dans le haut du classement est plus informatif. Nous calculons alors l'AUC jusqu'à que 20% des régions aléatoires aient été classées ($x_k = 0.2$). Ainsi, $AUC_{0.2}$ est donnée par :

$$AUC_{0.2} = \sum_{x_k=0}^{0.2} f(x_k) - x_k) \tag{4.10}$$

4.2.2.3 Analyse de combinaisons de marques

Afin d'identifier des combinaisons de marques spécifiquement associée aux CRM, j'ai procédé au classement des régions en comparant les rangs de chaque région r par rapport à la combinaison d'un ensemble ω d'expériences de localisation, définis comme suit :

$$\tilde{S}_r = \tilde{m}_{k_\omega}, \tag{4.11}$$

où $\tilde{m}_{k_\omega}$ est la médiane des rangs k obtenus dans le set d'expériences ω.

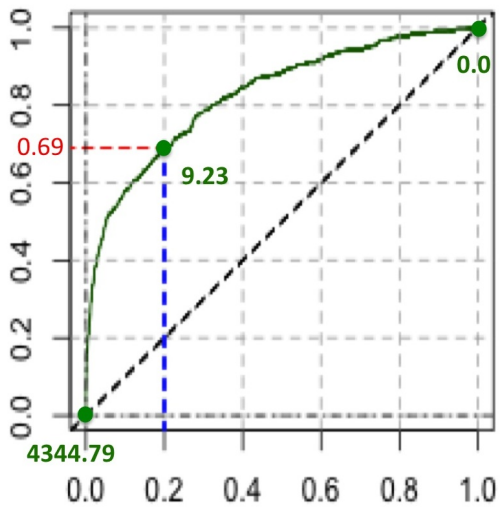


FIGURE 4.5: Courbe de ROC correspondant au classement des CRM AGZ (prédits dans les région en amont du TSS) parmi des régions non codantes choisies au hasard, en fonction des densités intégrées calculées sur la base des donnée de ChIP-seq du facteur Zelda (3h). - L'abscisse et l'ordonnée représentent respectivement la fraction de régions choisies au hasard et de CRM classés en fonction de S_r décroissante. Les nombres en vert indiquent les valeurs d'intensité correspondant au premier rang (i.e. valeur maximale de S_r, coordonnées (0,0)), au dernier rang (i.e valeur minimale de S_r, coordonnées (1,1)) et au rang auquel 20% des régions aléatoires (ligne pointillée bleue) et 69% des CRM (ligne pointillée rouge) sont détectés.

4.2.3 Résultats

J'ai tout d'abord procédé à l'analyse séparée de chaque jeux de données, afin de confronter les résultats obtenus entre les CRM prédits dans les régions non codantes des gènes AGZ (que je nommerais par la suite CRM AGZ) avec ceux obtenus avec les divers contrôles, et ainsi détecter les associations spécifiques de certaines marques avec les CRM AGZ. J'ai ensuite analysé le comportement de la liaison, de l'occupation et de l'accessibilité de la chromatine au cours du temps dans les CRM AGZ et les CRM présents dans RedFly, afin d'évaluer la spécificité des enrichissements découverts pendant la période couvrant l'AGZ. Enfin, j'ai calculé l'enrichissement pour des combinaisons de marques, afin de détecter des associations particulières. Je me limiterai ici à la présentation des résultats obtenus pour les CRM prédits dans les régions en amont des TSS. En effet, les CRM prédits dans les 5'UTR et dans les premiers introns montrent de faibles enrichissements. Par ailleurs, l'analyse de tous les CRM AGZ montre un enrichissement plus faible que les CRM prédits dans les régions en amont des TSS.

Le contrôle négatif consistant à scanner cinq groupes de gènes choisis au hasard avec les matrices AGZ ne retournent que très peu de CRER et ne révèle aucun enrichissement particulier (figure 4.6) ; les résultats ne sont montrés que pour un seul groupe (ceux obtenus pour les quatre autres groupes étant très similaires). Ceci suggère que les matrices AGZ sélectionnées sont spécifiques du groupe de gènes AGZ et prédisent des CRER dans des régions pourvues de signatures chromatiniennes particulières.

Par ailleurs, la permutation des matrices permet de générer des motifs aléatoires ayant le même contenu informationnel que les matrices de départ, ce qui constitue un second type de contrôle négatif. Les CRER prédits en scannant les régions non codantes des gènes AGZ avec les matrices AGZ permutées correspondent à des régions dotées de signatures chromatiniennes similaires à celles des CRM AGZ (figure 4.6). Environ 380 CRER ont été prédits avec les matrices permutées, dont 260 chevauchent au moins un CRM AGZ. Ce résultat s'explique au moins en partie par la sélection de plusieurs matrices AGZ riches en GA (Trl-like, motif2, motif5 et motif6). En effet, les permutations aléatoires de ces matrices produisent des motifs généralement peu différents.

4.2.3.1 La chromatine dans tous ses états

Marques liées à l'activité transcriptionnelle

Plusieurs expériences permettent d'analyser l'activité transcriptionnelle au niveau de la chro-

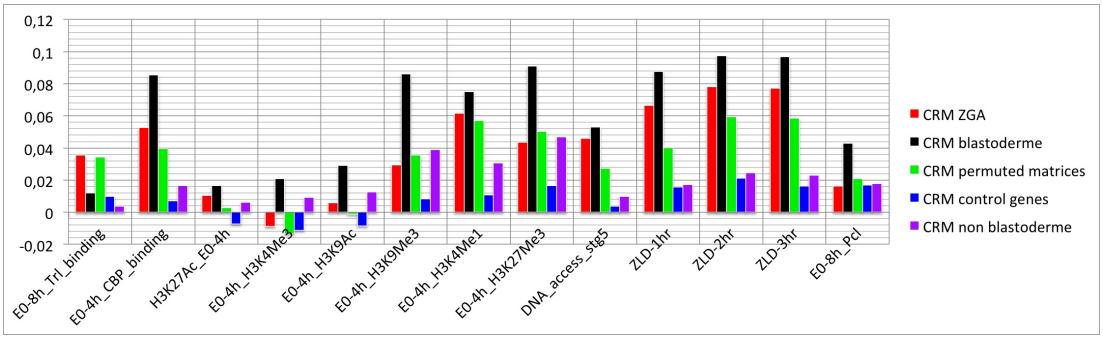


FIGURE 4.6: Distribution des valeurs d'AUC calculées en fonction des intensités de liaison, d'occupation ou d'accessibilité de la chromatine pour les CRM AGZ et contrôles. - L'abscisse indique les noms des facteurs (CBP, Trl et Pcl), les modifications d'histone H3, l'accessibilité de la chromatine (DNA_access), ainsi que les classes temporelles des expériences. L'ordonnée indique la valeur de l'AUC. ZLD : Zelda.

matine. Tout d'abord l'accessibilité de la chromatine peut être évaluée grâce aux expériences de DNAse1. Les CRM AGZ et blastoderme sont associés spécifiquement aux régions d'ouverture de la chromatine au stade 5 (figure 4.7A), par ailleurs, un faible enrichissement est obtenu dans les CRM contrôles négatifs. L'association des CRM AGZ et des CRM blastoderme avec cette marque diminue au cours du temps (figure 4.7 B-C).

L'activité transcriptionnelle peut également être révélée par la liaison de l'ARN polymérase II (polII). Cependant, je n'ai pas trouvé de données de ChIP-seq pour la polII dans un intervalle de temps englobant l'AGZ, les données les plus précoces disponibles correspondant à 4-8h après la fécondation (figure 4.7D). Durant cette période aucun enrichissement n'est détecté. L'analyse des profils d'occupation disponibles pour la polII (figure 4.7E-F), indique que les CRM AGZ et les CRM blastoderme présentent les mêmes profils d'enrichissement, particulièrement marqué pour l'intervalle temporel 8-12h.

Les régions activatrices de la transcription sont associées à différentes marques (124). En effet, Nègre et collaborateurs ont mis en évidence l'enrichissement en CBP et H3K4me1 dans les CRM blastoderme. Les résultats présentés dans la figure 4.8 sont en accord avec cette étude. L'enrichissement observé pour ces deux marques dans les CRM AGZ indique que ceux-ci correspondent à dans des régions activatrices de la transcription. L'analyse des marques CBP au cours du temps montre que les CRM AGZ et blastoderme ne restent pas associés à CBP. Ceci suggère que les CRM AGZ sont spécifiquement actifs dans les embryons âgés de 0 à 4h.

Nègre et collaborateurs ont également identifié plusieurs marques d'histones associées

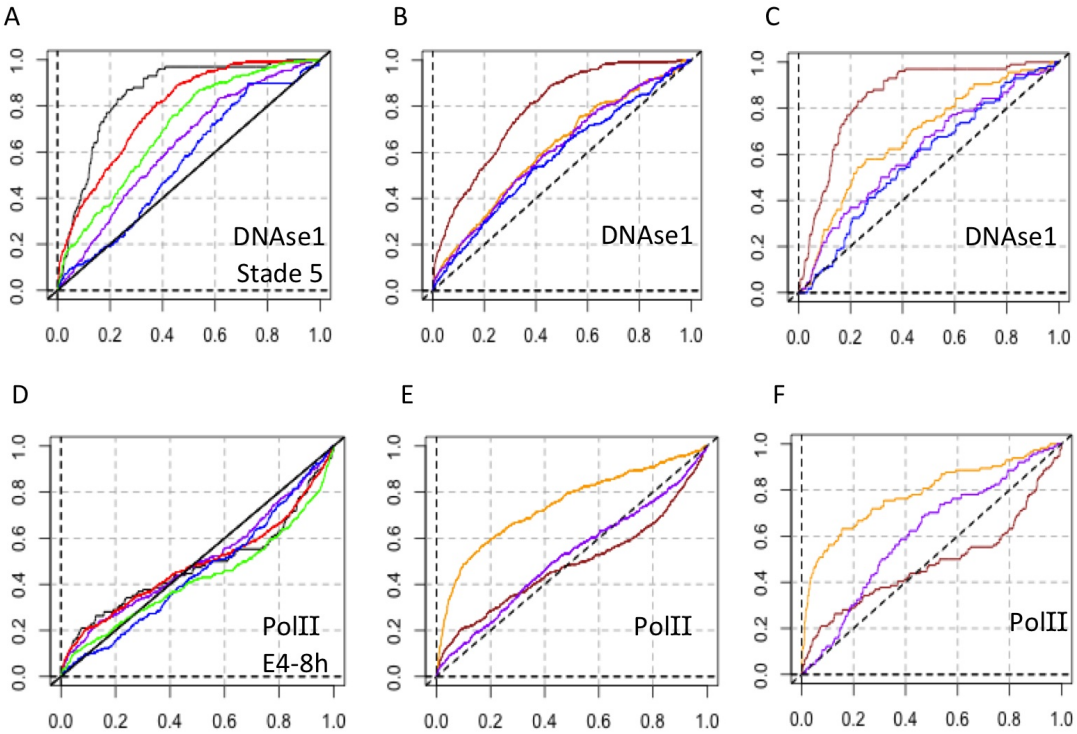


FIGURE 4.7: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques d'ouverture de la chromatine (DNAse1) et de liaison de l'ARN Polymérase II (polII). - A et D. La protéine analysée et l'intervalle temporel de chaque expérience sont indiqués dans le coin bas à droite de chaque graphe. La courbe rouge correspond au classement des CRM AGZ ; les courbes noires et violettes correspondent au classement des CRM blastoderme et non blastoderme, respectivement ; les courbes bleues et vertes correspondent aux classements des CRM contrôles obtenus avec un jeu de gènes aléatoires et avec les matrices permutées, respectivement. B-C, E-F. Résultats de l'analyse des données d'ouverture de la chromatine (B-C) et de l'occupation de la polII (E-F) au cours du temps, pour les CRM AGZ (B, E) et les CRM blastoderme (C, F). Les stades développementaux des expériences de DNAse1 sont représentés en rouge (stade 5), orange (stade 9), violet (stade 11) et bleu (stade 14). Les intervalles temporels des expériences d'occupation de la PolII sont représentés en rouge (4-8h), orange (8-12h) et violet (12-16h).

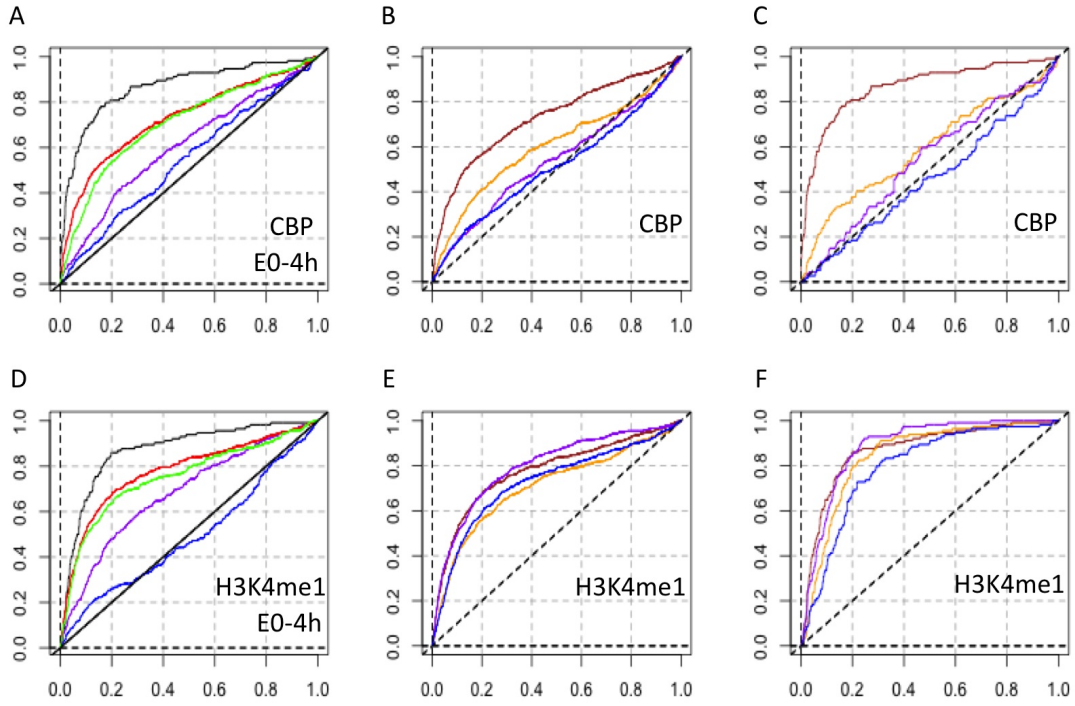


FIGURE 4.8: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques associées aux régions activatrices. - A et D. La protéine analysée et l'intervalle temporel de chaque expérience sont indiqués dans le coin bas à droite de chaque graphe. Le code couleur des courbes est identique à celui des figures 4.7A et D. B-C, E-F. Résultats d'analyses de CBP (B-C) et de H3K4me1 (E-F) au cours du temps pour les CRM AGZ (B, E) et les CRM blastoderme (C, F). La couleur des courbes indique les intervalles temporels des expériences. Dans l'ordre chronologique par intervalles de 4h de 0 à 16h : rouge, orange, violet et bleu.

aux promoteurs actifs (H3K4me3, H3K9ac et H3K27ac). Un enrichissement apparait dans les courbes ROC construites à partir des trois marques dans les CRM blastoderme (figure 4.9 A-C). Les valeurs des AUC (figure 4.6) suggèrent cependant une très faible spécificité. Les CRM AGZ ne montrent pas d'enrichissement particulier. En effet, la marque H3K4me3 semble même évitée (figure 4.9A). En ce qui concerne les autres marques, les courbes sont similaires aux courbes obtenues avec les CRM non blastoderme. Le contrôle construit à partir d'un groupe de gènes choisis au hasard montre clairement un évitement de ces marques. Ceci suggère que les gènes AGZ sont actifs, mais que les CRM AGZ, à l'inverse des CRM blastoderme, ne correspondent pas à des promoteurs actifs. Ceci suggère que ces CRM correspondraient plutôt à des régions enhancers.

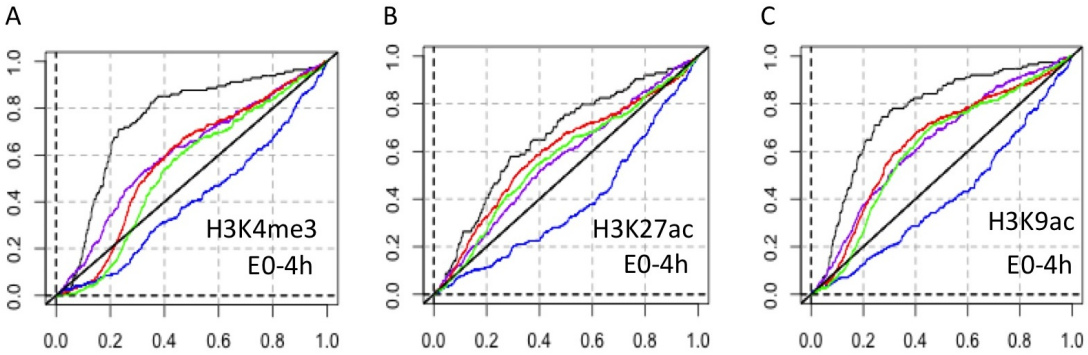


FIGURE 4.9: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques associées au promoteurs actifs. - A-C. La protéine analysée et l'intervalle temporel de chaque expérience sont indiqués dans le coin bas à droite de chaque graphe. Le code couleur des courbes est identique à celui des figures 4.7A et D.

L'analyse des régions non codantes des gènes AGZ a mis en évidence les motifs de liaison de Trl. L'enrichissement n'est pas aussi important que celui observé pour d'autres marques (CBP ou H3K4me1), mais l'enrichissement est clairement spécifique aux CRM AGZ (figure 4.10A), ce qui confirme la pertinence du motif de liaison de Trl découvert plus haut. Les intervalles temporels utilisés pour l'étude de la location génomique de Trl sont relativement longs (8h). Cette faible résolution temporelle pourrait diluer le signal, puisque les embryons sont récoltés sur l'ensemble de la période. De plus, notre analyse révèle une diminution de l'enrichissement de la liaison de Trl (figure 4.9B) dans les régions définies par les CRM AGZ au cours du temps. Sur la base des données disponibles, nous ne pouvons pas déterminer à partir de quel moment Trl pourrait cesser de se lier aux CRM AGZ.

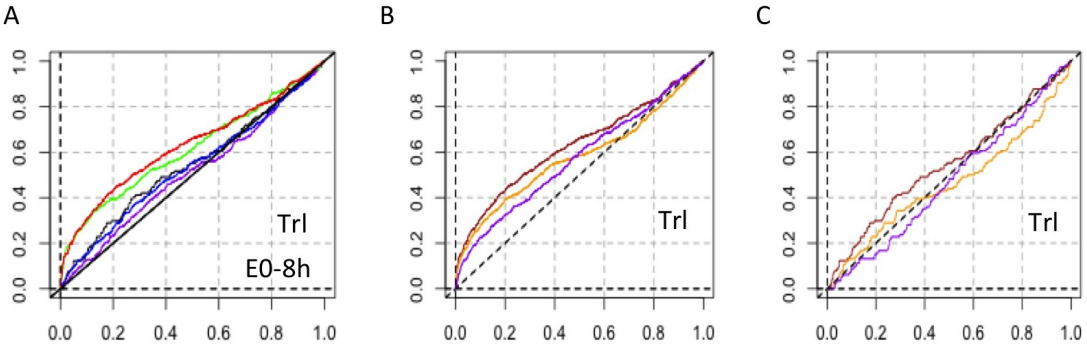


FIGURE 4.10: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles de la liaison de Trl. - Le code couleur des courbes est identique à celui des figures 4.7A et D. B-C. Résultats d'analyses de Trl au cours du temps pour les CRM AGZ (B) et les CRM blastoderme (C). Les intervalles temporels des expériences de liaison de Trl sont représentés en rouge (0-8h), orange (8-16h) et violet (16-24h).

Marques liées à l'état répressif de la chromatine

Les CRM blastoderme et les CRM AGZ diffèrent en ce qui concerne les modifications d'histones répressives (H3K27me3, H3K9me3) et la liaison du facteur Pcl associé à ces marques (figure 4.11A-C). Les CRM blastoderme contrôlent des gènes impliqués dans le développement qui sont activés très localement. L'accessibilité à ces séquences régulatrices est donc fortement réprimée dans la plupart des noyaux de l'embryon. Les marques H3K27m3 et H3K9m3 restent associées au CRM blastoderme au moins jusqu'à 16h et 12h après la fécondation, respectivement (figure 4.11D-E). J'ai testé la combinaison des deux marques répressives avec Pcl (figure 4.11F). Il ne semble pas y avoir d'association particulière entre la présence de Pcl et les tri-méthylations de H3K27 ou K9 (le résultat de H3K9me3 n'est pas montré mais est similaire à celui observé pour H3K27me3). En effet, si les CRM enrichis en H3K27me3 étaient les mêmes que ceux enrichis en Pcl, la courbe représentant les rangs médians (courbe noire) aurait montré une meilleure spécificité que H3K27me3 seul.

Combinaison de marques associées spécifiquement au CRM AGZ

Comme indiqué plus haut, les densité intégrées de reads provenant des expériences de liaison de Trl, du recrutement de CBP, d'occupation d'H3K4me1, et d'accessibilité de chromatine permettent de discriminer clairement les CRM AGZ des régions non codantes choisies au hasard. Pour affiner cette constatation, j'ai analysé séquentiellement la combinaison de ces différentes

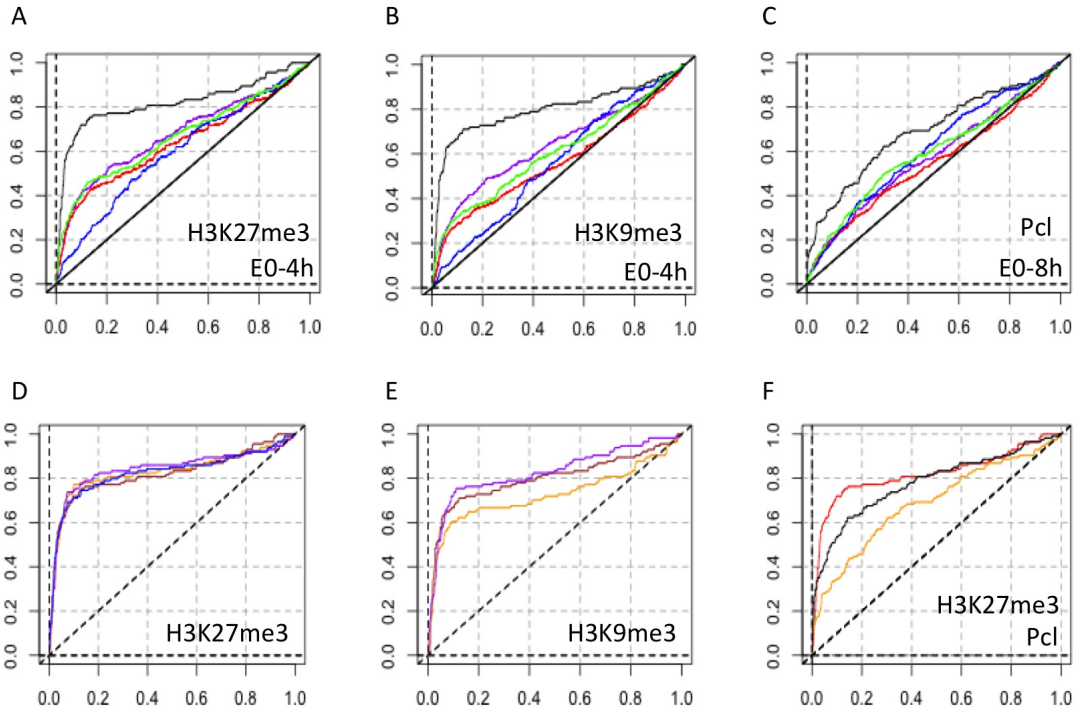


FIGURE 4.11: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques répressives. - A-C. La protéine analysée et l'intervalle temporel de chaque expérience sont indiqués dans le coin bas à droite de chaque graphe. Le code couleur des courbes est identique à celui des figures 4.7A et D. D-E. Classement des CRM blastoderme en fonction des intensités d'occupation de H3K27m3 et H3K9m3. La couleur des courbes indique les intervalles temporels des expériences. Dans l'ordre chronologique par intervalles de 4h de 0 à 16h : rouge, orange, violet et bleu. F. Analyse de la combinaison (courbe noire) des expériences de d'occupation de H3K27m3 (courbe rouge) et de liaison de Pcl (courbe orange).

marques. Une combinaison est considérée comme spécifique dans la mesure où elle améliore le classement de ces CRM par rapport au meilleur classement individuel (cf. 4.2.2.3).

Deux combinaisons améliorent le classement des CRM AGZ (figure 4.12). La première associe H3K4me1 et CBP, alors que la seconde associe H3K4me1, CBP, Trl et l'ouverture de la chromatine. Ceci suggère qu'il pourrait y avoir une collaboration entre CBP et Trl spécifiquement dans les CRM AGZ présents dans des régions accessibles et actives de la chromatine. Ces observations et les résultats présentés précédemment m'ont permis de proposer un modèle de régulation de l'AGZ. Ce modèle fera l'objet d'une discussion dans le chapitre suivant.

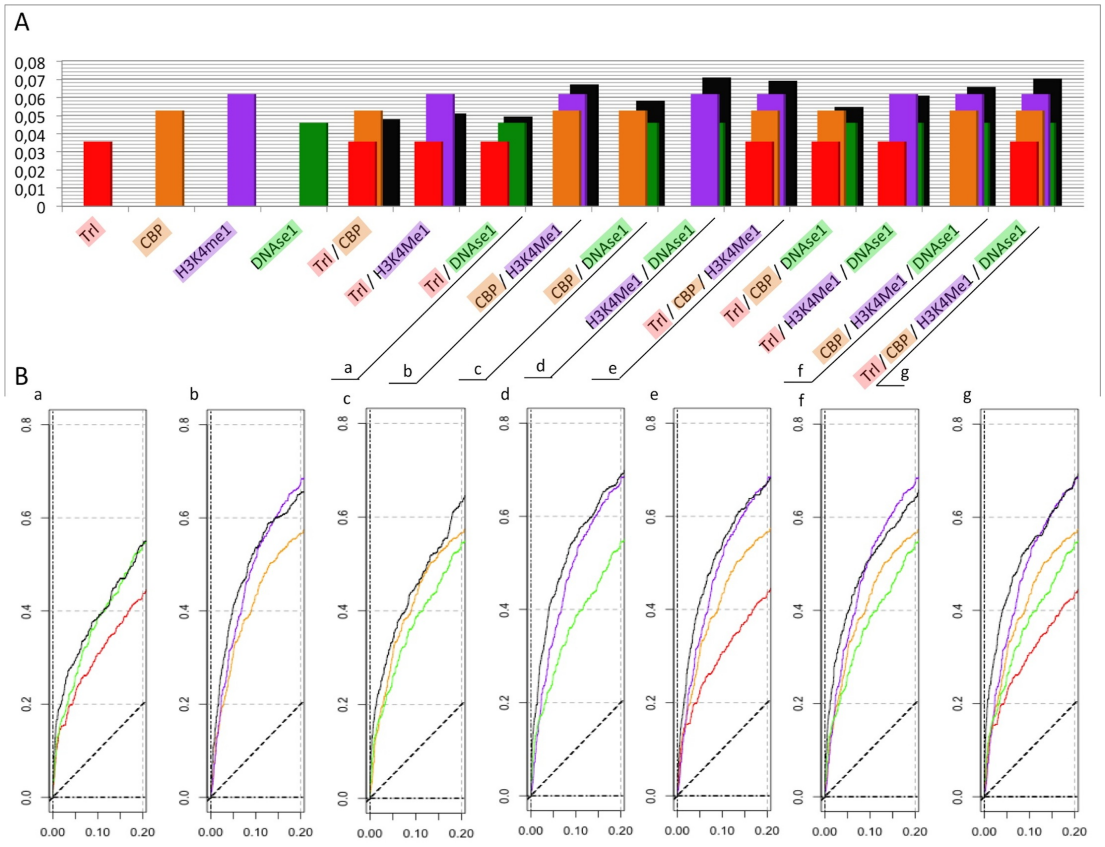


FIGURE 4.12: Résultats de l'analyse de la combinaison de différentes marques. - A. Distribution des AUC. L'abscisse indique les combinaisons analysées. B. Agrandissement de courbes de ROC pour l'analyse du classement contenant 20% des régions aléatoires. Dans les deux parties de la figure, les couleurs rouge, orange, violet, vert et bleu correspondent respectivement à Trl E0-8h, CBP E0-4h, H3K4me1 E0-4h et DNAse 1 stade 5. Les lettres a, b, c, d, e et f dans la partie du bas (B) correspondent aux combinaisons soulignées dans la partie du haut (A).

4.2.3.2 Localisation de Zelda

Les résultats obtenus à partir des profils de liaison de Zelda (figure 4.13B) confirme sa liaison spécifique dans les CRM blastoderme. La présence d'un motif ressemblant au motif de liaison de Zelda permet de prédire des sites effectivement reconnus par Zelda. Les densités de reads intégrées calculées à partir des données d'expérience de localisation de Zelda constituent le meilleur critère de discrimination des CRM AGZ (figure 4.13A). La combinaison de Zelda et H3K4me1 améliore encore légèrement le pouvoir de discrimination (figure 4.13D).

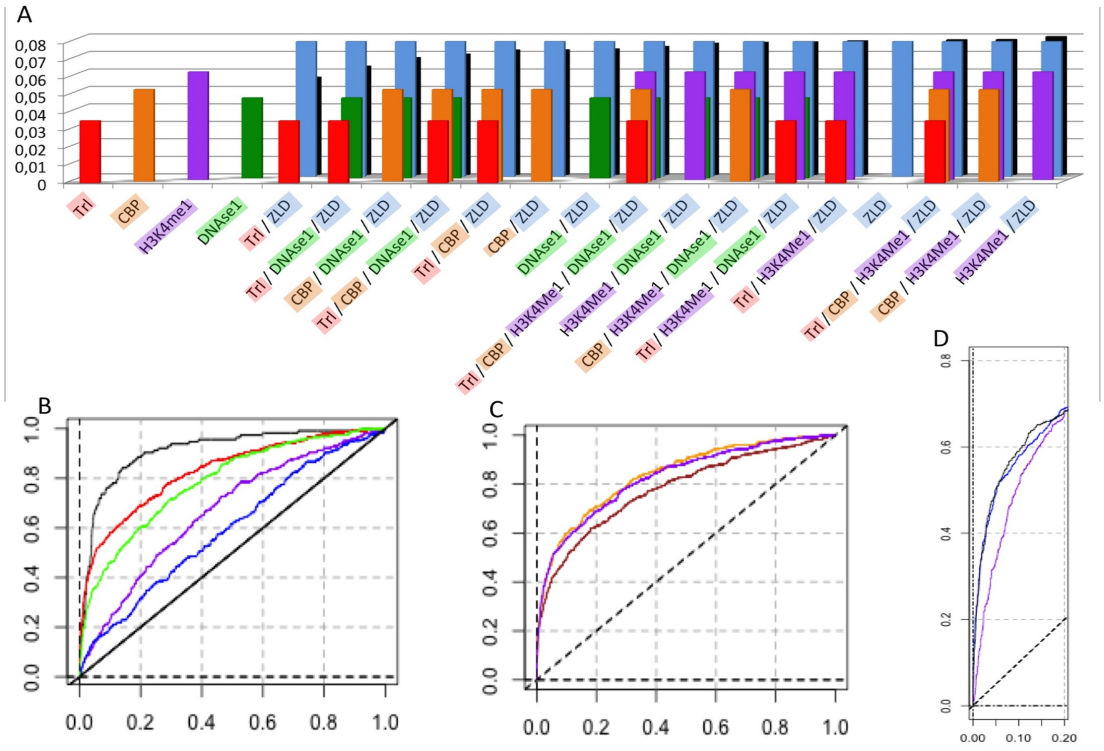


FIGURE 4.13: Analyse de l'enrichissement des CRM AGZ et contrôles de la liaison de Zelda. - A. Distribution des AUC. L'abscisse indique les combinaisons analysées. B. Le code couleur des courbes est identique à celui des figures 4.7A et D. C. Résultats de l'analyse des données de liaison de Zelda au cours du temps pour les CRM AGZ. Les intervalles temporels des expériences de liaison de Zelda sont représentés en rouge (1h), orange (2h) et violet (3h). D. Courbe ROC correspondant à la combinaison des classements obtenus à partir de densités intégrées de Zelda et H3K4me1.

4.3 Conclusion du chapitre

L'analyse des données venant d'expériences de localisation à grande échelle nous a permis de mettre en évidence que les CRM AGZ sont majoritairement localisés dans des régions accessibles et potentiellement actives de la chromatine. En effet, les CRM AGZ sont particulièrement associés avec les marques spécifiques d'enhancers actifs (CBP, H3K4me1). Trl s'est révélé également enrichi dans les CRM AGZ, bien qu'avec une spécificité relativement faible. Les embryons utilisés pour l'expérience de ChIP-seq de ce facteur étaient âgés de 0 à 8h. Ceci suggère une hétérogénéité des embryons au regard de leur stade développemental (figure 4.14). Malgré cette faible spécificité d'association entre Trl et les CRM AGZ, la combinaison de Trl avec CBP, H3K4me1 et les régions ouvertes de la chromatine montre un très fort enrichissement. La présence prédominante de Zelda entre 1h et 3h conforte la pertinence des CRM prédits. Les recoupements établis suggèrent que les différents motifs surreprésentés dans les CRM AGZ prédits sont spécifiquement associés aux régions activatrices accessibles de la chromatine durant l'AGZ.

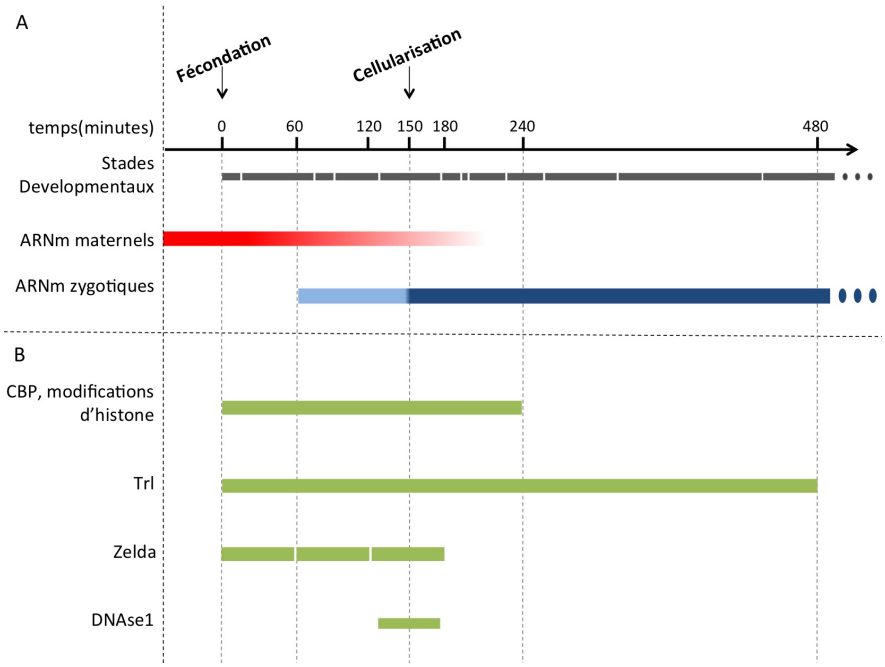


FIGURE 4.14: Organisation temporelle des données à haut débit analysées. - A. L'axe horizontal représente le temps en minutes après la fécondation. Les stades développementaux sont indiqués en gris sous l'axe temporel. Le gradient rouge représente la variation d'abondance des ARNm maternels au cours du temps. La ligne bleue correspond aux ARNm zygotiques synthétisés durant la première (bleu clair) et seconde (bleu foncé) vagues d'activation du génome zygotique. B. Intervalles correspondant aux jeux de données analysées dans ce chapitre (pour plus de détails, voir texte)

5

Conclusion

5.1 Proposition d'un mécanisme de régulation de l'AGZ

A partir de l'analyse et de l'intégration d'une série de données de transcriptome, j'ai sélectionné un groupe de 417 gènes dont l'expression est induite durant l'AGZ. L'analyse bioinformatique des régions non codantes associées à ces gènes (promoteurs, introns, UTR) m'a permis de détecter des motifs particulièrement surreprésentés, dont certains correspondent aux motifs de liaison de facteurs connus (Zelda, Trl et TTK). Ces motifs ont ensuite été utilisés pour prédire des éléments cis-régulateurs potentiels et des CRM (régions significativement enrichies en sites prédits), ce qui a permis d'identifier près d'un millier de régions potentiellement régulatrices. Enfin, différentes caractéristiques spécifiques de ces CRM ont pu être révélées par l'analyse de différents jeux de ChIP-seq concernant la localisation génomique de plusieurs marques chromatiniennes et facteurs associés. Ces résultats m'ont amenée à proposer un modèle de régulation de l'AGZ intégrant des mécanismes de remodelage de la chromatine et l'action de FT connus ou encore à caractériser. Chaque élément de ce modèle (figure 5.1) est discuté ci-dessous.

5.1.1 Avant l'AGZ

Durant la première heure de développement, la transcription du génome zygotique est silencieuse. Nous avons vu dans l'introduction que ceci peut être dû à différents facteurs : cycles mitotiques rapides, répression par des facteurs maternel, compaction de la chromatine. Pagans et collaborateurs ont montré que TTK peut inhiber l'action de Trl via une interaction protéine-protéine directe (figure 5.1Ab) dans le cadre de la régulation de l'expression du gène eve (102).

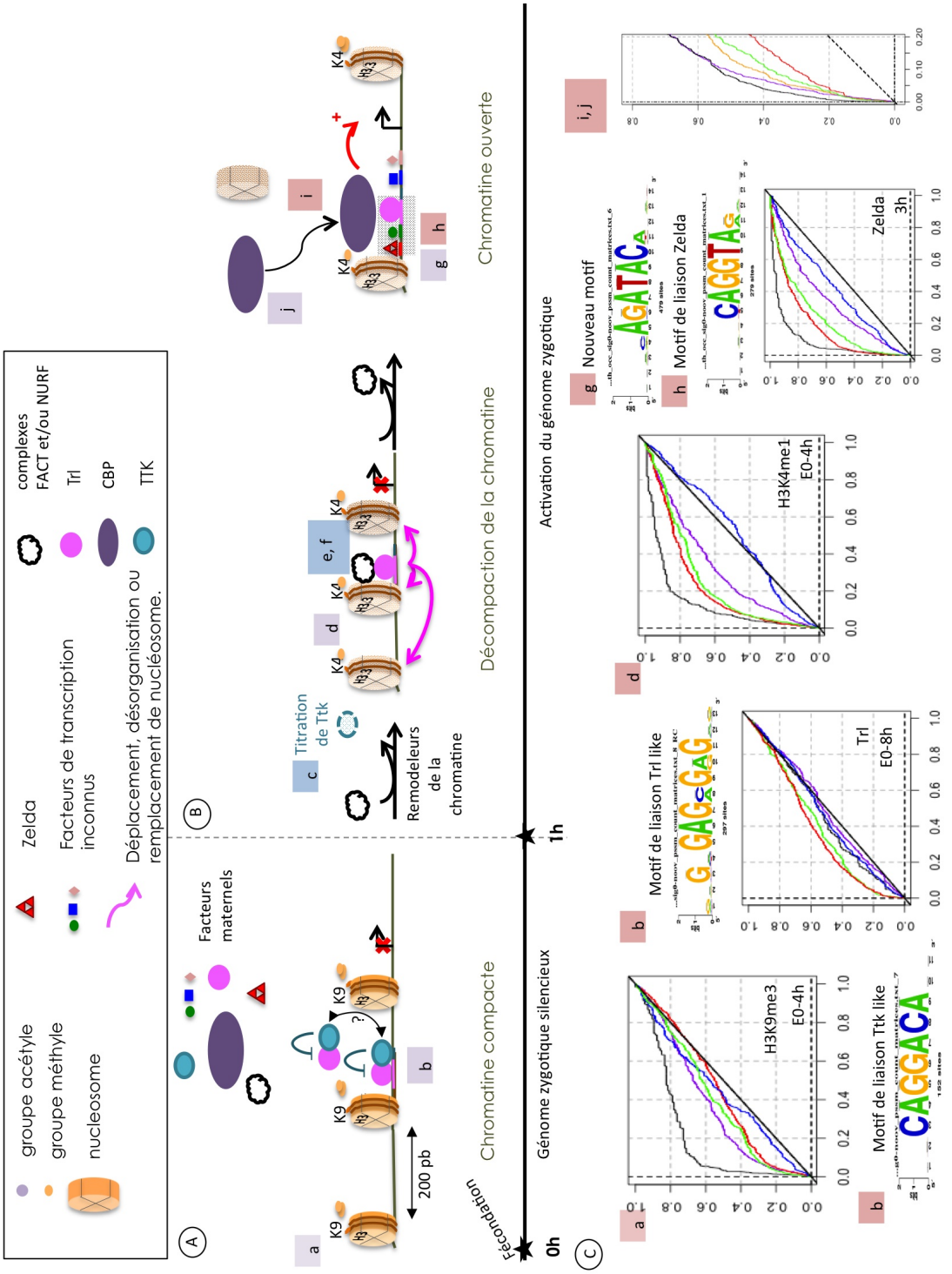


FIGURE 5.1: Modèle proposé pour la régulation de l'activation transcriptionnelle du génome zygotique (AGZ). - A. Modèle de l'état de la chromatine et du comportement de différents acteurs avant l'AGZ. B. Modèle d'ouverture de la chromatine et de l'activation de la transcription pendant l'AGZ. C. Indications apportées par mes analyses. Les lettres indiquent les événements qui ont été apportés par la recherche bibliographique (bleu), par mes analyses (rouge) ou par les deux (violet).

La présence de motifs ressemblant aux motifs de liaison de Trl et TTK dans les régions non-codantes AGZ (figure 5.1Cb) suggère que ce mécanisme pourrait être plus général. Des sites Trl et TTK sont prédits ensemble dans 22% des CRM comportant Trl et/ou TTK. Cependant, Pagans et al. n'ont pas réussi à déterminer si TTK piégeait Trl en dehors de la chromatine ou non.

La présence d'histone H3K9me3 (figure 5.1Aa) dans les promoteurs des gènes AGZ n'est pas évidente, puisque l'analyse de l'enrichissement de cette histone ne montre pas d'association particulière par rapport aux CRM non actifs dans le blastoderme (d'après RedFly) (figure 5.1Ca). Cependant, l'enrichissement est nettement supérieur que pour des CRM prédits dans un groupe de gènes aléatoires. De plus, l'expérience de localisation d'H3K9me3 a été menée entre 0 et 4h, or, si les CRM identifiés sont impliqués dans une activation générale des gènes, les histones H3K9me3 devraient être absentes de ces régions à partir de 1h (première vague de l'AGZ) ou 2h (deuxième vague de l'AGZ). Le signal de localisation de H3K9me3 pourrait ainsi être alors noyé. Bien sûr nous ne pouvons pas écarter une implication de l'avortement de la transcription due au cycles mitotiques très court. J'aurais pu tester cette hypothèse si j'avais trouvé des données de localisation de l'ARNpolII pour la période étudiée. La répression de l'activation des gènes AGZ résulte probablement d'une combinaison de ces différents éléments.

5.1.2 Activation de la transcription

5.1.2.1 Fin de la répression

La répression de la transcription jusqu'à l'AGZ peut-être due au moins aux deux phénomènes discutés précédemment. Tout d'abord, lors du titrage de TTK par l'augmentation du ratio NC, Trl pourrait être libéré et entrer en actio (figure 5.1 Bc) (78). De plus, les données de RNA-seq de Gelbart et Emmert (2010) montrent que l'abondance de Trl augmente entre 2 et 4h (101). TTK pourrait ne réprimer Trl que tant que sa concentration reste faible. Il y aurait donc une complémentarité entre la titration de TTK (soumise au ratio NC) et l'augmentation de Trl (indépendante du ratio NC) pour la régulation de l'activation des gènes AGZ. Ensuite, l'activation est probablement en partie due à la décompaction de la chromatine. Trl est impliqué dans le recrutement de divers complexes de remodelage de la chromatine (figure 5.1Be,f). En effet, Nakayama et al. ont montré que Trl interagissait avec le complexe FACT pour le remplacement ciblé des histones H3K9me3 par un variant H3.3 associé aux régions actives

de la chromatine et largement distribué dans les promoteurs des gènes dans l'embryon précoce (132, 133). Les variants H3.3 sont préférentiellement méthylés sur leur lysine 4 (K4). Nous avons identifié un enrichissement en H3K4me1 dans les CRM AGZ (figure 5.1Bd) qui pourrait être relié à la présence de ces variants. D'autre part, Trl est connu pour remodeler les nucléosomes de façon dépendante à l'ATP en collaborant avec le complexe NURF produisant ainsi des régions chromatiniennes accessibles (134). Enfin, l'ensemble des protéines formant ces complexes est présent dans l'embryon entre 2 et 4h (101).

Une fois la chromatine accessible, les différents facteurs activateurs pourraient venir se lier à leur séquences cibles (figure 5.1Bg,h). Des motifs de Zelda, Trl et "aAGATACa" sont non seulement sur-représentés dans les régions non codantes des gènes AGZ, mais aussi dans les pics chevauchant ces mêmes régions. Le modèle proposé implique (au moins) ces trois motifs.Les autres motifs sont peut-être facultatifs. Par ailleurs, les CRM sont spécifiquement enrichis pour la combinaison de Trl, CBP, H3K4me1 et ouverture de la chromatine (figure 5.1Bi,Ci,j). Ainsi, la combinaison de ces marques pourrait s'expliquer par la liaison des facteurs cités ci-dessus au niveau de régions accessibles de la chromatine. Ces facteurs pourraient recruter CBP, connu pour être recruté par des centaines de facteurs différents, probablement pas tous identifiés. Il est également connu que CBP a une activité acétyl-transférase ciblant les histones. Cependant, nous ne trouvons pas d'enrichissement spécifique des CRM prédits en histone H3 acétylée (K9 et K27). CBP pourrait néanmoins jouer un rôle de pont entre les facteurs de transcription et la machinerie basale de transcription. Enfin, les travaux récents sur Zelda (96) suggèrent clairement que ce facteur joue un rôle prépondérant et général dans l'AGZ. Le modèle proposé pourrait ainsi être applicable à l'activation de la transcription en plusieurs vagues.

5.1.3 Rôles de Trl et modèles alternatifs

Trl est impliqué dans beaucoup de processus lié à l'activation de la transcription.

5.1.3.1 Trl est un facteur de transcription et un co-activateur

Trl possède un domaine riche en glutamine (Q-rich), connu pour stimuler la transcription en interagissant avec TAF3 (un composant du complexe TFIID) au sein d'autres protéines, telles que CREB ou SP1 (135). A cet égard, Vaquero et al. ont montré que Trl augmente la transcription en stabilisant le PIC et en promouvant la réinitiation de la transcription via son

domaine Q-rich (97). Chez la souris, le facteur GAGA, un homologue de Trl, est impliqué dans l'activation du gène hsp70.1, en collaboration avec SP1, au début de l'AGZ (136).

5.1.3.2 Trl pourrait recruter CBP

Le domaine Q-rich est donc impliqué dans des interactions protéine-protéine et pourrait assurer les interactions entre Trl est d'autres FT ou co-facteurs. Même si aucune interaction entre CBP et Trl n'a encore été documentée, la forte représentation de sites riches en GA dans les pics CBP suggère que Trl pourrait interagir avec CBP, et probablement le recruter aux régions transcriptionnellement actives. De plus, CBP contient aussi un domaine riche en Q, qui pourrait être impliqué dans cette interaction.

5.1.3.3 Trl est impliqué dans la pause des ARN polymérase II

Il a été montré que Trl est impliqué dans la pause de l'ARNpolII en coopérant avec NELF au promoteur de hsp70 et d'un grand nombre d'autre gènes (137). Les ARNpolII en pause ont été détectées dans environ 1500 gènes des embryons de drosophile, incluant un grand nombre de gènes contrôlant le développement (138). Hendrix et collaborateurs ont suggéré que Trl induirait une conformation ouverte de la chromatine, permissive pour le recrutement et la pause de l'ARNpolII, en recrutant NURF par exemple (134). De plus, une forte association entre les CRM et les régions de chromatine liées à Trl a été révélée (figure 4.10). De manière intéressante, les signaux de liaison de Trl dans les CRM présente un pic à proximité du TSS (figure5.2A). Par ailleurs, les sites prédits avec la matrice Trl sont particulièrement concentrés entre 50 et 250 pb en amont du TSS (figure5.2B).

5.1.4 Utilisation du modèle pour la sélection de CRM

Afin de sélectionner les CRM les plus pertinents en vue d'une validation expérimentale, j'ai utilisé le classement des CRM obtenu en combinant les classements pour les marques individuelles. J'ai ensuite procédé à la visualisation de ces régions grâce au navigateur de génome du UCSC (http ://genome.ucsc.edu/) pour visualiser chaque région dans son contexte génomique, avec les annotations pertinentes (conservation inter-spécifique de séquence, environnement génique, etc.).

La figure 5.3 montre la région d'environ 6Kb en amont du gène croc (crocodile), impliqué dans la spécification du segment le plus antérieur de la tête. Il s'agit d'un gène purement zy-

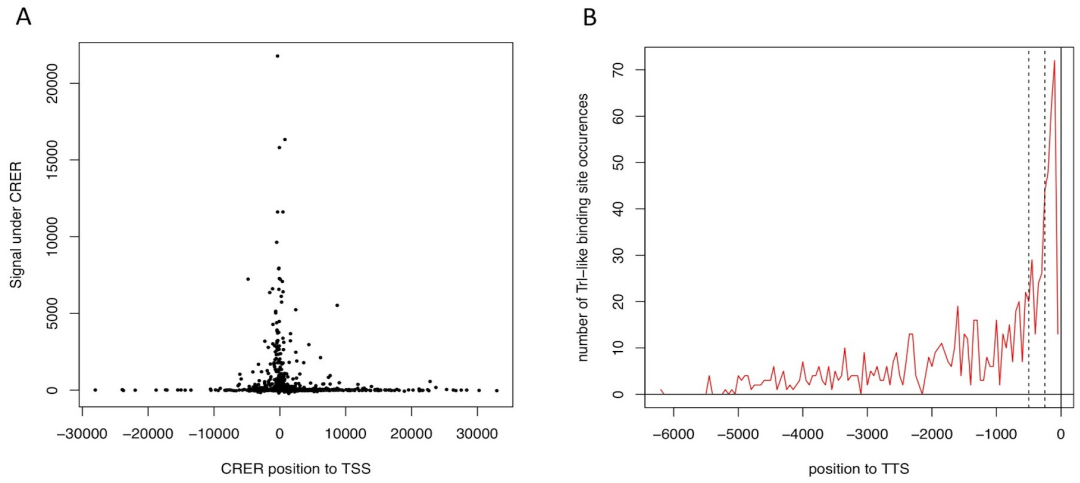


FIGURE 5.2: Analyse positionnelle de l'enrichissement des CRM en signaux de liaison de Trl et sites Trl par rapport au TSS. - A. Chaque point représente un CRM ; l'abscisse indique sa position relative au TSS et l'ordonnée la densité intégrée de reads ChIP-seq Trl sous ce CRM. B. L'abscisse indique la position par rapport au TSS et l'ordonnée le nombre d'occurrences de sites prédits pour Trl dans des fenêtres de 50 pb. Les pointillés indiquent les positions -250 et -500.

gotique dont l'activation est soumise au ratio NC. Nous pouvons noter qu'aucun site de liaison de Zelda n'a été prédit dans le premier CRM (à gauche) alors que deux pics de liaison ont été identifiés par ChIP-seq.

5.2 Conclusion générale

Les travaux menés durant cette thèse m'ont permis de proposer un mécanisme de régulation de l'activation du génome zygotique chez la Drosophile. À partir de données transcriptomiques publiées, j'ai proposé une méthode simple et adaptée à l'analyse de la variation de l'expression des gènes au cours du temps. L'analyse fonctionnelle (profils d'expression, analyse des régions non-codantes, enrichissement en classes fonctionnelles) des différents groupes de gènes co-exprimés m'a permis de sélectionner un groupe étendu de gènes activés durant l'AGZ. Les différents motifs détectés m'ont permis de proposer des facteurs et co-facteurs agissant potentiellement en *trans* et de prédire des régions potentiellement régulatrices. L'analyse de l'association spécifique de ces régions avec des marques chromatiniennes m'a amenée à proposer un modèle combinant l'action de différents facteurs (Zelda, Trl et CBP, ainsi que plusieurs facteurs inconnus) par leur liaison dans des régions accessibles et actives. L'apport de jeux de données

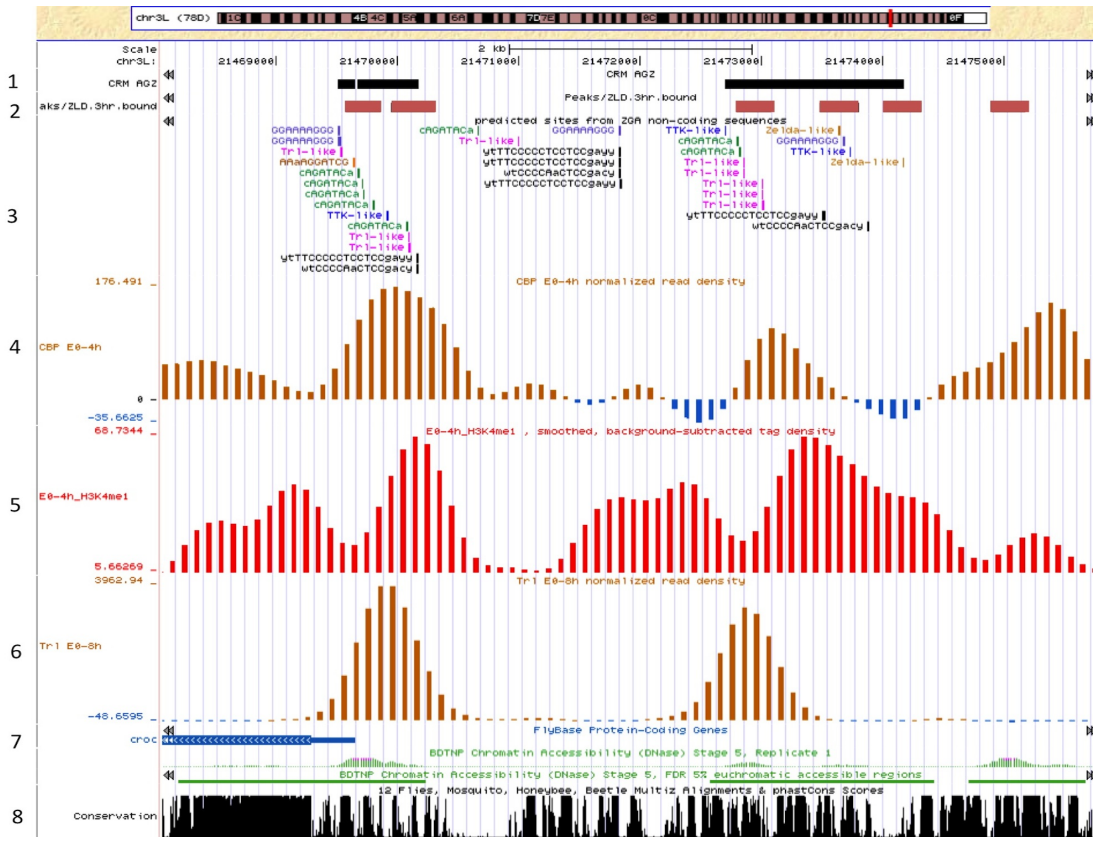


FIGURE 5.3: Visualisation de la région en amont du gène croc. - 1. CRM AGZ. 2. Pics de liaison de Zelda 3h provenant d'expérience de ChIP-seq. 3. Sites prédits en scannant les séquences avec le motif découvert dans le cluster AGZ. 4. Densité normalisée des reads provenant de l'expérience de ChIP-seq de CBP. 5. Densité normalisée des reads provenant de l'expérience de ChIP-seq de H3K4me1. 6. Densité normalisée des reads provenant de l'expérience de ChIP-seq de Trl. 7. Représentation de la partie 5' du gène (gène sur le brin anti-sens). 7. Ouverture de la chromatine. 8. Conservation entre les espèces de drosophiles.

de localisation génomique dans des intervalles temporels plus précis permettrait probablement d'affiner le modèle proposé ici. Enfin, une validation expérimentale permettrait d'apprécier la pertinence de l'ensemble des prédictions et hypothèses générées durant cette thèse.

References

[1] N Woychik and Michael Hampsey. **The RNA Polymerase II MachineryStructure Illuminates Function**. *Cell*, **108**(4) :453–463, February 2002. 2

[2] J. E.F. Butler and James T. Kadonaga. **Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs**. *Genes & Development*, **15** :2515–2519, October 2001. 2

[3] G D Stormo. **DNA binding sites : representation and discovery**. *Bioinformatics (Oxford, England)*, **16**(1) :16–23, January 2000. 3, 6

[4] D J Galas and A Schmitz. **DNAse footprinting : a simple method for the detection of protein-DNA binding specificity**. *Nucleic Acids Research*, **5**(9) :3157–3170, September 1978. 3

[5] A M Maxam and W Gilbert. **A new method for sequencing DNA**. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(2) :560–564, February 1977. 3

[6] R Stoltenburg, C Reinemann, and B Strehlitz. **SELEX,ÄîA (r)evolutionary method to generate high-affinity nucleic acid ligands**. *Biomolecular Engineering*, **24**(4) :381–403, October 2007. 3

[7] J. Wang, J. Lu, G. Gu, and Y. Liu. **In vitro DNA-binding profile of transcription factors : methods and new insights**. *Journal of Endocrinology*, **210**(1) :15–27, March 2011. 5

[8] Xiangdong Meng, Michael H Brodsky, and Scot A Wolfe. **A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors**. *Nature Biotechnology*, **23**(8) :988–994, August 2005. 5

[9] Xiangdong Meng and Scot A Wolfe. **Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system**. *Nature Protocols*, **1**(1) :30–45, June 2006. 5

[10] Martha L Bulyk. **Discovering DNA regulatory elements with bacteria**. *Nature Biotechnology*, **23**(8) :942–944, August 2005. 5

[11] Michael J Buck and Jason D Lieb. **ChIP-chip : considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments**. *Genomics*, **83**(3) :349–360, March 2004. 5

[12] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. **Genome-wide identification of in vivo protein,ÄìDNA binding sites from ChIP-Seq data**. *Nucleic acids research*, **36**(16) :5221, 2008. 5

[13] Dusan Štanojević, Timothy Hoey, and Michael Levine. **Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Krüppel in Drosophila**. *Nature*, **341**(6240) :331–335, September 1989. 6

[14] P Bucher and B Bryan. **Signal search analysis : a new method to localize and characterize functionally important DNA sequences.** *Nucleic acids research*, **12**(1 Pt 1) :287–305, January 1984. 6

[15] R Staden. **Computer methods to locate signals in nucleic acid sequences**. January 1984. 6

[16] G. B. Ehret, P Reichenbach, U Schindler, C M Horvath, S Fritz, M Nabholz, and P Bucher. **DNA binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites.** *The Journal of biological chemistry*, **276**(9) :6675–88, March 2001. 6

[17] W H Day and F R McMorris. **Critical comparison of consensus methods for molecular sequences**. *Nucleic Acids Research*, **20**(5) :1093–1099, March 1992. 6, 12

[18] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. **WebLogo : a sequence logo generator**. *Genome Research*, **14**(6) :1188–1190, June 2004. 6

[19] T D Schneider and R M Stephens. **Sequence logos : a new way to display consensus sequences**. *Nucleic Acids Research*, **18**(20) :6097–6100, October 1990. 7

[20] G Z Hertz and G D Stormo. **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences**. *Bioinformatics (Oxford, England)*, **15**(7-8) :563–577, August 1999. 7, 14

[21] T.L. Bailey, Mikael Boden, F.A. Buske, Martin Frith, C.E. Grant, Luca Clementi, Jingyuan Ren, W.W. Li, and W.S. Noble. **MEME SUITE : tools for motif discovery and searching**. *Nucleic Acids Research*, **37**(suppl 2) :W202–W208, 2009. 7, 73

[22] G Z Hertz, G W 3rd Hartzell, and G D Stormo. **Identification of consensus patterns in unaligned DNA sequences known to be functionally related**. *Computer Applications in the Biosciences : CABIOS*, **6**(2) :81–92, April 1990. 7

[23] Matthieu Defrance and J. Van Helden. **info-gibbs : a motif discovery algorithm that directly optimizes information content during sampling**. *Bioinformatics*, **25**(20) :2715–2722, 2009. 7

[24] Gert Thijs, Kathleen Marchal, Magali Lescot, Stephane Rombauts, Bart De Moor, Pierre Rouzé, and Yves Moreau. **A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes**. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, **9**(2) :447–464, 2002. 14

[25] Jean-Valery Turatsinze, Morgane Thomas-Chollier, Matthieu Defrance, and Jacques van Helden. **Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules.** *Nature protocols*, **3**(10) :1578–88, January 2008. 14, 17, 76, 81

[26] Jean Valery Turatsinze. **Développement et évaluation de méthodes bioinformatiques pour la détection de séquences cis-régulatrices impliquées dans le développement de la drosophile**. *Catalogue des theses electroniques de l ULB*, November 2009. 14, 18, 19

[27] Wyeth W. Wasserman and Albin Sandelin. **Applied bioinformatics for the identification of regulatory elements.** *Nature reviews. Genetics*, **5**(4) :276–87, April 2004. 14, 17

[28] Gabriela G Loots, Ivan Ovcharenko, Lior Pachter, Inna Dubchak, and Edward M Rubin. **rVista for Comparative Sequence-Based Discovery of Functional Transcription Factor Binding Sites**. *Genome Research*, **12**(5) :832–839, April 2002. 17

REFERENCES

[29] Robert K Bradley, Xiao-yong Li, Cole Trapnell, Stuart Davidson, Lior Pachter, Hou Cheng Chu, Leath A Tonkin, Mark D Biggin, and Michael B Eisen. **Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species.** *PLoS biology*, **8**(3) :e1000343, March 2010. 17, 34

[30] M.C. Frith, M.C. Li, and Zhiping Weng. **Cluster-Buster : Finding dense clusters of motifs in DNA sequences**. *Nucleic acids research*, **31**(13) :3666–3668, 2003. 17

[31] Benjamin P Berman, Barret D Pfeiffer, Todd R Laverty, Steven L Salzberg, Gerald M Rubin, Michael B Eisen, and Susan E Celniker. **Computational identification of developmental enhancers : conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.** *Genome biology*, **5**(9) :R61, January 2004. 17, 34

[32] Anthony A Philippakis, Fangxue Sherry He, and Martha L Bulyk. **Modulefinder : a tool for computational discovery of cis regulatory modules.** *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 519–30, January 2005. 17

[33] A. Sosinsky. **Target Explorer : an automated tool for the identification of new target genes for a specified set of transcription factors**. *Nucleic Acids Research*, **31**(13) :3589–3592, July 2003. 17

[34] Ian John Donaldson, Michael Chapman, and Berthold Göttgens. **TFBScluster : a resource for the characterization of transcriptional regulatory networks.** *Bioinformatics (Oxford, England)*, **21**(13) :3058–9, July 2005. 17

[35] Benjamin P. Berman, Yutaka Nibu, Barret D Pfeiffer, Pavel Tomancak, Susan E Celniker, Michael Levine, Gerald M Rubin, and Michael B Eisen. **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 of the United States of America*, **99**(2) :757–62, January 2002. 17, 34

[36] Peter Van Loo and Peter Marynen. **Computational methods for the detection of cis-regulatory modules.** *Briefings in bioinformatics*, **10**(5) :509–24, September 2009. 17

[37] Yutao Fu and Zhiping Weng. **Improvement of TRANSFAC matrices using multiple local alignment of transcription factor binding site sequences.** *Genome informatics. International Conference on Genome Informatics*, **16**(1) :68–72, January 2005. 20

[38] Elodie Portales-Casamar, Supat Thongjuea, Andrew T Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W Wasserman, and Albin Sandelin. **JASPAR 2010 : the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic acids research*, **38**(Database issue) :D105–10, January 2010. 20, 66, 82

[39] Lihua Julie Zhu, Ryan G Christensen, Majid Kazemian, Christopher J Hull, Metewo Selase Enuameh, Matthew D Basciotta, Jessie A Brasefield, Cong Zhu, Yuna Asriyan, David S Lapointe, Saurabh Sinha, Scot A Wolfe, and Michael H Brodsky. **FlyFactorSurvey : a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system.** *Nucleic acids research*, **39**(Database issue) :D111–7, January 2011. 20, 66, 81

[40] Ivan V Kulakovskiy, Alexander V Favorov, and Vsevolod J Makeev. **Motif discovery and motif finding from genome-mapped DNase footprint data.** *Bioinformatics (Oxford, England)*, **25**(18) :2318–25, September 2009. 20

[41] Socorro Gama-Castro, Heladia Salgado, Martin Peralta-Gil, Alberto Santos-Zavaleta, Luis Muñiz Rascado, Hilda Solano-Lira, Verónica Jimenez-Jacinto, Verena Weiss, Jair S García-Sotelo, Alejandra López-Fuentes, Liliana Porrón-Sotelo, Shirley Alquicira-Hernández, Alejandra Medina-Rivera, Irma Martínez-Flores, Kevin Alquicira-Hernández, Ruth Martínez-Adame, César Bonavides-Martínez, Juan Miranda-Ríos, Araceli M Huerta, Alfredo Mendoza-Vargas, Leonardo Collado-Torres, Blanca Taboada, Leticia Vega-Alvarado, Maricela Olvera, Leticia Olvera, Ricardo Grande, Enrique Morett, and Julio Collado-Vides. **RegulonDB version 7.0 : transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).** *Nucleic acids research*, **39**(Database issue) :D98–105, January 2011. 20

[42] Elodie Portales-Casamar, Stefan Kirov, Jonathan Lim, Stuart Lithwick, Magdalena I Swanson, Amy Ticoll, Jay Snoddy, and Wyeth W Wasserman. **PAZAR : a framework for collection and dissemination of cis-regulatory sequence annotation.** *Genome biology*, **8**(10) :R207, January 2007. 21

[43] Elodie Portales-Casamar, David Arenillas, Jonathan Lim, Magdalena I Swanson, Steven Jiang, Anthony McCallum, Stefan Kirov, and Wyeth W Wasserman. **The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences.** *Nucleic acids research*, **37**(Database issue) :D54–60, January 2009. 21

[44] Casey M Bergman, Joseph W Carlson, and Susan E Celniker. **Drosophila DNase I footprint database : a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster**. *Bioinformatics*, **21**(8) :1747–1749, November 2004. 21, 81

[45] Thomas A Down, Casey M Bergman, Jing Su, and Tim J P Hubbard. **Large-scale discovery of promoter motifs in Drosophila melanogaster.** *PLoS computational biology*, **3**(1) :e7, January 2007. 21, 82

[46] Marc S Halfon, Steven M Gallo, and Casey M Bergman. **REDfly 2.0 : an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila.** *Nucleic acids research*, **36**(Database issue) :D594–8, January 2008. 21

[47] Marcus B Noyes, Ryan G Christensen, Atsuya Wakabayashi, Gary D Stormo, Michael H Brodsky, and Scot A Wolfe. **Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites.** *Cell*, **133**(7) :1277–89, June 2008. 21, 81

[48] Marcus B Noyes, Xiangdong Meng, Atsuya Wakabayashi, Saurabh Sinha, Michael H Brodsky, and Scot A Wolfe. **A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system.** *Nucleic acids research*, **36**(8) :2547–60, May 2008. 21, 81

[49] Alan M Moses, Daniel A Pollard, David A Nix, Venky N Iyer, Xiao-Yong Li, Mark D Biggin, and Michael B Eisen. **Large-scale turnover of functional transcription factor binding sites in Drosophila.** *PLoS computational biology*, **2**(10) :e130, October 2006. 21

[50] Xiao-yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L Luengo Hendriks, Hou Cheng Chu, Nobuo Ogawa, William Inwood, Victor Sementchenko, Amy Beaton, Richard Weiszmann, Susan E Celniker, David W Knowles, Tom Gingeras, Terence P Speed, Michael B Eisen, and Mark D Biggin. **Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm.** *PLoS biology*, **6**(2) :e27, February 2008. 21, 34

[51] R J Kelleher, P M Flanagan, and R D Kornberg. **A novel mediator between activator proteins and the RNA polymerase II transcription apparatus.** *Cell*, **61**(7) :1209–15, June 1990. 20

[52] Amelia Casamassimi and Claudio Napoli. **Mediator complexes and eukaryotic transcription regulation : an overview.** *Biochimie*, **89**(12) :1439–46, December 2007. 22

[53] Floris Bosveld, Sjoerd van Hoek, and Ody C M Sibon. **Establishment of cell fate during early Drosophila embryogenesis requires transcriptional Mediator subunit dMED31.** *Developmental biology*, **313**(2) :802–13, January 2008. 22, 29, 34

[54] Karolin Luger, A W Mäder, R K Richmond, D F Sargent, and T J Richmond. **Crystal structure of the nucleosome core particle at 2.8 A resolution.** *Nature*, **389**(6648) :251–60, September 1997. 22

[55] Susumu Hirose. **Crucial roles for chromatin dynamics in cellular memory.** *Journal of biochemistry*, **141**(5) :615–9, May 2007. 22

[56] E. B. Lewis. **A gene complex controlling segmentation in Drosophila**. *Nature*, **276**(5688) :565–570, December 1978. 23

[57] Christiane Nüsslein-Volhard and Eric Wieschaus. **Mutations affecting segment number and polarity in Drosophila**. *Nature*, **287**(5785) :795–801, October 1980. 23

[58] Wael Tadros and Howard D Lipshitz. **The maternal-to-zygotic transition : a play in two acts.** *Development (Cambridge, England)*, **136**(18) :3033–42, September 2009. 25, 33

[59] Justin Crest, Nathan Oxnard, Jun-Yuan Ji, and Gerold Schubiger. **Onset of the DNA replication checkpoint in the early Drosophila embryo.** *Genetics*, **175**(2) :567–84, February 2007. 24, 33

[60] Beatrice Benoit, Chun Hua He, Fan Zhang, Sarah M Votruba, Wael Tadros, J Timothy Westwood, Craig A Smibert, Howard D Lipshitz, and William E Theurkauf. **An essential role for the RNA-binding protein Smaug during the Drosophila maternal-to-zygotic transition.** *Development (Cambridge, England)*, **136**(6) :923–32, March 2009. 24, 31, 33

[61] C A Smibert, J E Wilson, K Kerr, and P M Macdonald. **smaug protein represses translation of unlocalized nanos mRNA in the Drosophila embryo.** *Genes & development*, **10**(20) :2600–9, October 1996. 26

[62] Rongwen Xi, Jennifer R McGregor, and Douglas A Harrison. **A gradient of JAK pathway activity patterns the anterior-posterior axis of the follicular epithelium.** *Developmental cell*, **4**(2) :167–77, February 2003. 28

[63] Jens Januschke, Louis Gervais, Laurent Gillet, Guy Keryer, Michel Bornens, and Antoine Guichet. **The centrosome-nucleus complex and microtubule organization in the Drosophila oocyte.** *Development (Cambridge, England)*, **133**(1) :129–39, January 2006. 28

[64] Natalie Denef and Trudi Schüpbach. **Patterning : JAK-STAT signalling in the Drosophila follicular epithelium.** *Current biology : CB*, **13**(10) :R388–90, May 2003. 28

[65] Dierk Niessing, Stephen Blanke, and Herbert Jäckle. **Bicoid associates with the 5'-cap-bound complex of caudal mRNA and represses translation.** *Genes & development*, **16**(19) :2576–82, October 2002. 29

[66] Scott Gilbert. *Developmental biology*. Sinauer, Sunderland Mass., 6.ed. edition, 2000. 30

[67] Fanny Pilot, Jean-Marc Philippe, Céline Lemmers, Jean-Paul Chauvin, and Thomas Lecuit. **Developmental control of nuclear morphogenesis and anchoring by charleston, identified in a functional genomic screen of Drosophila cellularisation.** *Development (Cambridge, England)*, **133**(4) :711–23, February 2006. 29, 34, 37, 40, 43, 45, 126

[68] Stefano De Renzis, Olivier Elemento, Saeed Tavazoie, and Eric F Wieschaus. **Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo.** *PLoS biology*, **5**(5) :e117, May 2007. 29, 31, 32, 33, 34, 35, 38, 46, 52, 66, 67

[69] Sean D Hooper, Stephanie Boué, Roland Krause, Lars J Jensen, Christopher E Mason, Murad Ghanim, Kevin P White, Eileen E M Furlong, and Peer Bork. **Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis.** *Molecular systems biology*, **3**(3) :72, January 2007. 29

[70] Eric Lécuyer, Hideki Yoshida, Neela Parthasarathy, Christina Alm, Tomas Babak, Tanja Cerovina, Timothy R Hughes, Pavel Tomancak, and Henry M Krause. **Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function.** *Cell*, **131**(1) :174–187, 2007. 29, 31

[71] Xuemin Lu, Jennifer M Li, Olivier Elemento, Saeed Tavazoie, and Eric F Wieschaus. **Coupling of zygotic transcription to mitotic control at the Drosophila mid-blastula transition.** *Development (Cambridge, England)*, **136**(12) :2101–10, June 2009. 29, 34, 37, 46

[72] Wael Tadros, Aaron L Goldman, Tomas Babak, Fiona Menzies, Leah Vardy, Terry Orr-Weaver, Timothy R Hughes, J Timothy Westwood, Craig A Smibert, and Howard D Lipshitz. **SMAUG Is a Major Regulator of Maternal mRNA Destabilization in Drosophila and Its Translation Is Activated by the PAN GU Kinase**. *Developmental Cell*, **12**(1) :143–155, 2007. 31, 33

[73] Stefan Thomsen, Simon Anders, Sarath Chandra Janga, Wolfgang Huber, and Claudio R Alonso. **Genome-wide analysis of mRNA decay patterns during early Drosophila development.** *Genome biology*, **11**(9) :R93, January 2010. 31, 32

[74] Natascha Bushati, Alexander Stark, Julius Brennecke, and Stephen M Cohen. **Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in Drosophila.** *Current biology : CB*, **18**(7) :501–6, April 2008. 31

[75] André P Gerber, Stefan Luschnig, Mark A Krasnow, Patrick O Brown, and Daniel Herschlag. **Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster.** *Proceedings of the National Academy of Sciences of the United States of America*, **103**(12) :4487–92, March 2006. 31

[76] B A Edgar and S A Datar. **Zygotic degradation of two maternal Cdc25 mRNAs terminates Drosophila's early cell cycle program.** *Genes & development*, **10**(15) :1966–77, August 1996. 32

[77] B A Edgar and P H O'Farrell. **Genetic control of cell division patterns in the Drosophila embryo.** *Cell*, **57**(1) :177–87, April 1989. 32, 33

[78] D K Pritchard and G Schubiger. **Activation of transcription in Drosophila embryos is a gradual process mediated by the nucleocytoplasmic ratio.** *Genes & development*, **10**(9) :1131–42, May 1996. 32, 67, 113

[79] O C Sibon, V A Stevenson, and W E Theurkauf. **DNA-replication checkpoint control at the Drosophila midblastula transition.** *Nature*, **388**(6637) :93–7, July 1997. 33

REFERENCES

[80] D Read, T Nishigaki, and J L Manley. **The Drosophila even-skipped promoter is transcribed in a stage-specific manner in vitro and contains multiple, overlapping factor-binding sites.** *Molecular and cellular biology*, **10**(8) :4334–44, August 1990. 34

[81] Gregory T Reeves and Angelike Stathopoulos. **Graded dorsal and differential gene regulation in the Drosophila embryo.** *Cold Spring Harbor perspectives in biology*, **1**(4) :a000836, October 2009. 34

[82] John R ten Bosch, Joseph A Benavides, and Thomas W Cline. **The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription.** *Development (Cambridge, England)*, **133**(10) :1967–77, May 2006. 34, 66

[83] Hsiao-Lan Liang, Chung-Yi Nien, Hsiao-Yun Liu, Mark M Metzstein, Nikolai Kirov, and Christine Rushlow. **The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila.** *Nature*, **456**(7220) :400–3, December 2008. 34, 66

[84] Amy Tsurumi, Fan Xia, Jinghong Li, Kimberly Larson, Russell Lafrance, and Willis X Li. **STAT Is an Essential Activator of the Zygotic Genome in the Early Drosophila Embryo.** *PLoS genetics*, **7**(5) :e1002086, May 2011. 34

[85] E. Hubbell, W.-M. Liu, and R. Mei. **Robust estimators for expression analysis**. *Bioinformatics*, **18**(12) :1585–1592, December 2002. 38

[86] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. **Summaries of Affymetrix GeneChip probe level data.** *Nucleic acids research*, **31**(4) :e15, February 2003. 38, 40

[87] W.-m Liu, R. Mei, X. Di, T. B. Ryder, E. Hubbell, S. Dee, T. A. Webster, C. A. Harrington, M.-h. Ho, J. Baid, and S. P. Smeekens. **Analysis of high density expression microarrays with signed-rank call algorithms**. *Bioinformatics*, **18**(12) :1593–1599, December 2002. 38

[88] Stuart D Pepper, Emma K Saunders, Laura E Edwards, Claire L Wilson, and Crispin J Miller. **The utility of MAS5 expression summary and detection call algorithms.** *BMC bioinformatics*, **8** :273, January 2007. 38, 40

[89] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics (Oxford, England)*, **4**(2) :249–64, April 2003. 40

[90] Zhijin Wu and Rafael A. Irizarry. **Stochastic models inspired by hybridization theory for short oligonucleotide arrays.** *Journal of computational biology : a journal of computational molecular cell biology*, **12**(6) :882–93, July 2005. 40

[91] Bettina Harr and Christian Schlötterer. **Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons.** *Nucleic acids research*, **34**(2) :e8, January 2006. 40

[92] R Development Core Team. **R : A language and environment for statistical computing.** pages ISBN 3–900051–07–0, 2008. 40

[93] Eleanor Howe, Kristina Holton, Sarita Nair, Daniel Schlauch, Raktim Sinha, and John Quackenbush. **MeV : MultiExperiment Viewer**. *Biomedical Informatics for Cancer Research*, pages 267–277, 2010. 42

[94] Sylvain Brohée, Karoline Faust, Gipsi Lima-Mendez, Gilles Vanderstocken, and Jacques van Helden. **Network Analysis Tools : from biological networks to clusters and pathways.** *Nature protocols*, **3**(10) :1616–29, January 2008. 59, 69

[95] Morgane Thomas-Chollier, Olivier Sand, Jean-Valéry Turatsinze, Rekin's Janky, Matthieu Defrance, Eric Vervisch, Sylvain Brohée, and Jacques van Helden. **RSAT : regulatory sequence analysis tools.** *Nucleic acids research*, **36**(Web Server issue) :W119–27, July 2008. 60, 125

[96] Melissa M. Harrison, Xiao-Yong Li, Tommy Kaplan, Michael R. Botchan, and Michael B. Eisen. **Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition**. *PLoS Genetics*, **7**(10) :e1002266, October 2011. 66, 74, 86, 114

[97] Alejandro Vaquero, Marta Blanch, M Lluïsa Espinás, and Jordi Bernués. **Activation properties of GAGA transcription factor.** *Biochimica et biophysica acta*, **1779**(5) :312–7, May 2008. 66, 115

[98] H Granok, B A Leibovitch, C D Shaffer, and S C Elgin. **Chromatin. Ga-ga over GAGA factor.** *Current biology : CB*, **5**(3) :238–41, March 1995. 66

[99] Tsukasa Shimojima, Masahiro Okada, Takahiro Nakayama, Hitoshi Ueda, Katsuya Okawa, Akihiro Iwamatsu, Hiroshi Handa, and Susumu Hirose. **Drosophila FACT contributes to Hox gene expression through physical and functional interactions with GAGA factor.** *Genes & development*, **17**(13) :1605–16, July 2003. 66

[100] Michael Lehmann. **Anything else but GAGA : a nonhistone protein complex reshapes chromatin structure.** *Trends in genetics : TIG*, **20**(1) :15–22, January 2004. 66

[101] W.M. Gelbart and D.B. Emmert. **FlyBase High Throughput Expression Pattern Data Beta Version.**, 2010. 66, 67, 113, 114

[102] Sara Pagans, Miguel Ortiz-Lombardía, Ma Lluïsa Espinás, Jordi Bernués, and Fernando Azorín. **The Drosophila transcription factor tramtrack (TTK) interacts with Trithorax-like (GAGA) and represses GAGA-mediated activation.** *Nucleic acids research*, **30**(20) :4406–13, October 2002. 67, 76, 111

[103] F Hirose, M Yamaguchi, K Kuroda, A Omori, T Hachiya, M Ikeda, Y Nishimoto, and A Matsukage. **Isolation and characterization of cDNA for DREF, a promoter-activating factor for Drosophila DNA replication-related genes.** *The Journal of biological chemistry*, **271**(7) :3930–7, February 1996. 68

[104] Craig M. Hart, Olivier Cuvier, and Ulrich K. Laemmli. **Evidence for an antagonistic relationship between the boundary element-associated factor BEAF and the transcription factor DREF**. *Chromosoma*, **108**(6) :375–383, November 1999. 68

[105] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. **The DAVID Gene Functional Classification Tool : a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome biology*, **8**(9) :R183, January 2007. 69

[106] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nature protocols*, **4**(1) :44–57, January 2009. 69

[107] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. **GOToolBox : functional analysis of gene datasets based on Gene Ontology.** *Genome biology*, **5**(12) :R101, January 2004. 69

[108] Douglas A Hosack, Glynn Dennis, Brad T Sherman, H Clifford Lane, and Richard A Lempicki. **Identifying biological themes within lists of genes with EASE.** *Genome biology*, **4**(10) :R70, January 2003. 70

[109] Stein Aerts, Xiao-Jiang Quan, Annelies Claeys, Marina Naval Sanchez, Phillip Tate, Jiekun Yan, and Bassem A Hassan. **Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in Drosophila uncovers a regulatory basis for sensory specification.** *PLoS biology*, **8**(7) :e1000435, January 2010. 73

[110] Long Li, Qianqian Zhu, Xin He, Saurabh Sinha, and Marc S Halfon. **Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses.** *Genome biology*, **8**(6) :R101, January 2007. 73

[111] V Morel and F Schweisguth. **Repression by suppressor of hairless and activation by Notch are required to define a single row of single-minded expressing cells in the Drosophila embryo.** *Genes & development*, **14**(3) :377–88, February 2000. 76, 135

[112] Olivier Sand, Morgane Thomas-Chollier, and Jacques van Helden. **Retrieve-ensembl-seq : user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl.** *Bioinformatics (Oxford, England)*, **25**(20) :2739–40, October 2009. 78

[113] Matthieu Defrance, Rekin's Janky, Olivier Sand, and Jacques van Helden. **Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences.** *Nature protocols*, **3**(10) :1589–603, January 2008. 80

[114] Sheryl T Smith, Svetlana Petruk, Yurii Sedkov, Elizabeth Cho, Sergei Tillib, Eli Canaani, and Alexander Mazo. **Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex.** *Nature cell biology*, **6**(2) :162–7, February 2004. 85

[115] M. Mannervik. **Transcriptional Coregulators in Development**. *Science*, **284**(5414) :606–609, April 1999. 85

[116] Tobias Lilja, Hitoshi Aihara, Marianne Stabell, Yutaka Nibu, and Mattias Mannervik. **The acetyltransferase activity of Drosophila CBP is dispensable for regulation of the Dpp pathway in the early embryo.** *Developmental biology*, **305**(2) :650–8, May 2007. 85

[117] Hiroshi Akimaru, Yang Chen, Ping Dai, D X Hou, Maki Nonaka, Sarah M. Smolik, Steve Armstrong, Richard H. Goodman, and Shunsuke Ishii. **Drosophila CBP is a co-activator of cubitus interruptus in hedgehog signalling.** *Nature*, **386**(6626) :735–8, April 1997. 85

[118] H Akimaru, D X Hou, and S Ishii. **Drosophila CBP is required for dorsal-dependent twist gene expression.** *Nature genetics*, **17**(2) :211–4, October 1997. 85

[119] L Waltzer and M Bienz. **Drosophila CBP represses the transcription factor TCF to antagonize Wingless signalling.** *Nature*, **395**(6701) :521–5, October 1998. 85

[120] Hye-Kyung Lee, Ui-Hyun Park, Eun-Joo Kim, and Soo-Jong Um. **MED25 is distinct from TRAP220/MED1 in cooperating with CBP for retinoid receptor activation.** *The EMBO journal*, **26**(15) :3545–57, August 2007. 85

[121] N Vo and R H Goodman. **CREB-binding protein and p300 in transcriptional regulation.** *The Journal of biological chemistry*, **276**(17) :13505–8, April 2001. 86

[122] Feng Tie, Rakhee Banerjee, Carl A. Stratton, Jayashree Prasad-Sinha, Vincent Stepanik, Andrei Zlobin, Manuel O. Diaz, Peter C. Scacheri, and Peter J. Harte. **CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing.** *Development (Cambridge, England)*, **136**(18) :3131–41, September 2009. 86

[123] Nicholas T Crump, Catherine A Hazzalin, Erin M Bowers, Rhoda M Alani, Philip A Cole, and Louis C Mahadevan. **Dynamic acetylation of all lysine-4 trimethylated histone H3 is evolutionarily conserved and mediated by p300/CBP.** *Proceedings of the National Academy of Sciences of the United States of America*, **108**(19) :7814–9, May 2011. 86

[124] Nicolas Nègre, Christopher D Brown, Lijia Ma, Christopher Aaron Bristow, Steven W Miller, Ulrich Wagner, Pouya Kheradpour, Matthew L Eaton, Paul Loriaux, Rachel Sealfon, Zirong Li, Haruhiko Ishii, Rebecca F Spokony, Jia Chen, Lindsay Hwang, Chao Cheng, Richard P Auburn, Melissa B Davis, Marc Domanus, Parantu K Shah, Carolyn A Morrison, Jennifer Zieba, Sarah Suchy, Lionel Senderowicz, Alec Victorsen, Nicholas A Bild, A Jason Grundstad, David Hanley, David M MacAlpine, Mattias Mannervik, Koen Venken, Hugo Bellen, Robert White, Mark Gerstein, Steven Russell, Robert L Grossman, Bing Ren, James W Posakony, Manolis Kellis, and Kevin P White. **A cis-regulatory map of the Drosophila genome.** *Nature*, **471**(7339) :527–31, March 2011. 86, 100

[125] Maxim Nekrasov, Tetyana Klymenko, Sven Fraterman, Bernadett Papp, Katarzyna Oktaba, Thomas Köcher, Adrian Cohen, Hendrik G Stunnenberg, Matthias Wilm, and Jürg Müller. **Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes.** *The EMBO journal*, **26**(18) :4078–88, September 2007. 86

[126] Sean Thomas, Xiao-yong Li, Peter J Sabo, Richard Sandstrom, Robert E Thurman, Theresa K Canfield, Erika Giste, William Fisher, Ann Hammonds, Susan E Celniker, Mark D Biggin, and John A Stamatoyannopoulos. **Dynamic reprogramming of chromatin accessibility during Drosophila embryo development.** *Genome biology*, **12**(5) :R43, May 2011. 86, 87

[127] Xiao-Yong Li, Sean Thomas, Peter J Sabo, Michael B Eisen, John A Stamatoyannopoulos, and Mark D Biggin. **The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding.** *Genome biology*, **12**(4) :R34, April 2011. 86

[128] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. **Model-based analysis of ChIP-Seq (MACS).** *Genome biology*, **9**(9) :R137, January 2008. 87

[129] Aaron R. Quinlan and Ira M. Hall. **BEDTools : a flexible suite of utilities for comparing genomic features.** *Bioinformatics (Oxford, England)*, **26**(6) :841–2, March 2010. 87

[130] Pauline A Fujita, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, Galt P Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R Dreszer, Belinda M Giardine, Rachel A Harte, Jennifer Hillman-Jackson, Fan Hsu, Vanessa Kirkup, Robert M Kuhn, Katrina Learned, Chin H Li, Laurence R Meyer, Andy Pohl, Brian J Raney, Kate R Rosenbloom, Kayla E Smith, David Haussler, and W James Kent. **The UCSC Genome Browser database : update 2011.** *Nucleic acids research*, **39**(Database issue) :D876–82, January 2011. 87, 125

[131] Steven M Gallo, Dave T Gerrard, David Miner, Michael Simich, Benjamin Des Soye, Casey M Bergman, and Marc S Halfon. **REDfly v3.0 : toward a comprehensive database of transcriptional regulatory elements in Drosophila.** *Nucleic acids research*, **39**(Database issue) :D118–23, January 2011. 93

[132] Takahiro Nakayama, Kenichi Nishioka, Yi-Xin Dong, Tsukasa Shimojima, and Susumu Hirose. **Drosophila GAGA factor directs histone H3.3 replacement that prevents the heterochromatin spreading.** *Genes & development*, **21**(5) :552–61, March 2007. 114

[133] Emmanuelle Szenker, Dominique Ray-Gallet, and Geneviève Almouzni. **The double face of the histone variant H3.3.** *Cell research*, **21**(3) :421–34, March 2011. 114

[134] T Tsukiyama and C Wu. **Purification and properties of an ATP-dependent nucleosome remodeling factor.** *Cell*, **83**(6) :1011–20, December 1995. 114, 115

[135] Vivek Sarojkumar Chopra, Arumugam Srinivasan, Ram Parikshan Kumar, Krishnaveni Mishra, Denis Basquin, Mylène Docquier, Carole Seum, Daniel Pauli, and Rakesh Kumar Mishra. **Transcriptional activation by GAGA factor is through its direct interaction with dmTAF3.** *Developmental biology*, **317**(2) :660–70, May 2008. 114

[136] A Bevilacqua, M T Fiorenza, and F Mangia. **A developmentally regulated GAGA box-binding factor and Sp1 are required for transcription of the hsp70.1 gene at the onset of mouse zygotic genome activation.** *Development (Cambridge, England)*, **127**(7) :1541–51, April 2000. 115

[137] Chanhyo Lee, Xiaoyong Li, Aaron Hechmer, Michael Eisen, Mark D Biggin, Bryan J Venters, Cizhong Jiang, Jian Li, B Franklin Pugh, and David S Gilmour. **NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila.** *Molecular and cellular biology*, **28**(10) :3290–300, May 2008. 115

[138] David A Hendrix, Joung-Woo Hong, Julia Zeitlinger, Daniel S Rokhsar, and Michael S Levine. **Promoter elements associated with RNA Pol II stalling in the Drosophila embryo.** *Proceedings of the National Academy of Sciences of the United States of America*, **105**(22) :7762–7, June 2008. 115

6

Annexes

6.1 Résultats supplémentaires

6.2 Articles

6.2.1 From Peaks to Motifs : A complete workflow for full-sized ChIP-seq (and like) dataset

Ce protocole explique la façon d'utilser *peak-motifs* via son interface web (http ://rsat.ulb.ac.be/rsat/). *peak-motifs* est un programme dédié à la découverte de motifs sur-représentés dans des jeux de séquences complets provenant d'expériences de ChIP-seq et ChIP-chip. Ce programme intègre de nombreux d'outils de la suite RSAT (95). En effet, quatre algorithmes de découverte de motifs sont disponibles afin d'identifier des motifs significativement sur-représentés dans les jeux de pics. Ces motifs sont ensuite comparés à des motifs connus présents dans des bases de données. Les séquences sont ensuite scannés afin de prédire des sites de liaison, d'analyser leur enrichissement et leur position par rapport au centre des pics. Les pics analysés et les sites prédits à partir des motifs découverts peuvent être exportés vers le navigateur de génome de UCSC (130) afin d'être visualiser dans leur contexte génomique.

J'ai contribué à cet article en analysant divers cas d'études et en produisant certaines figures.

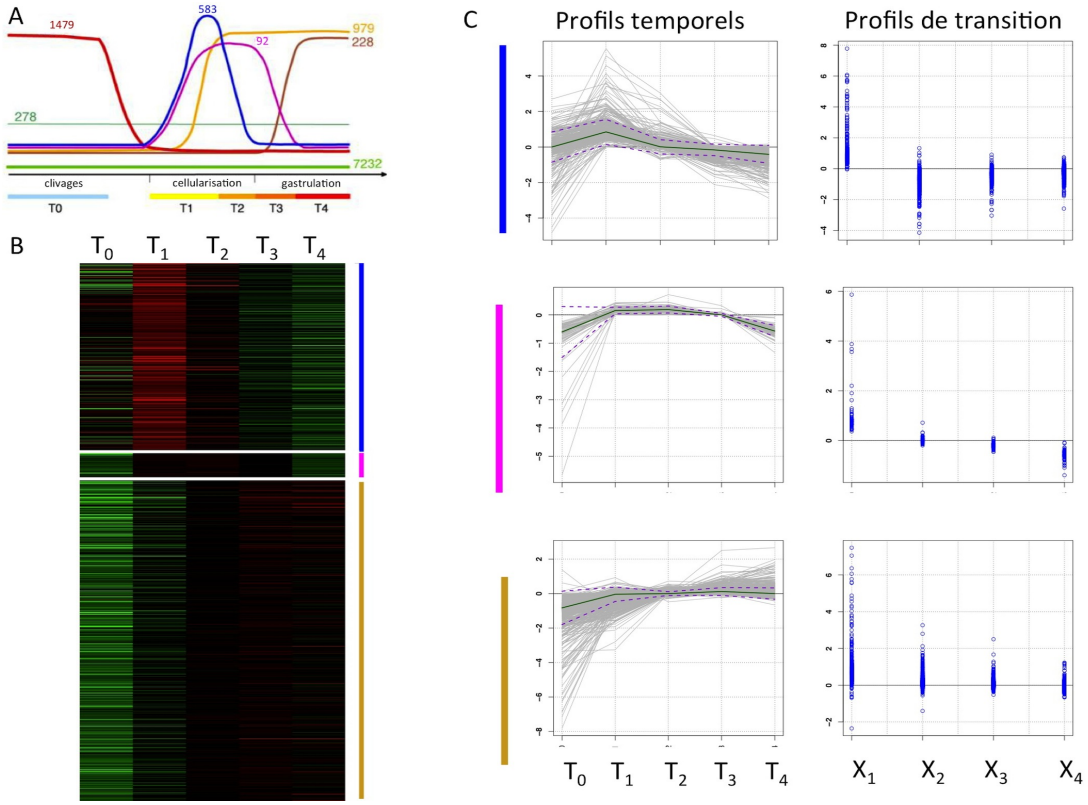


FIGURE 6.1: Exploration des clusters groupant les gènes potentiellement activés pendant l'AGZ selon Pilot et collaborateurs obtenus par la méthode de détection MAS5.0. - A. Représentation des différents groupes de co-expressions (D'après Pilot et al. (67)). B. Heatmap des profils temporels entre T0 et T4, vert, noir et rouge : intensité inférieure, égale ou supérieure à par l'intensité médiane du gène au travers de tous les points temporels, respectivement des trois groupes de gènes activés pendant la cellularisation (profils bleu, rose et jaune du panneau A). C. Profils temporels correspondant aux clusters d'intérêts où l'ordonnée indique la valeur d'intensité (log2) pour chaque gène (chaque courbe grise correspond à un gène) standardisée par l'intensité médiane du gène au travers de tous les points temporels. La ligne verte représente le profil médian et les lignes pointillées représentent la 1 déviation standart par rapport à la médiane. À droite : Profils de transition (X1 à X4) entre points temporels consécutifs. L'ordonnée indique la valeur du log ratio et chaque point représente un gène.

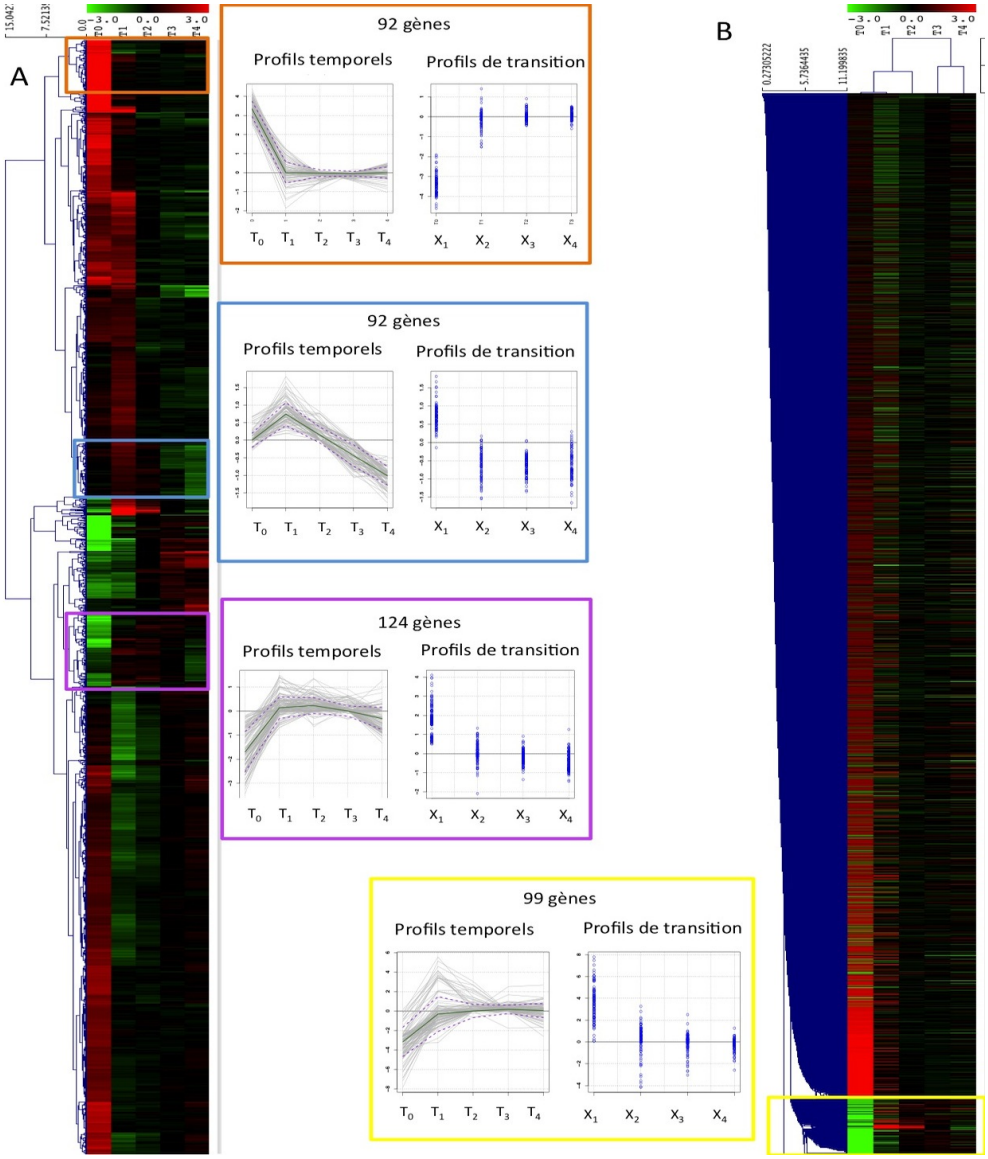


FIGURE 6.2: Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression. - Pour chaque panel A et B, les cadres de couleur indiquent les mêmes clusters représentés sous différentes formes : à droite, heatmap des profils temporels entre T0 et T4, vert, noir et rouge : intensité inférieure, égale ou supérieure à l'intensité médiane du gène au travers de tous les points temporels, respectivement. Au milieu : profils temporels, où l'ordonnée indique la valeur d'intensité (log2) pour chaque gène (chaque courbe grise correspond à un gène) standardisée par l'intensité médiane du gène dans l'ensemble des classes temporelles. La ligne verte représente le profil médian et les lignes pointillées représentent la 1 écart-type par rapport à la médiane. À gauche : profils de transition (X_{i1} à X_{i4}) entre classes temporelles consécutives. L'ordonnée indique la valeur du log ratio et chaque point représente un gène. A. regroupement par clustering hiérarchique, distance : euclidienne, règle d'agglomération : complète. B. regroupement par clustering hiérarchique, distance : produit scalaire moyen, règle d'agglomération : simple.

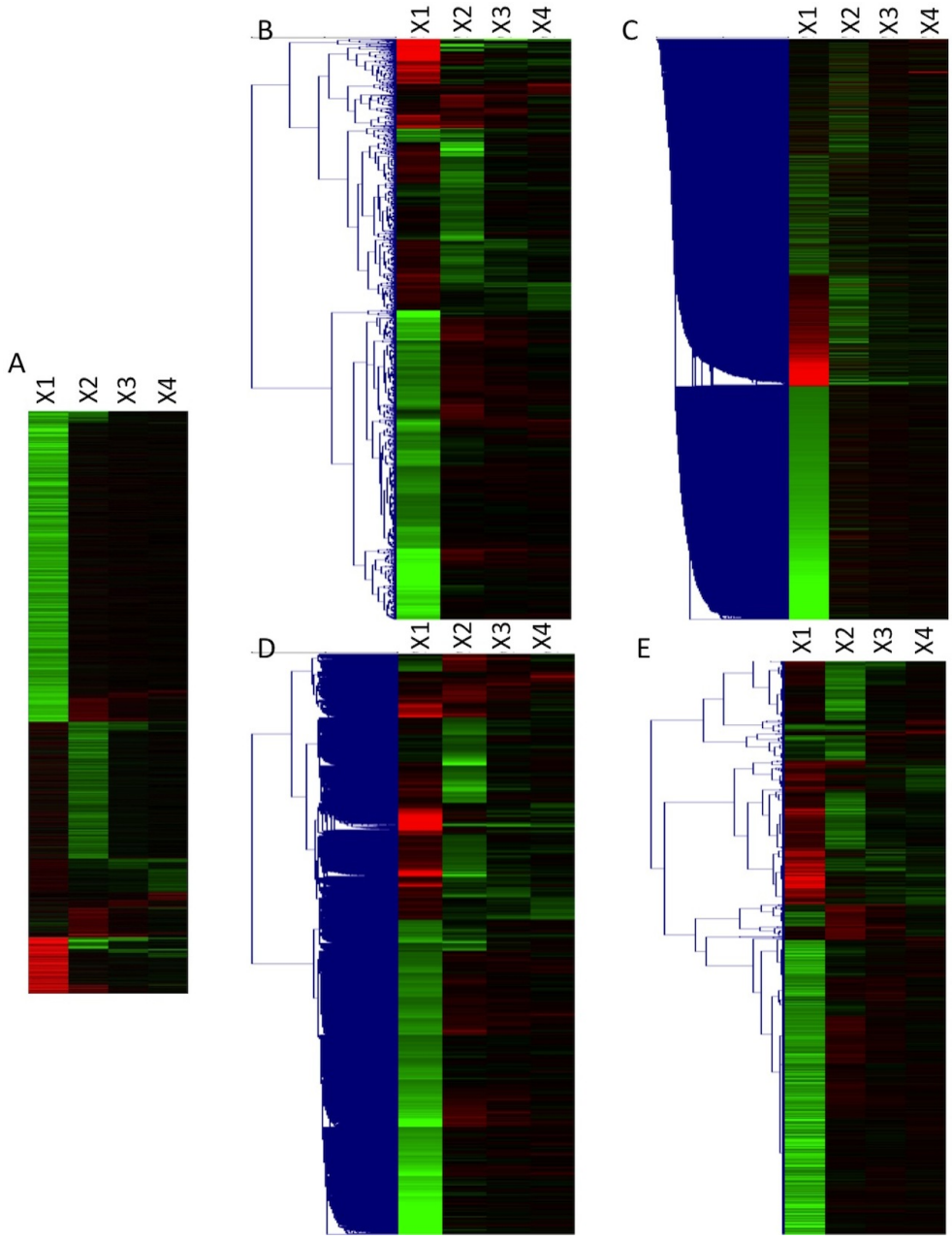


FIGURE 6.3: Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression à partir des profils de transitions. - Heatmap représentant les clusters obtenus à partir des profils de transition (X_{i1} à X_{i4}) entre points temporels consécutifs avec différentes méthodes. A. Groupement par profils de transition discrets identiques. B. regroupement par clustering hiérarchique, distance : euclidienne, règle d'agglomération : complète. C. regroupement par clustering hiérarchique, distance : produit scalaire moyen, règle d'agglomération : simple. D. regroupement par clustering hiérarchique, distance : produit scalaire moyen, règle d'agglomération : complète. E. regroupement par clustering hiérarchique, distance : corrélation de Pearson, règle d'agglomération : moyenne.

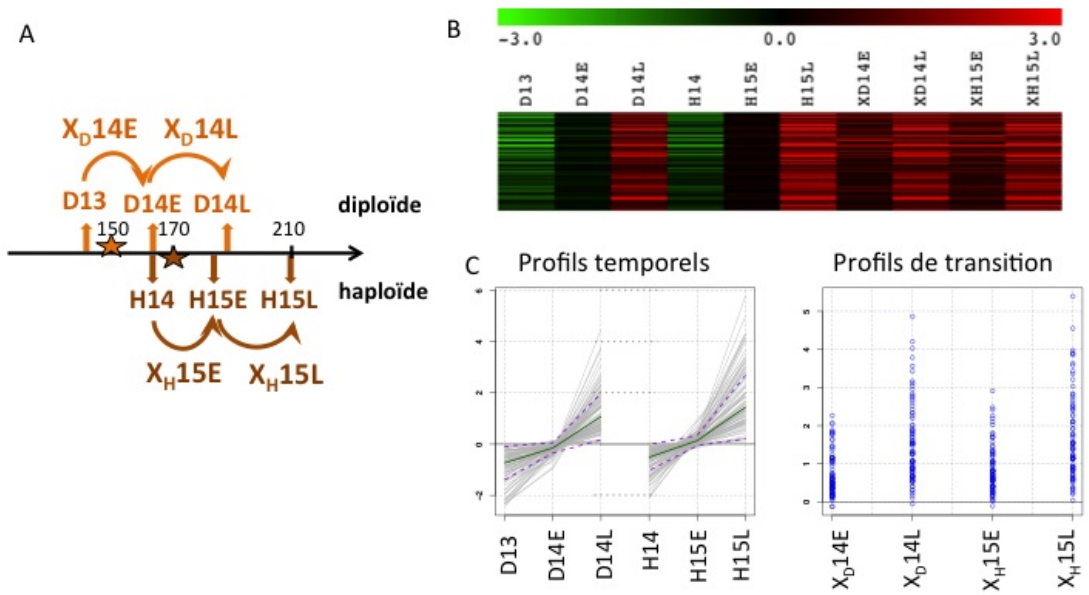


FIGURE 6.4: Exploration des clusters groupant les gènes potentiellement activés dépendamment du ratio NC pendant l'AGZ selon Lu et collaborateurs. - A. Echelle temporelle des données de puces à partir d'embryons sauvages diploïdes (haut), et mutants haploïdes (bas). L'axe horizontal représente le temps en minutes. Les étoiles représentent le moment de la cellularisation dans chaque génotype. B. Heatmap représentant les profils temporels (D13 à H15L) et les profils de transition (X_D14E à X_H15L). C. Représentation graphique des profils temporels et de transition.

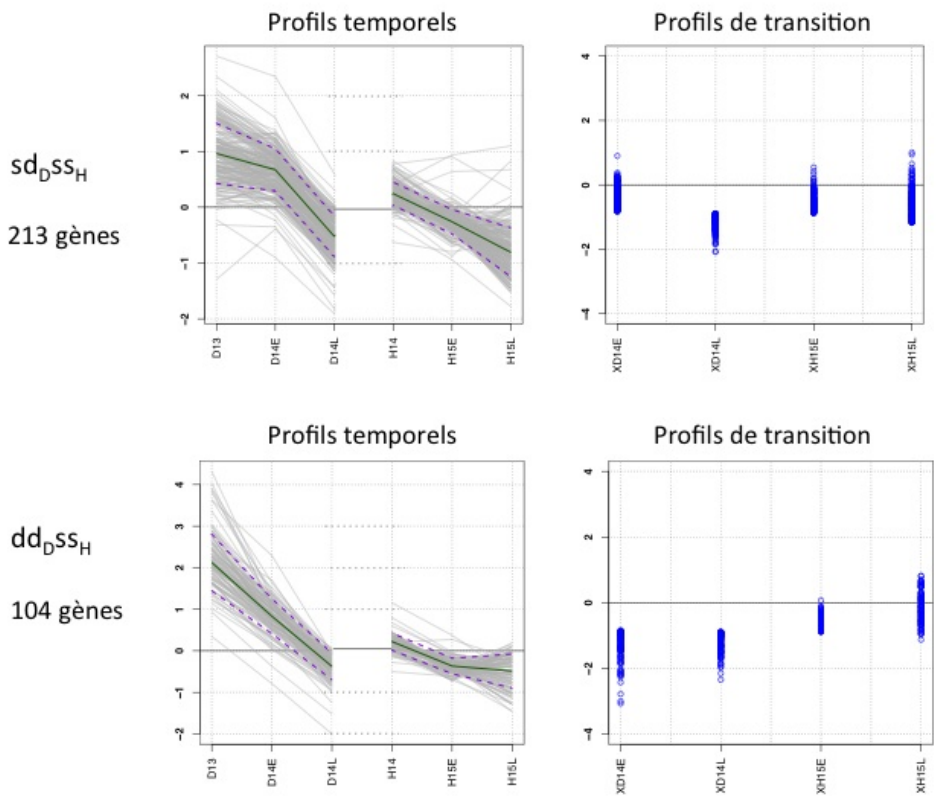


FIGURE 6.5: Profils temporels et de transition des clusters les plus importants ne correspondant à aucun modèle de régulation connu obtenus à partir des données de Lu et collaborateurs -

TABLE 6.1: Codes des évidences pour l'annotation des gènes dans GO et composition des classes.

Evidence code	Signification	Biological Process	Fonction Moléculaire	Composante Cellulaire	Total
	Évidences expérimentales				
IDA:	Inferred from Direct Assay	190 (1000)	680 (1230)	363 (2429)	1733 (3175)
IPI:	Inferred from Physical Interaction	69 (82)	113 (589)	60 (155)	242 (687)
IMP:	Inferred from Mutant Phenotype	1971 (4550)	197 (237)	35 (54)	2203 (4568)
IGI:	Inferred from Genetic Interaction	409 (949)	37 (50)	4 (6)	450 (967)
IEP:	Inferred from Expression Pattern	153 (428)	0 (0)	0 (0)	153 (428)
	Computational Analysis Evidence Codes				
ISS:	Inferred from Sequence or Structural Similarity	557 (2671)	1203 (5259)	305 (2788)	2065 (6125)
ISO:	Inferred from Sequence Orthology	0 (0)	1 (1)	1 (1)	2 (2)
ISA:	Inferred from Sequence Alignment	12 (61)	19 (74)	2 (2)	33 (75)
ISM:	Inferred from Sequence Model	3 (30)	53 (302)	7 (104)	63 (425)
IGC:	Inferred from Genomic Context	0 (0)	0 (0)	0 (0)	0 (0)
IBA:	Inferred from Biological aspect of Ancestor	0 (0)	0 (0)	0 (0)	0 (0)
IBD:	Inferred from Biological aspect of Descendant	0 (0)	0 (0)	0 (0)	0 (0)
IKR:	Inferred from Key Residues	0 (0)	0 (0)	0 (0)	0 (0)
IRD:	Inferred from Rapid Divergence	10 (1)	5 (1)	2 (1)	17 (2)
RCA:	inferred from Reviewed Computational Analysis	12 (2)	6 (19)	8 (3)	26 (21)
	Author Statement Evidence Codes				
TAS:	Traceable Author Statement	776 (988)	197 (388)	107 (294)	1080 (1147)
NAS:	Non-traceable Author Statement	864 (2406)	681 (2595)	255 (1950)	1810 (3724)
	Curator Statement Evidence Codes				
IC:	Inferred by Curator	162 (224)	62 (50)	44 (147)	268 (346)
ND:	No biological Data available	103 (1318)	37 (2262)	21 (2323)	161 (2673)
	Automatically-assigned Evidence Codes				
IEA:	Inferred from Electronic Annotation	519 (4145)	521 (5209)	150 (2396)	1190 (6931)
	Obsolete Evidence Codes				
NR:	Not Recorded	218 (40)	49 (41)	26 (40)	293 (43)
Total		2977 (11068)	1905 (11251)	626 (9146)	5513 (13090)

TABLE 6.2: Top 10 des 7-mères surreprésentés suivant divers modèle de background. Modèle basé sur les séquences en amont du TTS (upstream) de *Drosophila melanogaster* (dm) (A), de *Spirochaeta thermophila* (st) (B), de *Mus musculus* (mm) (C), le modèle de markov d'ordre 2 (D), les séquences 3'UTR *Drosophila melanogaster* (E), séquences 5'UTR *Drosophila melanogaster* (F). 0 indique une significativité ≤ 0.

A	bg all upstream dm	bg mkv2 ZGA dm	bg all 5utr dm	bg all firstintron dm	bg all 3utr dm	bg all non coding dm	bg all upstream mm	bg all upstream st
caggtag	40,71	57,78	45,48	44,97	74,22	44,8	0,62	45,24
gcaggta	20,5	15,91	22,25	24,43	37,95	23,01	0	14,39
aggtaga	11,61	20,37	12,68	10,71	9,84	11,47	0	0
caggtaa	10,48	8,87	11,54	13,94	32,67	12,75	0	6,52
ccaggta	7,11	2,96	9,83	12,31	24,98	10,07	0	0
tacctga	5,71	2,69	8,95	8,72	10,2	7,65	0	0
aggtagg	5,3	9,99	6,16	6,37	3,1	5,67	0	0
acaggta	5,19	5,17	7,83	7,93	16,14	7,2	0	0
aggatcg	5	0,1	8,21	8,88	0	6,04	15,48	0
ctcgctc	4,83	25,8	1,58	5,04	27,19	4,74	16,99	0

B	bg all upstream dm	bg mkv2 ZGA dm	bg all 5utr dm	bg all firstintron dm	bg all 3utr dm	bg all non coding dm	bg all upstream mm	bg all upstream st
attaaat	0	6,44	0	0	0	0	99,88	350
aattaaa	0	11,47	0	0	0	0	99,09	350
caattaa	0	2,65	0	0	0	0	94,84	350
aatttaa	0	0	0	0	0	0	82,15	350
aaattaa	0	1,81	0	0	0	0	76,77	350
aaattta	0	0	0	0	0	0	75,02	350
tatttaa	0	6,95	0	0	0	0	56,9	350
attaaaa	0	2,79	0	0	0	0	56,6	350
caaatta	0	0	0	0	0	0	56,31	350
aattggc	0	0	0	0	0,76	0	47,36	350

C	bg all upstream dm	bg mkv2 ZGA dm	bg all 5utr dm	bg all firstintron dm	bg all 3utr dm	bg all non coding dm	bg all upstream mm	bg all upstream st
attttcg	0	0	0	0	0	0	258,85	49,39
gcgaaaa	0	0	0	0	11,52	0	241,14	45,41
aatcgaa	0	1,65	0	0	0	0	239,6	18,6
aaatcga	0	0	0	0	0	0	218,02	45,71
atttcga	0	0	0	0	0	0	190,21	26,97
atcgaaa	0	0	0	0	0	0	183,21	0
aatttcg	0	0	0	0	0	0	177,33	31,12
aacgaaa	0	6,39	0	0	0	0	175,34	19,3
ccgaaaa	0	0	0	0	11,56	0	173,23	0
aaaatcg	0	0	0	0	0	0	172,58	36,81

D	bg all upstream dm	bg mkv2 ZGA dm	bg all 5utr dm	bg all firstintron dm	bg all 3utr dm	bg all non coding dm	bg all upstream mm	bg all upstream st
atatgta	0	73,95	0	0	0	0	37,55	185,2
caggtag	40,71	57,78	45,48	44,97	74,22	44,8	0,62	45,24
cacacac	0	56,18	0	0	0	0	0	0
atacata	0	52,39	0	0	0	0	14,63	232,29
acacaca	0	37,79	0	0	0	0	0	29,56
cgctctc	1,41	34,4	1,66	4,64	37,54	3,42	23,14	0
ataaata	0	32,68	0	0	0	0	34,62	121,15
gtatgta	0	29,47	0	0	0	0	3,63	13,13
aaacaaa	0	28,78	0	0	0	0	0	303,75
acaacaa	0	28,19	0	0	0	0	38,45	198,67

E	bg all upstream dm	bg mkv2 ZGA dm	bg all 5utr dm	bg all firstintron dm	bg all 3utr dm	bg all non coding dm	bg all upstream mm	bg all upstream st
caggtag	40,71	57,78	45,48	44,97	74,22	44,8	0,62	45,24
gcaggta	20,5	15,91	22,25	24,43	37,95	23,01	0	14,39
cgctctc	1,41	34,4	1,66	4,64	37,54	3,42	23,14	0
caggtaa	10,48	8,87	11,54	13,94	32,67	12,75	0	6,52
gggaaaa	0	7,89	0	0	32,15	0	8,17	0
ggccaaa	0	2,49	0	0	31,28	0	55,33	164,13
gccataa	0	3,07	0	0	30,15	0	75,33	228,96
ccataaa	1,01	11,84	0,66	0	29,75	1,02	83,82	74,01
ccttttc	2,76	13,1	0	0	28,97	1,57	0	0
aaaaggg	0,53	6,31	0	0	27,86	0	5,7	0

F	bg all upstream dm	bg mkv2 ZGA dm	bg all 5utr dm	bg all firstintron dm	bg all 3utr dm	bg all non coding dm	bg all upstream mm	bg all upstream st
caggtag	40,71	57,78	45,48	44,97	74,22	44,8	0,62	45,24
gcaggta	20,5	15,91	22,25	24,43	37,95	23,01	0	14,39
aggtaga	11,61	20,37	12,68	10,71	9,84	11,47	0	0
caggtaa	10,48	8,87	11,54	13,94	32,67	12,75	0	6,52
ccaggta	7,11	2,96	9,83	12,31	24,98	10,07	0	0
tacctga	5,71	2,69	8,95	8,72	10,2	7,65	0	0
aggatcg	5	0,1	8,21	8,88	0	6,04	15,48	0
acaggta	5,19	5,17	7,83	7,93	16,14	7,2	0	0
aggacat	1,04	8,08	6,63	6,48	8,36	4,23	0	3,69
aggtagg	5,3	9,99	6,16	6,37	3,1	5,67	0	0

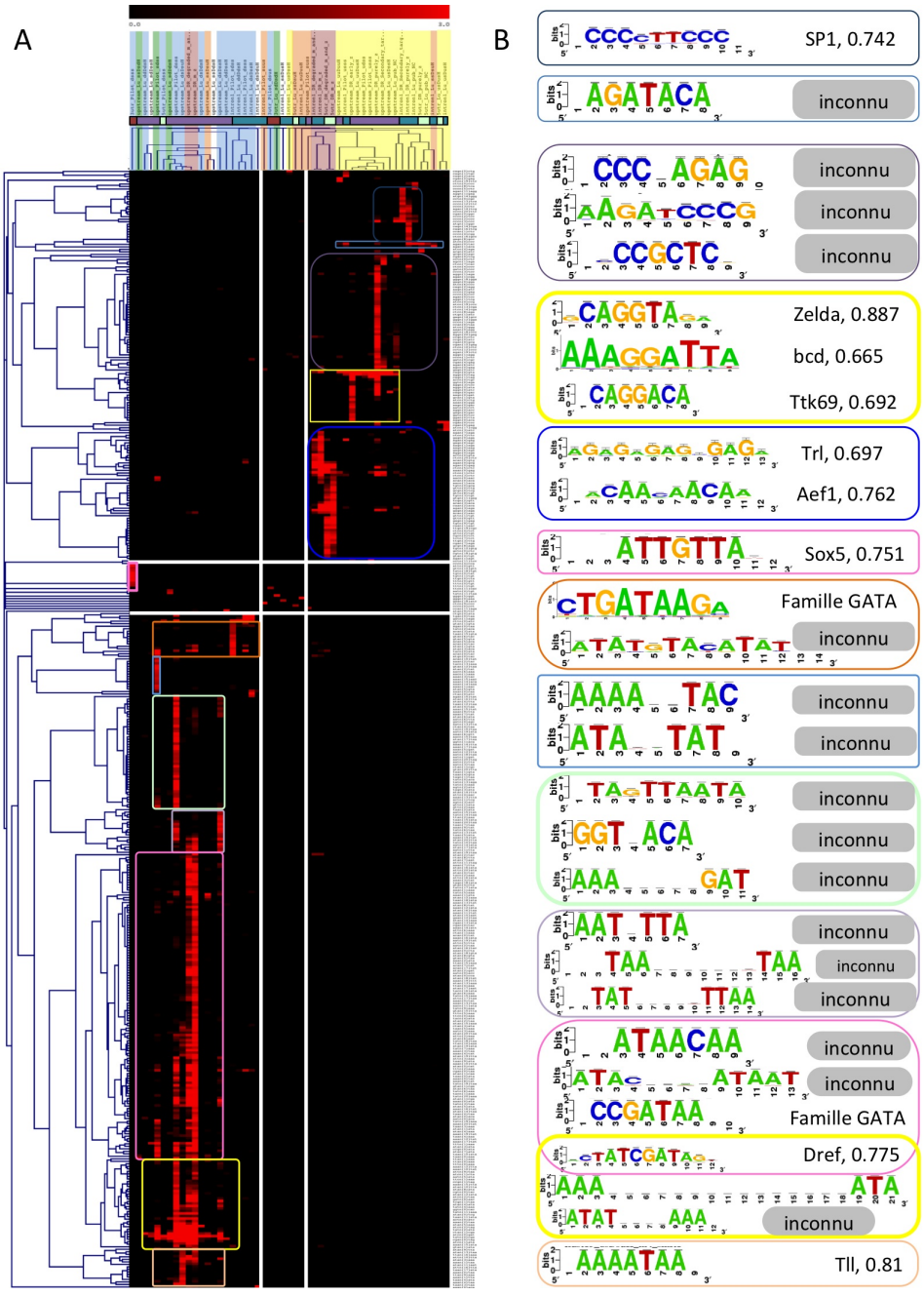


FIGURE 6.6: Clustering des dyades et des groupes de gènes en fonction de la significativité de la surreprésentation des dyades. -

Motif	Enrichment score	Logos	Recovery curves
vfl_SANGER_5_FBgn0259789	12.759135723067		
CAGGTAG	11.1380836470664		
PF0076.1-CAGGTA	11.0717022210421		
vfl_SOLEXA_5_FBgn0259789	10.2502212575165		
CASGTAR	9.74366411671557		
TACCTGC	8.85525369244997		
MA0340.1-MOT3	4.24295891331343		
GAGAGAG	3.46645717797162		
CTCTCTC	3.46645717797162		
M00043-I-DL_01	3.35453118268841		
dl_FlyReg_FBgn0000462	3.24723752294856		
dl	3.24723752294856		
RAAMGGRTTA-Kruppel	3.24664910672761		
M00774-V-NFKB_Q6_01	3.19416468733339		
Gsc_Cell_FBgn0010323	3.17917161127211		
bcd_FlyReg_FBgn0000166	3.13986886898448		
bcd	3.13986886898448		
Trl_FlyReg_FBgn0013263	3.09551843852693		
Trl	3.09551843852693		
MA0205.1-Trl	3.09551843852693		
M00723-I-GAGAFACTOR_Q6	3.04762020438423		
Oc_SOLEXA_FBgn0004102	3.00892895346348		
Ptx1_SOLEXA_FBgn0020912	2.98415778431874		
TIFDMEM0000009	2.93688834790237		
CGTGNGAA	2.93685373518348		
Bcd_SOLEXA_FBgn0000166	2.93221563085364		
MA0022.1-dl_1	2.92012425439175		
M01111-V-RBPJK_Q4	2.9091693288664		
M00412-V-AREB6_01	2.86589189269411		
Bcd_Cell_FBgn0000166	2.85202949878289		
AGAGAGCG	2.83936124367301		
Kr_FlyReg_FBgn0001325	2.79895089438123		
Kr	2.79895089438123		
PF0086.1-GGGNNTTTCC	2.79666645493519		
MA0306.1-GIS1	2.7539658974109		
MA0372.1-RPH1	2.73781906405363		
AGAGAGAG	2.73031964162975		
M01112-V-RBPJK_01	2.65071615699347		
Kr_NAR_FBgn0001325	2.60858670933069		
M00234-I-SUH_01	2.59301098583493		
MA0023.1-dl_2	2.58319251124613		
Gsc_SOLEXA_FBgn0010323	2.5705069497768		
M00935-V-NFAT_Q4_01	2.56416705343538		
PH0035.1-Gsc	2.53477508632023		

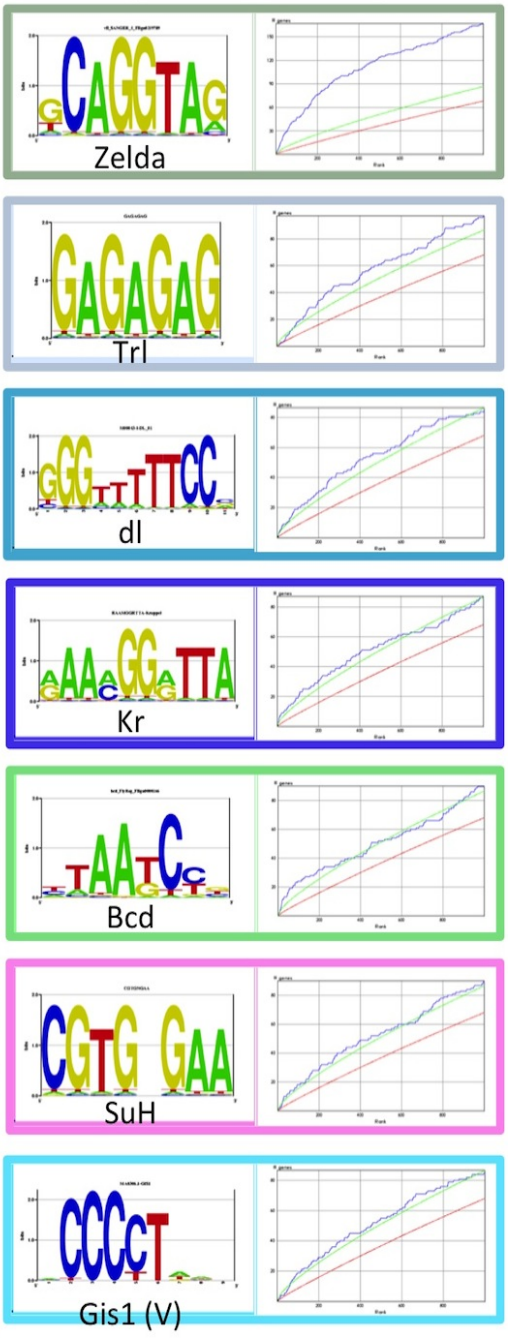


FIGURE 6.7: Résumé des résultats CisTargetX pour le cluster AGZ. - Les lignes marquées de la même couleur de fond indiquent des motifs similaires. Le logo et la courbe de recouvrement représentent le motif ayant le plus haut score d'enrichissement dans chaque groupe et le nom indique le facteur de transcription de drosophile s'y liant, sauf Gis1 ou le (V) indique que c'est un facteur vertébré.

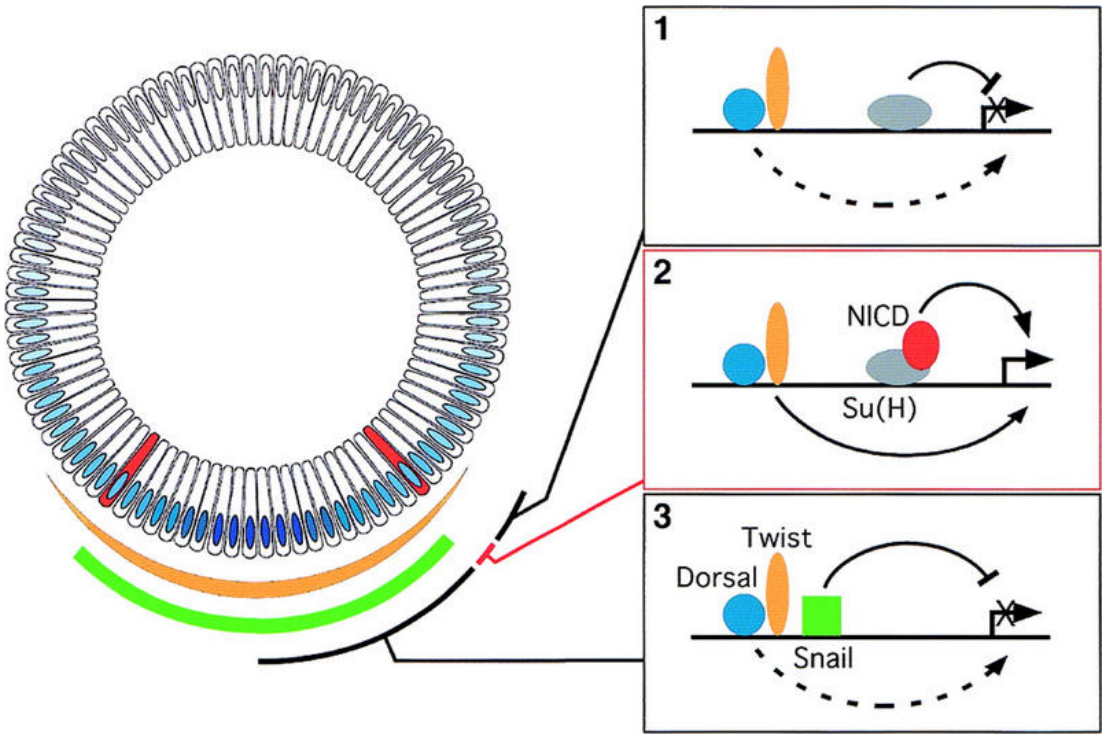


FIGURE 6.8: Action de Su(H) dans la définition du mésectoderme. - En haut : partie dorsale ; en bas : partie ventrale. Gradient orange : Twist, Gradient bleu : Dorsal, vert : Snail. Les cellules rouges représentent le mésectoderme. (D'après Morel et Schweisguth, 2000 (111))

From peaks to motifs: a complete workflow for full-sized ChIP-seq (and like) datasets

Morgane Thomas-Chollier[1], Elodie Darbo[2], Carl Herrmann[2], Matthieu Defrance[3], Denis Thieffry[2,4] and Jacques van Helden[5,2]

1. Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany. Email: thomas-c@molgen.mpg.de
2. Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Université de la Méditerranée. Campus de Luminy, F - 13288 Marseille, France. Email: herrmann@tagc.univ-mrs.fr , darbo@tagc.univ-mrs.fr
3. Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México. Av. Universidad, Cuernavaca, Morelos 62210, Mexico. Email: defrance@ccg.unam.mx
4. IBENS - UMR ENS & CNRS 8197 & INSERM 1024, 46 rue d'Ulm, 75005 Paris, France. Email: thieffry@ens.fr
5. Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium. Email: Jacques.van.Helden@ulb.ac.be

ABSTRACT

This protocol explains how to use the online integrated pipeline *peak-motifs* (http://rsat.ulb.ac.be/rsat/) to predict motifs and binding sites in full-sized peak sets obtained by ChIP-seq or related technologies. The workflow combines four time- and memory-efficient motif discovery algorithms to extract significant motifs from the sequences. Discovered motifs are compared to databases of known motifs to identify potentially bound transcription factors. Sequences are scanned to predict binding sites, analyze their enrichment and positional distribution relative to peak centers. Peaks and binding sites are exported as BED tracks that can be uploaded to the UCSC genome browser to be visualized in their genomic context. This protocol is illustrated with the analysis of a set of 6000 peaks (8Mb in total) bound by the Drosophila transcription factor Krüppel. The complete workflow is achieved in about 25 minutes on the RSAT Web server. This protocol can be followed in about one hour.

ABBREVIATIONS

PSSM	Position-Specific Scoring Matrix (also called position-weight matrices depending on the authors).
BED	A standard format for files describing a list of genomic features (e.g. peaks, sites, gene coordinates, etc).
FASTA	A standard format for sequence files.
ChIP-seq	Combination of chromatin Immunoprecipitation and massively parallel sequencing to characterize the DNA fragments bound to a tagged protein.

INTRODUCTION

The ChIP-seq technology[1,2] enables genome-wide detection of transcription factor binding sites and epigenetic marks. The method typically returns several millions of short reads, which are mapped onto the reference genome and analyzed to extract peak regions, *i.e.* regions presenting a significantly high

density of reads. The typical result is a list of several thousands peak regions of varying sizes (from a few tens base pairs to several kilobases). There is a crucial need for efficient and user-accessible tools to extract relevant information from high-throughput sequencing data[3,4]. While various programs have been developed to perform read mapping and peak calling[5], the subsequent steps have not yet reached proper maturation: identifying relevant transcription factor binding motifs and the precise location of their binding sites remains a bottleneck. Most existing tools present limitations on sequence size, and typically restrict motif discovery to a few hundreds peaks [6,7], or to the central-most part of the peaks [8].

In order to interpret genome-wide location analyses, there is a crucial need of time- and memory-efficient algorithms equipped with in a user-friendly interface. To this purpose, we developed the software tool *peak-motifs*, which takes as input a set of peak sequences of interest ("test sequences"), discovers significant motifs, compares them with transcription factor binding motifs from various databases, predicts the location of binding sites within the peaks and exports them in a format suitable for visualization in the UCSC genome browser. Importantly, all those steps including motif discovery are performed on the full-sized sets of peak sequences, without restriction on the number of peaks or on their width.

The main analytical steps of the workflow (Figure 1) are summarized hereafter, and will be illustrated and further discussed in the next sections.

1. **Sequence purging.** Sequences are automatically purged to discard redundant fragments (peak overlaps, duplications), which would bias the detection of over-represented motifs.
2. **Peak length distribution.** The distribution of sequence lengths provides a useful way to detect outlier peaks (i.e. exceptionally long peaks that may “dilute” the motif signal) or irregular lengths distributions resulting from problems during the peak calling procedure. Such indications may lead to redo the pre-processing in order to refine the peaks (*e.g.* with PeakSplitter [9]) before using *peak-motifs*.
3. **Sequence composition.** Nucleotide and dinucleotide compositions are computed and displayed in the form of heat maps and positional profiles (**Box 1**).
4. **Motif discovery.** The workflow combines four word-counting algorithms relying on two complementary criteria (over-representation and positional bias) to detect exceptional words (oligonucleotides) and spaced pairs of words (dyads) (**Box 2**). Significant words are used as seeds to build probabilistic description of motifs (position-specific scoring matrices), indicating residue variability at each position of the motif.
5. **Comparisons with motif databases**. Discovered motifs are compared to one or several public databases of annotated motifs, to predict associated transcription factors. Comparison results are displayed as multiple motif alignments to highlight matches with several annotated motifs (*e.g.* factors belonging to the same family, composite motifs bound by protein complexes).
6. **Reference motif(s).** In some cases, ChIP-seq experiments are done with transcription factors for which some binding motifs have already been characterized and annotated in specialized databases [10-12]. Even in such cases, it is always interesting to discover motifs in peak sets, for

several reasons.

a. Motifs discovered from large peak collections are generally more robust than those annotated from a handful of binding sites, leading to a significant refinement of their predictive power [13].

b. The discovery can result in multiple motifs corresponding to transcription factors interacting with the targeted factors [14].

Users can enter one or several reference motifs (*i.e.* motifs expected to be found in the result) to ease the evaluation of the motif discovery results.

7. **Motif similarity networks.** All discovered motifs and their matches with reference or database motifs are displayed in the form of networks, where nodes represent motifs and edges their similarity. This visualization enables to grasp the groups of similar motifs returned by the different algorithms.
8. **Binding site predictions.** Sequences are scanned with the discovered motifs to predict binding sites. Peaks and predicted sites can be uploaded in the UCSC genome browser [15], in order to visualize them in their genomic context.

Main advantages of *peak-motifs*

1. **Time efficiency.** The processing time of the word-counting algorithms increases linearly with sequence size, whereas the complexity of most other algorithm is quadratic or worse. Our benchmarking shows that peak-motifs is able to treat peak sets of several tens of Mb in a few minutes on a personal computer.
2. **User-friendliness.** Whereas each component of *peak-motifs* can be used as a separate tool of the RSAT suite, their organization within the pipeline makes them available for non-experts, with a user-friendly interface and pre-selected parameters suited for analyzing ChIP-seq sequences. Results are reported as a summary Web page with expandable sections and links to the detailed results of each analysis step.
3. **Multiple motif detection.** The detection of multiple motifs provides clues about composite motifs and potential cofactors.
4. **Reliability**. The significance tests underlying pattern detection ensure a control of the rate of false positives, with suitable multi-testing corrections.
5. **Motif comparisons**. Discovered motifs can be compared with user-specified reference motifs (the motifs expected to bound the pulled down factor), but also with several public motif databases.
6. **Automation.** All the operations can be readily integrated in automatic workflows either as stand-alone applications, or as Web services invoked from a remote client via a SOAP/WSDL programmatic interface.

Main limitations

1. The workflow combines many analytic steps, each assorted with some parameters that may

strongly affect the outcome. Tuning those parameters and interpreting the results requires some experience, if one wants to go beyond the superficial analysis of motif logos and predicted site maps, in order to fully exploit the richness of the results. The goal of this protocol is precisely to guide users about the choice of parameters and the interpretation of the results.

2. The output presents motifs in a somewhat redundant form, since the same motif can be discovered by multiple algorithms. We however chose to maintain this partly redundant presentation because detecting a motif by several independent programs indicates the robustness of the result, and can show that this motif is both over-represented (*oligo-analysis*, *dyad-analysis*) and concentrated in the center of the peaks (*position-analysis*, *local-word-analysis*).

3. The motif discovery algorithms consider all peak regions as equivalent, and cannot take into consideration the actual peak shape. Such information provided as a coverage file can be taken into account by other programs such as ChIPMunk[16].

Study case

To illustrate this protocol, we use a ChIP-seq dataset obtained by pulling down the transcription factor Krüppel in 2-3 hours old embryos of *Drosophila melanogaster* [17]. Krüppel is the product of a gap gene, which plays a central role in antero-posterior patterning during early embryogenesis. Importantly, the products of gap genes and maternal factors such as Bicoid and Hunchback are known to bind neighboring sites on the genome, within regulatory regions or enhancers driving precise spatio-temporal patterns.

This study case will illustrate how motif discovery can identify the motif corresponding to the targeted transcription factor, but also highlight potential co-factors. The starting point of our procedure is a set of ~6,000 peak coordinates returned by the peak-calling algorithm MACS with a p-value threshold of 1e-5. Because peak-calling artifacts can generate very long regions that are unlikely to correspond to single binding sites, longer peaks are truncated to a maximum of 2000 bp, as described in the step 2 of the PROCEDURE section.

Although this study case is based on ChIP-seq peaks, this protocol would also be suitable to analyze other data types, such as ChIP-on-chip, CLIP-seq or sets of promoter sequences centered on the TSS (e.g. +/- 250bp around TSS).

MATERIALS

Equipment

- A computer connected to the Internet and a web browser.
- A collection of peak sequences of interest, hereafter called "test sequences" (in FASTA format). See EQUIPMENT SETUP for an example of FASTA file.
- Optional: a collection of “control sequences” (in FASTA format), for differential analysis.
- Optional: one or more position-specific scoring matrices (PSSMs) representing already known reference motifs, against which discovered motifs should be compared (see EQUIPMENT

SETUP).

- Optional: a custom motif database against which the results should be compared (see EQUIPMENT SETUP).

Equipment setup

Peak collections

Peak sequences can be produced by custom experiments, or obtained from publicly available datasets. Such datasets can be obtained for example from UCSC[15] (http://genome.ucsc.edu/), GEO [18] (http://www.ncbi.nlm.nih.gov/geo/) or Galaxy[19] (http://main.g2.bx.psu.edu/).

Download the peak sequences for testing this protocol

The coordinates of the peak sequences used to illustrate this protocol were retrieved from the GEO database (http://www.ncbi.nlm.nih.gov/geo/, entry GSM511084) in BED format, and uploaded in Galaxy to retrieve the corresponding sequences in FASTA format. In the Galaxy export, the header of each sequence indicates its genomic coordinates, which are parsed by peak-sequences in order to enable uploading the sites on the UCSC genome browser. As an example the header ">dm3_chr2L_26210_29479_+" indicates a region comprised between positions 26,210 and 29,479 on the forward (+) strand of the left arm of the second chromosome (chr2L) of *Drosophila melanogaster*.

> **CAUTION:** some peak calling programs such as QuEST export the position of peak centers instead of their left and right limits. In such case, peak coordinates have to be extended by adding a fixed interval (e.g. +/-200bp) around each peak center.

The FASTA sequence file exported by Galaxy can be downloaded from the supporting Web site:

http://rsat.ulb.ac.be/~rsat/suppmat_peak-motifs_Protocol/data/sequences/peak_sequences/Kr_D.mel_E01-03h_Eisen_rep1.fasta

This file is quite large (9.2Mb) and should be downloaded on your computer before running the protocol.

Reference matrices* and *custom motif databases (optional)

Reference matrices correspond to the motifs previously known for the studied protein, which are expected to be found by peak-motifs in the test sequences. For the study case, we will use two reference motifs representing the binding specificity of Kruppel, obtained from JASPAR and FlyReg, respectively. These reference matrices can be downloaded from the supporting Web site.

http://rsat.ulb.ac.be/~rsat/suppmat_peak-motifs_Protocol/data/matrices/Kr_JASPAR_FlyReg.tf

In addition, discovered motifs can be matched against a user-loaded custom motif collection (e.g. a set of user-collected motifs, or a licensed database). The Web interface requires for these user-loaded motifs to be provided as TRANSFAC-formatted files. This format was chosen because its syntax permits to document matrices with detailed information (ID, description, bound factor, site sequences, etc.). Matrices coming from other sources can be converted from a wide variety of formats ((JASPAR, MEME, MotifSampler, AlignACE, ClusterBuster, etc.) to the TRANSFAC format with the RSAT tool

convert-matrix.

PROCEDURE

Access the *peak-motif* web form

1. Open a connection to the RSAT Web server (http://rsat.ulb.ac.be/rsat/). Depending on your geographic location, use one of the mirrors available from the main page. The left menu bar provides access to the various RSAT programs. In this menu, click on the title `NGS-ChIP-seq`, and select the tool `peak-motifs`. This will open the *peak-motifs* form (Figure 4).

Sequences

2. Specify a title for this analysis in the `title` field of the top panel 'Peak sequences'. For the study case, type 'Kr D.mel 1-3h Markov-2'. When performing a differential analysis using two datasets, the title can be 'treatment_vs_control', or 'factorX_vs_factorY', which will help you to remember which datasets were given as input.
3. In the left side of the panel, under `peak sequences`, click on the `browse` button and select the file containing the test sequences. Peak sequence(s) is the only mandatory option to run *peak-motifs* with default parameters. You can optionally perform a differential analysis by selecting a second sequence file with the `browse` button of the right side of the panel, under `control sequences`. For the study case, locate on your computer the downloaded file containing the Krüppel peak sequences (see section *Equipment setup*).

   CAUTION. We strongly advise to use the `browse` button to upload your file, rather than pasting the sequences in the box. The web browser will freeze or crash if thousands of peak sequences are pasted in the box.

   CRITICAL STEP. The test (and optionally control) sequence files(s) should contain sequences of reasonably well-defined peaks. We explain hereby three traps to avoid.

   a) Make sure that the sequence file contains **peak sequences** and **not the raw reads**. A peak file should have a size in the range of several Mbytes whereas a read file with millions of reads has a size of hundreds Mbytes to few Gbytes. It is crucial to run peak-motifs on peak sequences, since the reads generally correspond to short fragments (typically 30bp) *on the left and on the right sides* of the actual binding sites. In addition, files containing several million reads are too large for online treatment. Files containing read sequences should first be treated with a read mapping program (e.g. Bowtie [20]) that will align the reads on the reference genome. The resulting mapped reads should be processed with a peak-calling program (e.g. MACS [21]) to obtain the peak coordinates, and the corresponding sequences can be retrieved from specialized Web resources (UCSC, Galaxy).

   b) The program expects peak sequences (in FASTA format) and not peak coordinates

(BED files). See Table 1 for more information of how to obtain sequences from a coordinate file.

c) Depending on the peak-calling program used, peaks may span several hundreds to thousands base pairs. Long peaks often result from the merging of a series of neighboring peaks. In this case, *peak-motifs* will perform better if these peaks are refined into individual peaks, for example with PeakSplitter[9]. This will increase the performance of *position-analysis* and *local-word-analysis*, as both algorithms search for motifs with positional biases, which are diluted when the peak regions are too broad.

TROUBLESHOOTING (see Table 1)

4. *Peak-motifs* offers the possibility to easily reduce the analyzed dataset, by focusing on a given number of top sequences, and/or by trimming the sequences to a desired length around peak centers. Click on the triangle on the right of 'Reduce peak sequences' to expand the hidden panel (Figure 5, top panel). For the study case, the option `Number of top sequences to retain` is left blank, as we will use the full-size set of sequences.

5. The option `Cut peak sequences` restricts the analysis to the most central region of each peak (e.g. peak center +/- 200bp), supposed to be dense in binding sites. This assumption is nevertheless highly dependent on the peak-calling program, and on whether the peak centers actually correspond to their summits (where the binding site is supposed to be found). We thus advise to generally leave this option blank, or run both the restricted and complete analyses, and compare the results. However in the Krüppel dataset, peaks have variable sizes up to 13,205bp, which probably reflects a problem with the peak-calling procedure rather than the natural extension of Krüppel binding regions. For this particular study case, restrict the analysis to 1000bp around the center of each peak.

   Note that input sequences are automatically purged to discard redundant fragments (peak overlaps, duplications). Motif discovery is performed on the purged sequences, but the sequence scanning is done on the non-purged sequences, in order to locate all the putative binding sites.

Motif discovery parameters

6. Click on the triangle on the right of 'Change motif discovery parameters' to expand the motif-discovery option panel (Figure 5, bottom panel). For the study case, keep the options *oligo-analysis* and *position-analysis* checked (the other algorithms may be checked for a full analysis, but this takes more time). Check the values 6 and 7 for `'Oligomer length'`, in order to detect significant hexanucleotides and heptanucleotides.

7. The background model must be specified when analyzing a single set of peaks. In the case of differential analysis, the second set of peaks (control set) will serve as background to estimate the random expectation of each oligonucleotide. In single-dataset mode, the background model is built from the test sequences based on frequencies of smaller words. For the study case, select the most stringent background model (`oligo length -2`).

CRITICAL STEP For single datasets, the background model must be chosen carefully, as this parameter strongly affects the results. **Box 1** explains how the background model is calculated from the input sequences. The order of the Markov model should be adapted to the sequence size: we recommend low-order models (*m=1*) to increase the sensitivity for small datasets (a few hundred kb), and higher order models (*m=k-2*, where *k* is the oligonucleotide length) to increase the specificity for large sequence sets (>= 1Mb).

TROUBLESHOOTING (see Table 1)

Comparisons of discovered motifs with motif databases

8. Click on the triangle on the right of 'Compare discovered motifs with databases' to reveal the 'Compare motifs' panel. This section displays a list of public motif databases, such as JASPAR[10], that are directly supported by *peak-motifs*. Each discovered motif will be compared to the selected collection(s) of motifs, in order to identify which transcription factors may correspond to these binding motifs, or to pinpoint the currently unknown motifs. Motif databases should be chosen according to the studied organism. For the study case, deselect the default database ('JASPAR core Vertebrates') and select all the databases related to drosophila ('JASPAR core Insects', 'Drosophila DMMPMM' and 'Drosophila IDMMPMM'), as illustrated in Figure 6.

   CAUTION: Due to limitations in annotating resources, motif databases are very incomplete and should not be considered as comprehensive knowledge repositories. We strongly encourage users working on a specific factor to independently search the literature for reference motifs, and provide these to *peak-motifs* as reference motifs (see Step 9).

9. If you dispose of your own motif collections (*e.g.* licensed databases, custom matrices), ensure that they are formatted as TRANSFAC files (if not, use the tool *convert-matrices* on the RSAT Web site) and upload the files by clicking on the `browse` button of the section 'Add your own motif database'. A title should be specified for this custom database in the field on the left on the 'browse' button. For the study case, we will only use the public databases available on RSAT, so this option will be left blank.

10. One or several reference motifs can be uploaded (as a TRANSFAC-formatted file) by clicking on the browse button in the section 'Add known reference motifs for this experiment'. For the study case, use the file `Kr_JASPAR_FlyReg.tf`.

Search for binding sites and export as UCSC custom track

11. Click on the triangle on the right of 'Locate motifs and export as UCSC custom track' to expand the panel with the options for searching putative binding sites in the peak sequences (Figure 6, bottom). Check the box 'Search putative binding sites in the peak sequences'.

12. Optionally, the sites can be exported as a custom track (BED file) that can be uploaded in the UCSC genome browser[15] or Ensembl[22], to visualize these putative binding sites in their annotated genomic environment. By default, this very helpful way to interpret the results is disabled, as it requires information in addition to peak sequences (genome assembly version and coordinates of

the peaks). The required information can be provided in either of two ways: (i) if your sequences have been fetched from Galaxy, check the radio button 'peak coordinates specified in the headers of the sequence file'; (ii) check the radio button ' Peak coordinates provided in a separate bed file', specify the file with the 'Choose file' button, and indicate the 'Assembly version (UCSC)'. For the study case, sequences were previously downloaded from the Galaxy server.

Submit the form

13. Check the 'email output' option and provide your email address, to be notified when the results are ready. Alternatively, you can keep the display output to obtain the results directly in the web browser. The email output is generally preferred for large datasets or when results are compared to many motif collections, because the whole processing can take a few tens of minutes.
14. Click on the GO button to run the analysis.

Viewing the results

15. A new page appears in place of the form, indicating that the task has been submitted to the server. A link to the results is displayed; click on this link to follow the analysis.
16. Results are displayed on this page progressively, so that it is possible to start analyzing the results before completing the whole analysis. Regularly refresh the page to update it. When the whole analysis is finished, the top of the page displays a summary of the results instead of the "Status: running" message.

TIME TAKEN

The processing time depends on the server load (the number of jobs currently running on the server), on the selected tasks, and the sequence size. For the 8Mb peak sets of the Krüppel case study, the complete analysis (sequence composition, motif discovery, motif comparisons, sequence scanning) took 27 minutes from the job submission to the reception of the completion email.

TROUBLESHOOTING

See Table 1.

ANTICIPATED RESULTS

The result of peak-motif is presented as a synthetic report with clickable links to the detailed result files. To ease the interpretation of the results, the report is organized in thematic sections as presented hereafter.

Sequence length distribution and composition

The first section of the report (Figure 7) displays the length distribution and composition of the peak sequences. The distribution of sequence lengths gives some hints about the pre-processing (peak-calling). In the study case, the original peaks ranged from 506bp to 13,205bp, but were clipped to a maximal size of 2kb (1,000bp on each side of each peak center). The mean length of the clipped peaks is still 1,347bp, which exceeds by far the length of a single binding site, or even the lengths of typical drosophila enhancers. The nucleotide and dinucleotide compositions displayed on Figure 7 are typical of drosophila non-coding sequences (**Box 1**). The positional profiles show the depletion of some nucleotides (A and T) and dinucleotides (AA, TT, AT, TA) at the center of the peaks, suggesting a general avoidance of A/T-rich sequences in the Krüppel sites.

Reference motifs

For the study case, we use as reference two Krüppel binding motifs extracted from JASPAR [10] and FlyReg[23], respectively. Logos are displayed in both direct and reverse complementary orientations (Figure 8). The colored consensus sequences are shown above the logos, and can be searched for in the html output using the browser search function.

Discovered motifs (by algorithm)

Figure 9 shows the discovered motifs, grouped by algorithm. Each motif is represented by its direct and reverse logo, its colored consensus and its significance. In the present case, several highly significant motifs have been found, among one strongly similar to the canonical Krüppel motif. This view enables a comparison between the outputs of different algorithms, in order to spot motifs found by several or a single algorithm. On the right hand side of this section, various links provide access to the primary results: the list of identified words, their assembly into longer motifs, and the resulting matrices in various formats.

Discovered motifs (with motif comparison)

Figure 10 displays the same discovered motifs as in the previous section, documented with a comparison with motifs provided as reference or uploaded as additional database. For each database, the first 3 best matches are displayed (additional matches can be accessed by clicking the links 'match table' and 'alignment logos' on the right). The table summarizes information about the alignments: percentage of motifs aligned, Pearson correlation and normalized Pearson correlation. One should be aware that a high correlation coefficient can be misleading if it only concerns a fraction of the matrices. The goal of the normalized correlation is to avoid this effect by weighting the correlation according to

the mutual coverage of the motifs. The colored consensus indicates the aligned parts of each motif.

In this study case, the Krüppel motif is only detected by its positional bias, but escapes detection by *oligo-analysis*. We interpret this as a consequence of the very large size of the peaks obtained from the GEO database. Consistently, when peaks are trimmed to the 200 central-most base pairs, the Krüppel motif appears as significantly over-represented (it is detected by *oligo-analysis*) but its positional bias is not detected anymore, because sites were found on the whole range of the trimmed peaks.

In other cases – e.g. mouse samples proposed as DEMO on the Web site – the correct motifs are consistently detected by several motif discovery approaches, indicating their robustness (Thomas-Chollier *et al.*, submitted). For this protocol, we deliberately chose a more difficult example to illustrate the variety of the cases, and to demonstrate the importance of combining different statistical criteria (global over-representation, position bias, local over-representation) in order to increase the sensitivity of motif discovery.

Motif comparisons with multiple logo alignments

A more detailed view of the alignments is obtained by clicking on the html link on the right hand side (Alignments (logos)), which display a HTML with "one-to-n" alignments between one discovered motif and one or several database motifs (Figure 11). It is very advisable to check the relevance of the results using this view, as the human eye turns out to be more accurate in detecting similarities than any measure. Several similarity metrics are indicated in this view. Besides the similarity score, we provide, for each measure, the rank of the motif with respect to this measure. An integrated rank which combines the ranks for all measures is also provided.

Predicted sites and enrichment profiles

The bottom of the per-motif summary indicates the positions of predicted sites (left) and enrichment of peaks (right). The spatial distribution of predicted sites can be very informative for some motifs (Figure 10). In our study case, the various motifs have very different profiles: the motif corresponding to Krüppel (positions_6nt_m2) presents a very sharp peak whose maximum coincides with the peak centers, and a deviation of about +/- 100 bp. This nicely confirms that the peak centers are indeed enriched in Krüppel binding sites, since Krüppel was targeted in the experiment.

Some other motifs show a "volcano-like" profile, with a high enrichment on each of the peaks center. These might correspond to transcription factors that are co-factors of Krüppel. For example, the second motif displayed on Figure 10 (oligos_7nt_mkv5_m1) possibly corresponds to hunchback, a transcription factor known to interact with Kr in early embryogenesis. This motif was detected by *oligo-analysis* as strongly over-represented, yet its positional profile shows a strong avoidance in the middle of the peaks, where the Krüppel binding sites show the highest concentration. Altogether, these observations suggest that the peaks contain a high concentration of well-centered Krüppel sites flanked by Hunchbak-like motifs.

By clicking on the UCSC button on the right side of the position profiles, predicted sites can be uploaded to the UCSC genome browser in order to visualize them in their context and compare them

the annotation tracks (Figure 12).

In summary, despite the large dispersion of peak sizes in the initial dataset, peak-motif is able to identify the correct motif, to reveal some heterogeneity in AT composition at the center of the peaks, with a consistent avoidance of several AT-rich motifs, one of which may correspond to the hunchback co-factor.

REFERENCES

1. G. Robertson, M. Hirst, M. Bainbridge et al., *Nat Methods* **4** (8), 651 (2007).
2. D. S. Johnson, A. Mortazavi, R. M. Myers et al., *Science* **316** (5830), 1497 (2007).
3. N. Rusk, *Nat Methods* **6** (11 Suppl), S1 (2009).
4. J. D. McPherson, *Nat Methods* **6** (11 Suppl), S2 (2009).
5. S. Pepke, B. Wold, and A. Mortazavi, *Nat Methods* **6** (11 Suppl), S22 (2009).
6. V. Boeva, D. Surdez, N. Guillon et al., *Nucleic Acids Res* (2010).
7. P. Machanick and T. L. Bailey, *Bioinformatics* (2011).
8. T. L. Bailey, *Bioinformatics* (2011).
9. M. Salmon-Divon, H. Dvinge, K. Tammoja et al., *BMC Bioinformatics* **11**, 415 (2010).
10. E. Portales-Casamar, S. Thongjuea, A. T. Kwon et al., *Nucleic Acids Res* **38** (Database issue), D105 (2010).
11. E. Wingender, *Brief Bioinform* **9** (4), 326 (2008).
12. S. Gama-Castro, H. Salgado, M. Peralta-Gil et al., *Nucleic Acids Res* **39** (Database issue), D98 (2011).
13. A. Medina-Rivera, C. Abreu-Goodger, M. Thomas-Chollier et al., *Nucleic Acids Res* **39** (3), 808 (2011).
14. X. Chen, H. Xu, P. Yuan et al., *Cell* **133** (6), 1106 (2008).
15. P. A. Fujita, B. Rhead, A. S. Zweig et al., *Nucleic Acids Res* **39** (Database issue), D876 (2011).
16. I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov et al., *Bioinformatics* **26** (20), 2622 (2010).
17. R. K. Bradley, X. Y. Li, C. Trapnell et al., *PLoS Biol* **8** (3), e1000343 (2010).
18. T. Barrett, D. B. Troup, S. E. Wilhite et al., *Nucleic Acids Res* **39** (Database issue), D1005 (2011).
19. J. Goecks, A. Nekrutenko, and J. Taylor, *Genome Biol* **11** (8), R86 (2010).
20. B. Langmead, C. Trapnell, M. Pop et al., *Genome Biol* **10** (3), R25 (2009).
21. Y. Zhang, T. Liu, C. A. Meyer et al., *Genome Biol* **9** (9), R137 (2008).
22. P. Flicek, M. R. Amode, D. Barrell et al., *Nucleic Acids Res* **39** (Database issue), D800 (2011).
23. C. M. Bergman, J. W. Carlson, and S. E. Celniker, *Bioinformatics* **21** (8), 1747 (2005).
24. J. van Helden, B. Andre, and J. Collado-Vides, *J Mol Biol* **281** (5), 827 (1998).
25. J. van Helden, A. F. Rios, and J. Collado-Vides, *Nucleic Acids Res* **28** (8), 1808 (2000).
26. J. van Helden, M. del Olmo, and J. E. Perez-Ortin, *Nucleic Acids Res* **28** (4), 1000 (2000).
27. M. Thomas-Chollier, M. Defrance, A. Medina-Rivera et al., *Nucleic Acids Res* **Web software issue 2011**, in press (2011).
28. M. Defrance, R. Janky, O. Sand et al., *Nat Protoc* **3** (10), 1589 (2008).
29. J. V. Turatsinze, M. Thomas-Chollier, M. Defrance et al., *Nat Protoc* **3** (10), 1578 (2008).

BOXES

Box 1. Sequence composition and background models.

The choice of an appropriate background model is one of the most important criteria for predicting cis-regulatory elements. The analysis of sequence compositions in nucleotides and oligonucleotide provides useful hints for the choice of this model. Figure 3 displays the sequence compositions of two collections of peak sequences obtained by ChIP-seq with two orthologous proteins (drosophila CBP and mouse p300) that act as co-factors by interacting with multiple transcription factors. The heatmaps indicate the probability to observe a given residue (suffix, displayed in columns) following another residue (prefix, displayed in rows). The drosophila heatmap shows typical aggregative tendency of As and Ts: there is a much higher probability to observe an A after another A (33%) than after a T (18%). A striking feature of the mouse heatmap is the avoidance of CpG dinucleotides, typical of mammalian sequences: the probability of observing a G after a C is only 8%, whereas it is 30% after any other residue. The composition not only depends on the organism but also on the sequence type (promoters, introns, coding exons, etc), and on local particularities of the sequences. For example, the positional profiles of dinucleotide occurrences further show a specific depletion of AA, TT, TA and AT in the centers of the drosophila CBP peaks.

Such dependencies have an important impact on the computational analysis of cis-regulatory elements: the probability of a given site will strongly differ depending on the genomic context in which it is found. For instance, for the sequence ATCGCGAT, the probability estimated from dinucleotide composition is 1.4926e-05 in drosophila CBP peaks, and 9.9424e-07 in mouse p300 peaks. The same sequence is thus expected to occur by chance once every 66Kb in drosophila p300 peaks (1/1/1.4926e-05=66,997), versus once per Mb in mouse CBP peaks (1/9.9424e-07= 1,005,793).

The program *peak-motifs* automatically computes sequence composition for words of various sizes in order to estimate the background probabilities. Background models based on simple nucleotide composition (called Bernoulli models) are not suited, since they fail to capture dependencies between adjacent nucleotides. Markov models of order 1 take such dependencies into account by estimating the probability of each residue depending on the preceding nucleotide (Figure 3). By extension, a Markov model of order *m* can be built by computing the probabilities of each residue as a function of the *m* preceding residues. The program *oligo-analysis* uses Markov models of higher order to estimate the expected frequency of longer oligonucleotides (e.g. hexanucleotides) on the basis of the frequencies of shorter words (e.g. tetranucleotides). Higher-order background models are more stringent, and return less false positives but can result in a loss of sensitivity for small data sets.

Box 2. Motif discovery algorithms

Peak-motifs combines several previously described motif discovery algorithms that detect exceptional words on the basis of distinct criteria: global over-representation of words (*oligo-analysis*) or spaced word pairs (*dyad-analysis*), local over-representation of words in positional windows (*local-word-analysis*) or heterogeneity of the word count distribution along the peak sequences (*position-analysis*). A great advantage of those word-based algorithms is their low memory requirements and their linear

time complexity regarding the data set size (computing time increases linearly with the peak sequences size). Significant words are aligned to build "position-specific scoring matrices", which can be used to scan sequences for predicting sites.

Global over-representation of words (oligo-analysis)

The program *oligo-analysis*[24] (Figure 2a) counts the number of occurrences of each oligonucleotide ("word") of a given length (typically 6 or 7 nucleotides) in the test set ("observed occurrences"), and compares it with the number of occurrences that would be expected by chance, according to a given background model. By default, background models for motif discovery are estimated from the oligonucleotide composition of the test sequences. Optionally, a second sequence set ("control sequences") can be entered to estimate the random expectation of each oligonucleotide and dyad. The statistical significance of each word is estimated based on the binomial distribution.

Global over-representation of spaced pair words (dyad-analysis)

Spaced motifs are characteristic of some classes of transcription factors that bind DNA in the form of homo- or heterodimers. The program *dyad-analysis*[25] extends the principle of *oligo-analysis*, by counting the number of occurrences of pairs of trinucleotides separated by a spacing of fixed width but variable content. The program applies the binomial test to estimate the over-representation of each pair of trinucleotides with all possible spacing values from 0 to 20.

Positional biases (position-analysis)

The program *position-analysis*[26] (Figure 2b) detects exceptional words based on their positional biases, i.e. non-homogeneous distribution relative to some reference point. For the analysis of peaks, positions are computed relative to peak centers. For other applications, reference positions can be the right extremity of the sequence (e.g. to detect upstream transcriptional signals in upstream sequences of genes) or the left extremity (e.g. for the analysis of 3' untranslated regions).

The program counts the observed number of occurrences of each oligonucleotide in non-overlapping windows and compares it to the count that would be expected from a homogeneous repartition. Since the peaks can have variable lengths, the homogenous distribution is generally non-flat: expected occurrences typically decrease on both sides with increasing distances from peak centers (Figure 2b, green curve). The significance of the difference between the observed and the homogeneous distributions is estimated with a chi-squared test.

Local over-representation (local-words)

The program *local-words*[27] somehow combines the principles of *position-analysis* and *oligo-analysis*: it applies the binomial test to detect over-represented words in positional windows of variable or fixed size. (Figure 2c)

Statistical significance of exceptional words

All the above programs return lists of words associated with a P-value (binomial or chi-squared depending on the program). The P-value represents the nominal risk of false positive, i.e. the

probability for one particular word to show a given level of over-representation or positional bias by chance, according to the background model. Since the significance test is applied on several thousands of words, a multi-testing correction is applied by converting the P-value to an E-value (E-value = P-value * number tested words) representing the expected number of false positives. This E-value is in turn converted to a significance index *sig=-log10(E-value)*, providing an intuitive feeling of the reliability of the result (the higher the better).

Building matrices from lists of words

Each of the word-based motif discovery algorithms described above returns a set of exceptional words (oligonucleotides or dyads) sorted by significance. This list generally includes groups of mutually overlapping words, which reveal shifted fragments of the same motif. Those words can be used as seeds to collect putative sites, which are then aligned to build a position-specific scoring matrix. The final result of the motif discovery is a set of such PSSM, which can be used to scan sequences for predicting binding sites. See our previous protocols for the principle of matrix building from words [28] and sequence scanning with matrices[29].

TABLES

Table 1. Troubleshooting

Step	Symptom	Cause	Solution
33	The peak files does not contain any sequence, but only genomic coordinates.	Peak calling programs often returns the genomic coordinates (generally in "BED" format), but not the sequences directly.	Use these genomic coordinates to retrieve the sequences from the Galaxy server (http://main.g2.bx.psu.edu/). We provide a step-by-step explanation at the bottom of the `Peak sequences` panel, through the link '*I only have coordinates in a BED file, how to get sequences ?*'
77	The motif discovery programs return not a single motif, or only weakly significant motifs.	The order of the Markov model may be too high for the sequence size.	Check the total sequence size in the 'Sequence Composition' box of the result page: if the total size is smaller than 1Mb, reduce the order of the Markov model.

LEGENDS TO FIGURES

Figure 1. **Flow chart of the *peak-motifs* workflow**.

Figure 2. **Motif discovery approaches**. Schematic representation of the criteria for detecting exceptional words. (a) Over-representation of words (*oligo-analysis*). Right: schematic view of the principle underlying the test of over-representation for a given word. Left: occurrences observed for each word in the test set (Y axis) are compared to the occurrences expected according to the background model (X axis). Each dot represents a hexanucleotide. Significant words are highlighted in red (binomial sig >= 10). (b) Positional bias (*position-analysis*). Left: sequences are aligned relative to the peak centers, and the occurrences of each word are counted in non-overlapping windows of fixed width. Right: positional distribution of word occurrences that would be expected under a homogeneous distribution. Since peak sequences have varying widths, the number of sequences decreases with distance to peak centers, and the expected occurrences (green curve) decrease accordingly. (c) Local over-representation (*local-word-analysis*): word occurrences are counted in windows of increasing widths centered on peak centers (left), and compared to the occurrences that would be expected under a homogeneous model (right). In this example (Sox2 peaks from [14]), a 50bp window contains strongly over-represented words (sig >= 5, highlighted in red), corresponding to different fragments of the Sox2 binding motif.

Figure 3. **Dinucleotide composition and derived background models**. Peaks bound by (a) Drosophila transcriptional co-regulator CBP (a) and its mouse ortholog p300 (b), respectively. The left heatmap represents transition frequencies between prefix (rows) and suffix (columns) residues.

Figure 4. ***peak-motifs* web form**. By default, a simplified form is displayed. The four last sections can be expanded to display the parameters for each analytic step.

Figure 5. **Input sequences treatment (top) and motif discovery (bottom) options.** An essential parameter is the choice of the background model, whose stringency should be adapted to the sequence size. The display shows the way to fill-in these options for the Krûppel example.

Figure 6. **Options for motif comparisons (top) and predicted sites visualization (bottom).** The options are filled-in for the Krüppel peaks example.

Figure 7. **Sequence composition.** From top to bottom: distribution of peak lengths, nucleotide and di-nucleotide composition heatmaps (left) and position profiles (right).

Figure 8. **Reference motifs**. Reference motifs can be entered to indicate which motifs would be expected to be found (the "correct" answer). Note that reference motifs are not ignored during the motif discovery step (motif discovery is an "*ab initio*" process). Rather, they are used a posteriori for validating the discovered motifs.

Figure 9. **Discovered motifs grouped by algorithm.** Motifs discovered by *oligo-analysis* (top) and *position-analysis* (bottom). The motif identifier (first column) indicates the algorithm, oligonucleotide length and Markov model order (only for *oligo-analysis*). For each program the three best-scored motif

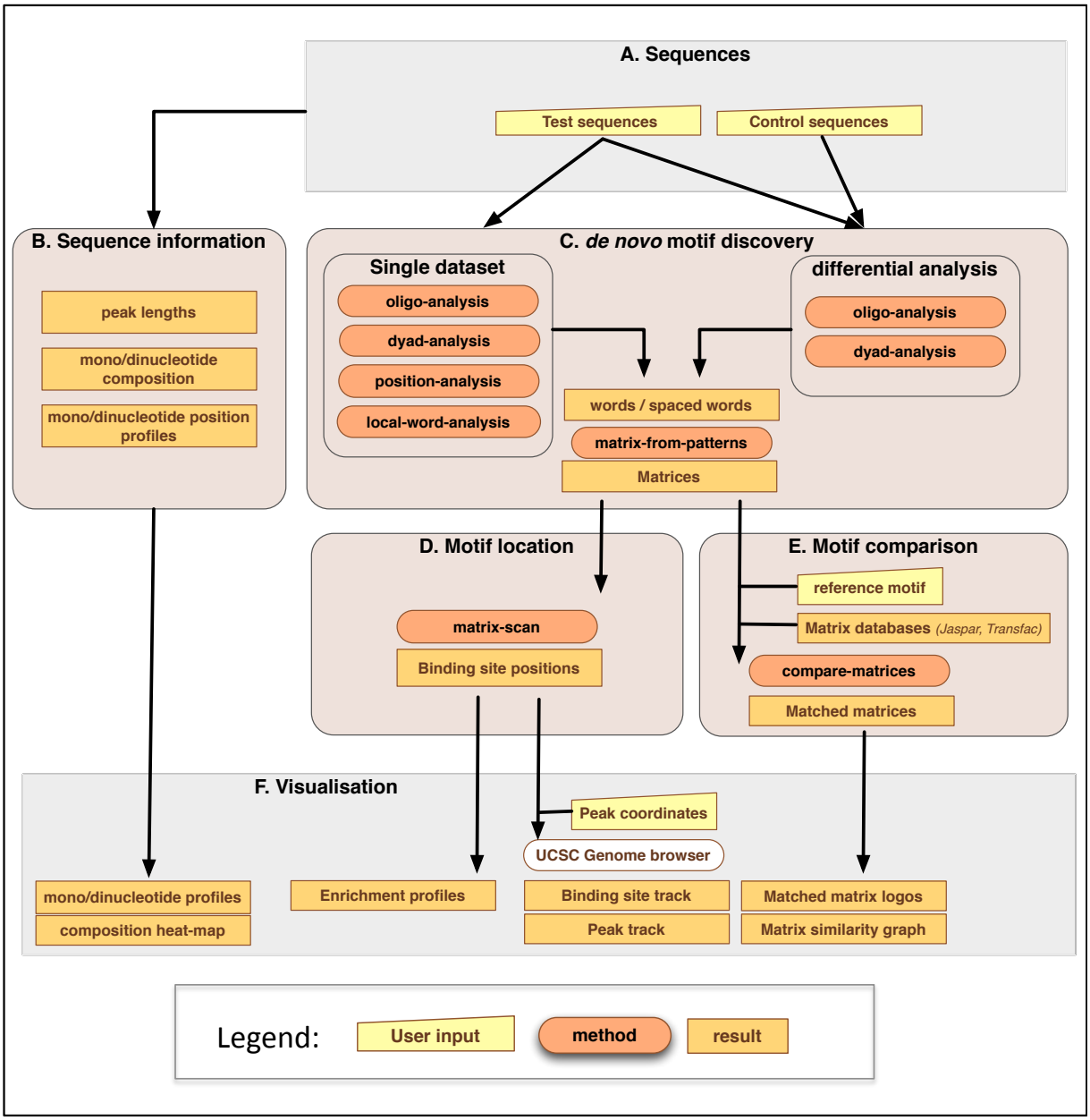
logos are displayed. The last column contains links toward intermediate results (over-represented words, assemblies of overlapping words and significance matrices) and matrices in TRANSFAC and tab formats.

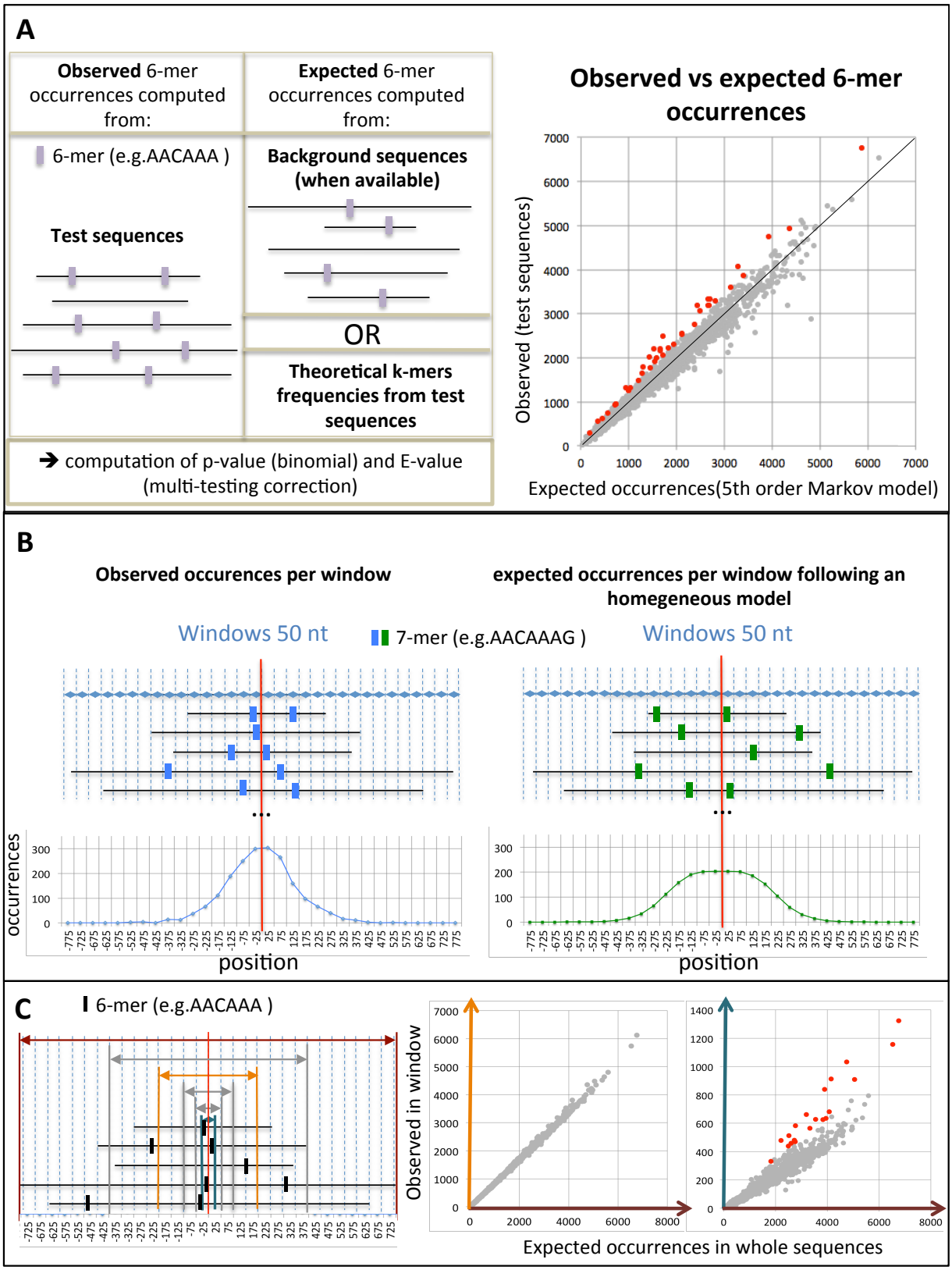
Figure 10. **Discovered motifs with motif comparisons.**

The snapshot displays the summary for two discovered motifs, returned by *position-analysis* (positions_6nt_m2) and *oligo-analysis* (oligos_7nt_mkv5_mA) respectively. The first row displays the logos (direct and reverse complementary) and colored consensus of the discovered motif, along with links to the text files. The next rows summarize the results of comparison between the discovered motif and those provided as reference or found in the selected databases. The three best matches of each comparison are summarized on the report, indicating the motif identifier and names, matching strand, number of aligned columns, and various similarity metrics. The last column displays colored consensuses restricted to the aligned positions (non-aligned positions are replaced by dots). The complete list of matches with the detailed matching statistics can be accessed by clicking the link below the comparison summary. Tables showing all the correlation statistics, count matrices and logos alignments are available through links in the left panel. The last row contains information about predicted sites. The plot 'Distribution of sites' shows the number of occurrences (Y axis) per position (X axis) along the centered peak sequences. The 'Enrichment in binding sites' plot indicates the binomial significance of the overrepresentation (Y axis) for all possible score values (X axis).

Figure 11. **Motif comparisons**. Logo alignments and scores of the matches between a motif discovered in Kruppel peaks and the insect section of the JASPAR database. The logo of the query motif (discovered motif) is aligned with the logos of multiple matching database motifs, in order to highlight partial correspondences. The table contains multiple similarity metrics: cor: Pearson correlation; ncor: normalized Pearson correlation; logoDP: dot product between the logo scores; NIcor: normalized correlation between information content values; SSD: Squared Sum of Deviations; NSW: normalized Sandelin-Wasserman. The next columns indicate the ranks of the previous columns (rcor: rank of the cor; rNcor: rank of Ncor, etc). The rank mean provides a robust measure of the overall similarity between two motifs.

Figure 12. **Predicted sites visualized in their genomics context on the UCSC genome browser**. Positions of Kruppel predicted sites appear in red, along with the corresponding peaks in green, in the light of relevant annotation tracks made available through the UCSC genome browser. Most of the information available for Drosophila has been generated by the ModENCODE consortium.


A. Sequences
Test sequences
Control sequences
B. Sequence information
peak lengths
mono/dinucleotide composition
mono/dinucleotide position profiles
C. de novo motif discovery
Single dataset
oligo-analysis
dyad-analysis
position-analysis
local-word-analysis
differential analysis
oligo-analysis
dyad-analysis
words / spaced words
matrix-from-patterns
Matrices
D. Motif location
matrix-scan
Binding site positions
E. Motif comparison
reference motif
Matrix databases (Jaspar, Transfac)
compare-matrices
Matched matrices
F. Visualisation
Peak coordinates
UCSC Genome browser
mono/dinucleotide profiles
composition heat-map
Enrichment profiles
Binding site track
Peak track
Matched matrix logos
Matrix similarity graph
Legend:
User input
method
result

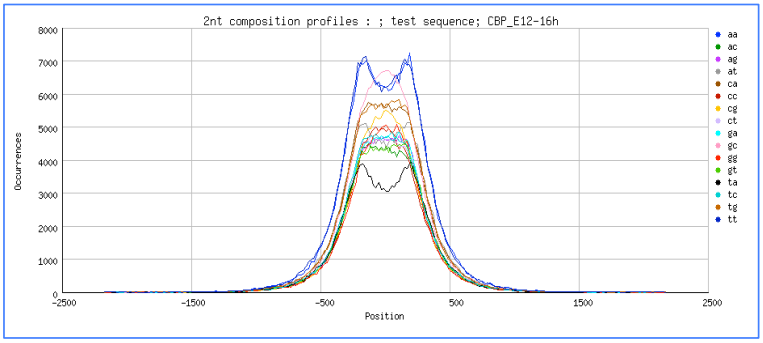

A
Observed 6-mer occurrences computed from:
Expected 6-mer occurrences computed from:
6-mer (e.g.AACAAA)
Test sequences
Background sequences (when available)
OR
Theoretical k-mers frequencies from test sequences
➔ computation of p-value (binomial) and E-value (multi-testing correction)
Observed vs expected 6-mer occurrences
Observed (test sequences)
Expected occurrences(5th order Markov model)
B
Observed occurences per window
expected occurrences per window following an homegeneous model
Windows 50 nt
7-mer (e.g.AACAAAG)
Windows 50 nt
occurrences
position
position
C
6-mer (e.g.AACAAA)
Observed in window
Expected occurrences in whole sequences

A. Drosophila transcriptional co-activator: CBP

Position profile

Transition frequencies

pr\suf	a	c	g	t	P_prefix
a	0.33409	0.20472	0.21368	0.24750	0.262
c	0.29181	0.22875	0.24375	0.23570	0.237
g	0.23796	0.30353	0.22994	0.22857	0.238
t	0.18549	0.21663	0.26421	0.33367	0.263

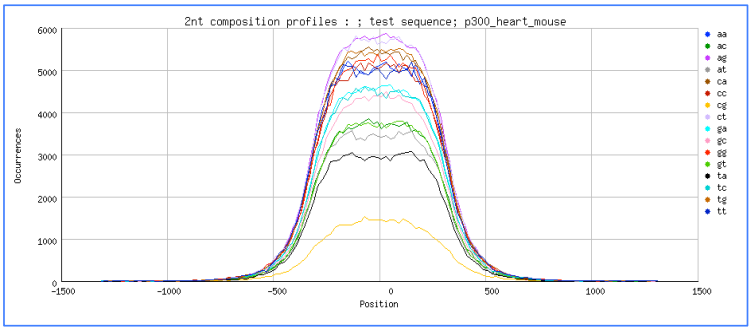


B. Mouse ortholog transcriptional co-activator: p300

Position profile

Transition frequencies

pr\suf	a	c	g	t	P_prefix
a	0.28516	0.20576	0.31378	0.19530	0.255
c	0.30680	0.28831	0.08174	0.32316	0.245
g	0.25648	0.24099	0.29096	0.21157	0.247
t	0.17107	0.24799	0.29850	0.28244	0.253



RSA-tools - peak-motifs

Pipeline for discovering motifs in massive ChIP-seq peak sequences.

Conception[c], implementation[i] and testing[t]: Jacques van Helden[cit] , Morgane Thomas-Chollier[cit] , Matthieu Defrance[ci] , Olivier Sand[i] , Denis Thieffry[ct] and Carl Herrmann[ct]

▶ **Information on the methods used in peak-motifs**

Peak Sequences

Title Kr D.mel 1-3h Markov -2

Peak sequences Paste your sequence in fasta format in the box below

Or select a file to upload (.gz compressed files supported)

Choisissez un fichier Kr_D.mel...p1.fasta

Optional: *control dataset for differential analysis (test vs control)*

Control sequences Paste your sequence in fasta format in the box below

Or select a file to upload (.gz compressed files supported)

Choisissez un fichier Aucun fichier choisi

(I only have coordinates in a BED file, how to get sequences ?)

▶ **Reduce input peak sequences**

▶ **Change motif discovery parameters**

▶ **Compare discovered motifs with databases (e.g. against Jaspar) or custom reference motifs**

▶ **Locate motifs and export as UCSC custom track**

Output ◉ display ○ email

Note: email output is preferred for very large datasets or many comparisons with motifs collections

▼**Reduce input peak sequences**

Restrict the input dataset

Number of top sequences to retain

Cut peak sequences: +/- 1000 **bp around the center of each peak**

▼**Change motif discovery parameters**

Discover motifs

Continuous words

- [x] **Discover over-represented words** [oligo-analysis]
- [ ] **Discover words with local over-representation** [local-word-analysis]
- [x] **Discover words with a positional biais** [position-analysis]

Oligomer length for the three programs above [x] 6 [x] 7

Note: motifs can be larger than word sizes (words are used as seed for building matrices)

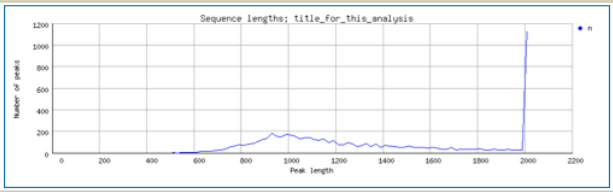
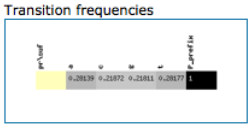
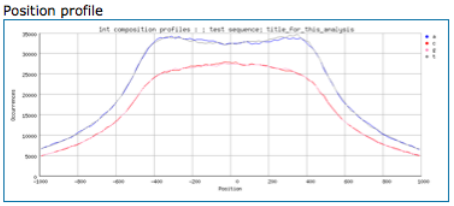
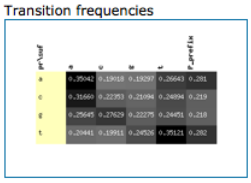
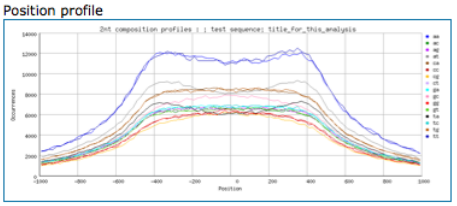
Spaced words pairs

- [ ] **Discover over-represented spaced word pairs** [dyad-analysis]

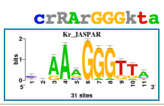
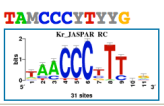
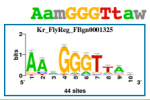
Markov order of the background model *(only for single-input analysis, will be ignored if control set is provided)*

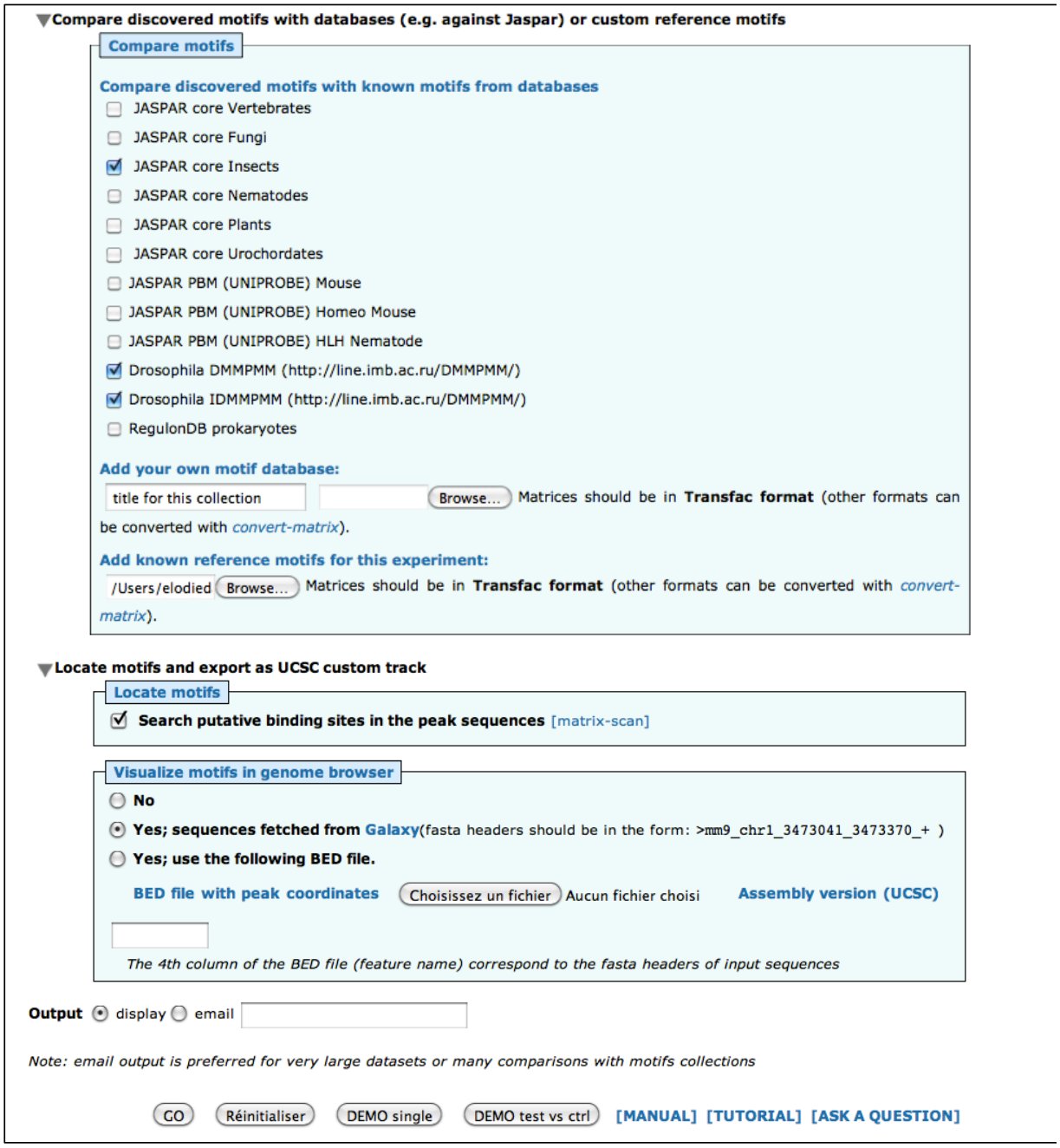
oligo length –2 (more stringent for large data sets e.g. > 1Mb)

0 (generally not ideal)
1 (more sensitive for small data sets, e.g. 100kb)
oligo length –3 (intermediate size data sets)
oligo length –2 (more stringent for large data sets e.g. > 1Mb)

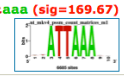
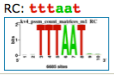
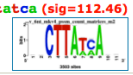
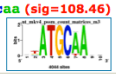
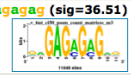
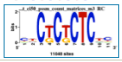
Sequence composition (test sequences)		
Nb of peaks: 6003 Total seq. size: 8087 kb Min length: 506 bp Mean length: 1347.13 bp Max length: 2000 bp		[seq: converted purged] [lengths: list distrib graph]
1nt composition	Transition frequencies  Position profile 	[1nt: freq transitions] [bg model: Inclusive format] [profile: table html(individual)]
2nt composition	Transition frequencies  Position profile 	[2nt: freq transitions] [bg model: Inclusive format] [profile: table html(individual)]

Reference motifs

Reference motif(s)		
Reference motif(s)	Kr_JASPAR Kr_JASPAR   Kr_FlyReg_FBgn0001325 Kr_FlyReg_FBgn0001325  	[transfac format] [tab format]


▼Compare discovered motifs with databases (e.g. against Jaspar) or custom reference motifs
Compare motifs
Compare discovered motifs with known motifs from databases
JASPAR core Vertebrates
JASPAR core Fungi
JASPAR core Insects
JASPAR core Nematodes
JASPAR core Plants
JASPAR core Urochordates
JASPAR PBM (UNIPROBE) Mouse
JASPAR PBM (UNIPROBE) Homeo Mouse
JASPAR PBM (UNIPROBE) HLH Nematode
Drosophila DMMPMM (http://line.imb.ac.ru/DMMPMM/)
Drosophila IDMMPMM (http://line.imb.ac.ru/DMMPMM/)
RegulonDB prokaryotes
Add your own motif database:
title for this collection
Browse...
Matrices should be in Transfac format (other formats can be converted with convert-matrix).
Add known reference motifs for this experiment:
/Users/elodied
Browse...
Matrices should be in Transfac format (other formats can be converted with convert-matrix).
▼Locate motifs and export as UCSC custom track
Locate motifs
Search putative binding sites in the peak sequences [matrix-scan]
Visualize motifs in genome browser
No
Yes; sequences fetched from Galaxy(fasta headers should be in the form: >mm9_chr1_3473041_3473370_+)
Yes; use the following BED file.
BED file with peak coordinates
Choisissez un fichier
Aucun fichier choisi
Assembly version (UCSC)
The 4th column of the BED file (feature name) correspond to the fasta headers of input sequences
Output
display
email
Note: email output is preferred for very large datasets or many comparisons with motifs collections
GO
Réinitialiser
DEMO single
DEMO test vs ctrl
[MANUAL] [TUTORIAL] [ASK A QUESTION]

Discovered motifs (by algorithm)

motif discovery			
oligos_6nt_mkv4	oligos_6nt_mkv4_m1	asmb: **attaaa (sig=169.67)**  RC: **tttaat** 	[discovered words: text] [assembly: text sig matrices] [matrices: tab format transfac format]
	oligos_6nt_mkv4_m2	asmb: **cttatca (sig=112.46)**  RC: **tgataag** 	
	oligos_6nt_mkv4_m3	asmb: **atgcaa (sig=108.46)**  RC: **ttgcat** 	
positions_6nt	positions_6nt_m1	asmb: **aataatttatattaaaatat (sig=300)**  RC: **atattttaatataaattatt** 	[discovered words: text] [assembly: text sig matrices] [matrices: tab format transfac format]
	positions_6nt_m2	asmb: **gaaagggtta (sig=300)**  RC: **taacccttttc** 	
	positions_6nt_m3	asmb: **gagagag (sig=36.51)**  RC: **ctctctc** 	

Motif 5 positions_6nt_m2

[matrix: tab format transfac format]

Reference motifs

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rvrAAAGGGTTrar	aligned col. match
Kr_JASPAR	Kr_JASPAR	D	11	0.7857	0.945	0.743	.vrAAAGGGTTr..	crAAaGGGTTa
Kr_FlyReg_FBgn0001325	Kr_FlyReg_FBgn0001325	D	10	0.7143	0.959	0.685	...AAAGGGTTra.	AAmGGGTtaw

Total maches= 2

[match table: html text]
[alignments (logos): html text]

jaspar_core_insects

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rvrAAAGGGTTrar	aligned col. match
Kr	MA0452.1	D	11	0.7857	0.945	0.743	.vrAAAGGGTTr..	crAAaGGGTTa

Total maches= 1

[match table: html text]
[alignments (logos): html text]

DMMPMM_drosophila

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rvrAAAGGGTTrar	aligned col. match
Kr	Kr	D	10	0.7143	0.967	0.691	..rAAAGGGTTr..	aAAaGGGTta

Total maches= 1

[match table: html text]
[alignments (logos): html text]

IDMMPMM_drosophila

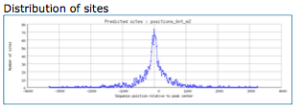
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif rvrAAAGGGTTrar	aligned col. match
Kr	Kr	R	11	0.7857	0.967	0.760	.vrAAAGGGTTr..	YRAAAGGGTTA

Total maches= 1

[match table: html text]
[alignments (logos): html text]

Predicted sites on input peaks

Distribution of sites



Enrichment in binding sites

[view in genome browser : UCSC]
[sites: text BED (UCSC track)]
[distribution: text]
[enrichment: text]

Motif 7 oligos_7nt_mkv5_m1

[matrix: tab format transfac format]

Reference motifs

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif wtATTTkTAww	aligned col. match

Total maches= 0

No match

jaspar_core_insects

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif wtATTTkTAww	aligned col. match
Abd-B	MA0165.1	D	6	0.5000	0.875	0.438	.....TKTAww	TTTATk.
cad	MA0216.1	D	6	0.5000	0.846	0.423	.....TKTAww	TTTATT.
hb	MA0049.1	R	9	0.7500	0.782	0.586	..ATTTKTAww	TTTTTTWYK.

Total maches= 3

[match table: html text]
[alignments (logos): html text]

DMMPMM_drosophila

name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif wtATTTkTAww	aligned col. match
zen	zen	R	6	0.5455	0.817	0.445	.....TKTAww	TTWATT
hb	hb	R	9	0.8182	0.805	0.659	..ATTTKTAww	TTTTTTRTK
br-Z1	br-Z1	R	8	0.7273	0.807	0.587	...TTTKTAww	TTWGTATY

Total maches= 4 (1 more)

[match table: html text]
[alignments (logos): html text]

IDMMPMM_drosophila

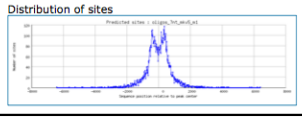
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif wtATTTkTAww	aligned col. match
cad	cad	D	7	0.5000	0.805	0.403	....TTKTAww	tTTTATk...

Total maches= 1

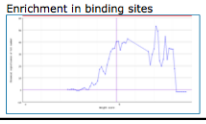
[match table: html text]
[alignments (logos): html text]

Predicted sites on input peaks

Distribution of sites



Enrichment in binding sites

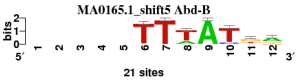
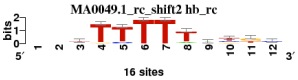


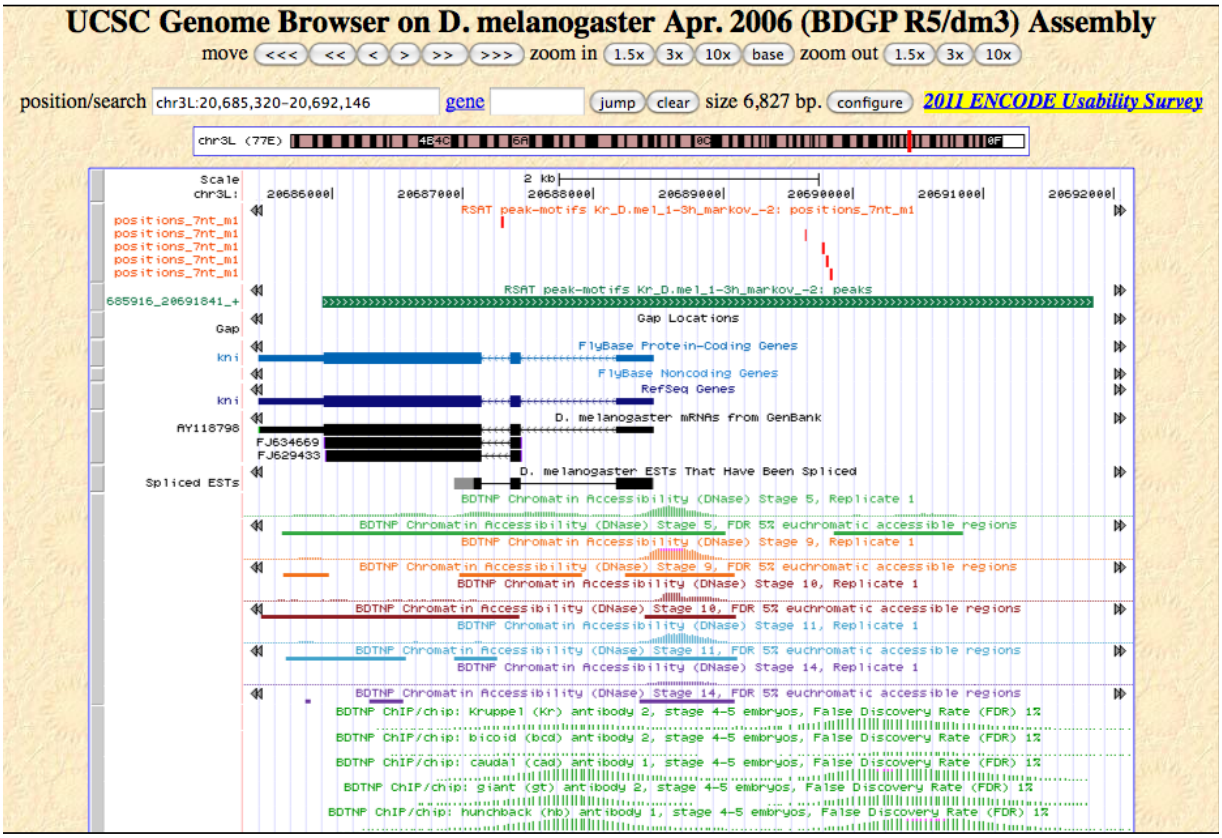
[view in genome browser : UCSC]
[sites: text BED (UCSC track)]
[distribution: text]
[enrichment: text]

One-to-n alignments

Command: compare-matrices -v 1 -mode matches -format1 transfac -file1 /home/rsat/rsa-tools/public_html/tmp/peak-motifs.2011_05_17.155419/results/discovered_motifs/oligos_7nt_mkv5_m1/peak-motifs_oligos_7nt_mkv5_m1.tf -format2 tf -file2 /home/rsat/rsa-

One-to-n matrix alignment; reference matrix: oligos_7nt_mkv5_m1_shift0 ; 4 matrices ; sort_field=rank_mean

Matrix name	Aligned logos	cor	Ncor	logoDP	NIcor	NsEucl	SSD	NSW	rcor	rNcor	rlogoDP	rNIcor	rNsEucl	rSSD	rNSW	rank_mean	match_rank
oligos_7nt_mkv5_m1_shift0 (oligos_7nt_mkv5_m1)	 8640 sites																
MA0165.1_shift5 (Abd-B)	 21 sites	0.875	0.438	5.092	0.429	0.898	0.753	0.937	1	2	1	1	1	1	1	1.143	1
MA0216.1_shift5 (cad)	 38 sites	0.846	0.423	3.758	0.394	0.890	0.879	0.927	2	3	2	2	3	2	2	2.286	2
MA0049.1_rc_shift2 (hb_rc)	 16 sites	0.782	0.586	2.177	0.053	0.891	1.921	0.893	3	1	3	3	2	3	3	2.571	3


UCSC Genome Browser on D. melanogaster Apr. 2006 (BDGP R5/dm3) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr3L:20,685,320-20,692,146 gene jump clear size 6,827 bp. configure 2011 ENCODE Usability Survey
chr3L (77E)
Scale
chr3L:
2 kb
20686000 20687000 20688000 20689000 20690000 20691000 20692000
RSAT peak-motifs Kr_D.mel_1-3h_markov_-2: positions_7nt_m1
positions_7nt_m1
positions_7nt_m1
positions_7nt_m1
positions_7nt_m1
RSAT peak-motifs Kr_D.mel_1-3h_markov_-2: peaks
685916_20691841_+
Gap
Gap Locations
FlyBase Protein-Coding Genes
kni
FlyBase Noncoding Genes
RefSeq Genes
kni
D. melanogaster mRNAs from GenBank
AY118798
FJ634669
FJ629433
D. melanogaster ESTs That Have Been Spliced
Spliced ESTs
BDTNP Chromatin Accessibility (DNase) Stage 5, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 5, FDR 5% euchromatic accessible regions
BDTNP Chromatin Accessibility (DNase) Stage 9, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 9, FDR 5% euchromatic accessible regions
BDTNP Chromatin Accessibility (DNase) Stage 10, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 10, FDR 5% euchromatic accessible regions
BDTNP Chromatin Accessibility (DNase) Stage 11, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 11, FDR 5% euchromatic accessible regions
BDTNP Chromatin Accessibility (DNase) Stage 14, Replicate 1
BDTNP Chromatin Accessibility (DNase) Stage 14, FDR 5% euchromatic accessible regions
BDTNP ChIP/chip: Kruppel (Kr) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
BDTNP ChIP/chip: bicoid (bcd) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
BDTNP ChIP/chip: caudal (cad) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%
BDTNP ChIP/chip: giant (gt) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
BDTNP ChIP/chip: hunchback (hb) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%

6.2.2 Evolution of major histocompatibility complex by "en bloc" duplication before mammalian radiation.

Durant mon année de master 2 et le début de ma première année de thèse au sein du laboratoire LATP UMR 6632 CNRS Evolution biologique et Modélisation sous la direction de M. Pontarotti, je me suis intéressé à l'évolution des génomes en étudiant particulièrement les duplications segmentales (duplication "en bloc"). Les duplications sont une source importante de nouveautés génétiques et fonctionnelles. Les mécanismes de duplication concernant l'ensemble du génome (polyploïdisation) ou un gène unique commençaient à être compris. Cependant, lors de l'étude de la duplication d'un gène, l'environnement génomique du gène était rarement étudié ce qui ne nous permettais pas de savoir si le gène avait été dupliqué seul ou avec les gènes environnement.

Cet article résume une stratégie pour la détection de tels événements par l'analyse phylogénétique et statistique de deux segments chomosomiques suspectés d'être issus d'une duplication. Nous avons étudier deux segments présents dans la région du complexe majeur d'histocompatibilité du chromosome 6 humain et suggéré qu'ils étaient issus d'une duplication en "bloc" qui avait eu lieu entre la divergence des amniotes et la séparation méthatériens/euthériens. Nous avons emis l'hypothèse que lors de duplication "en bloc", si un (ou plusieurs) gène était soumis à la sélection positive, il pourrait aider la fixation des gènes l'entourant. Cependant nous n'avons pas approndi cette hypothèse.

Immunogenetics (2008) 60:423–438
DOI 10.1007/s00251-008-0301-7

ORIGINAL PAPER

Evolution of major histocompatibility complex by "en bloc" duplication before mammalian radiation

Elodie Darbo · Etienne G. J. Danchin · Michael F. P. Mc Dermott · Pierre Pontarotti

Received: 12 February 2008 / Accepted: 28 April 2008 / Published online: 17 June 2008

Abstract Duplications are an important mechanism for the emergence of genetic novelties. Reports on duplicated genes are numerous, and mechanisms for polyploidization or local gene duplication are beginning to be understood. When a local duplication is studied, searches are usually done gene-by-gene, and the size of duplicated segments is not often investigated. Therefore, we do not know if the gene in question has duplicated alone or with other genes, implying that "en bloc" duplications are poorly studied. We propose a method for identification of "en bloc" duplication using mapping, phylogenetic and statistical analyses. We show that two segments present in the major histocompatibility complex (MHC) region of human chromosome 6 have resulted from an "en bloc" duplication that took place between divergence of amniotes and methaterian/eutherian separation. These segments contain members of the same multigenic families, namely olfactory receptors genes, genes encoding proteins containing B30.2 domain, genes encoding proteins containing immunoglobulin V domain and MHC class I genes. We will discuss the fact that olfactory receptors and MHC genes have undergone positive selection, which could have helped in fixation of the surrounding genes.

Keywords Evolution · "En bloc" duplication · MHC · Phylogeny · Hitchhiking effect

E. Darbo · P. Pontarotti (✉)
LATP UMR 6632 CNRS Evolution biologique et Modélisation,
Université de Provence,
case 19, 3 place Victor Hugo,
13331 Marseille Cedex 03, France
e-mail: Pierre.Pontarotti@univ-provence.fr

E. Darbo
e-mail: Elodie.Darbo@etu.univ-provence.fr

E. G. J. Danchin
UMR IBSV, INRA, UNSA, CNRS,
Centre de recherche de Sophia-Antipolis,
400 route des Chappes, BP 167,
06903 Sophia-Antipolis Cedex, France
e-mail: Etienne.Danchin@sophia.inra.fr

M. F. P. Mc Dermott
Leeds Institute of Molecular Medicine (LIMM),
Wellcome Trust Brenner Building,
St. James's University Hospital,
Beckett Street,
Leeds LS9 7TF, UK
e-mail: M.McDermott@leeds.ac.uk

Introduction

Genomes undergo many mutations, such as substitutions, deletions, insertions, inversions and duplications, during species evolution. It is well known that gene duplications provide a source of new genetic material that is relaxed from selective constraint and can evolve novel functions. Small-scale duplications may involve part of a gene, a whole gene, or several genes ("en bloc" duplication). Duplication may also involve the whole genome; the name of this process is polyploidization (Otto and Yong 2002). Species usually do not stay polyploid and eventually return to a diploid state. The diploidization process occurs via, in particular, recombination between the different chromosomes. At the end of this process, homologous chromosomes resulting from the polyploidization process will not be similar along their whole length but only in portions, defining paralogous regions along the genome (Lundin 1993).

Local duplication and polyploidization correspond to distinct evolutionary forces, with different impacts at the biological level. Local duplication occurs in all eukaryotes more or less continuously, as is the case with other types of mutations, such as point mutations, following the laws of

population genetics (Lynch and Conery 2003). In contrast, polyploidizations occur at different rates and not continuously; for example, three polyploidization events occurred in the last 150 millions years in the *Arabidopsis thaliana* lineage (Cui et al. 2006), whereas no polyploidization seems to have occurred in the *Drosophila* lineage for the last 600 million years (Rubin et al. 2000).

A common feature of local duplication and polyploidization is that after duplication, most of the copies are lost (the duplicate return to a single copy status; Lynch et al. 2001; Dehal and Boore 2005). The duplicates are lost, either by deletion or by pseudogenization, and, on average, only about 15% survived (Blomme et al. 2006). The conserved duplicates are retained, either because one of the duplicate shifted toward a new function (biochemical function, new expression territory) or because the two duplicates split their original function, a process known as subfunctionalization (Force et al. 1999).

Depending on the mode of duplication, local duplication versus polyploidization, different types of genes will be retained. If we consider, for example, a gene whose product is involved in a huge interaction network, in the case of local duplication, one of the members of the network will be supernumerary and will therefore alter the whole network stoichiometry. In this case, therefore, the duplications tend to be counter-selected (gene dosage hypothesis; Papp et al. 2003; Freeling and Thomas 2006). However, in the case of polyploidization, all the members of the network are duplicated. Therefore, if one of the members is lost, the stoichiometry will be altered. Hence, in this second case, the duplicate whose products are involved in a network will tend to be retained.

The remains of polyploidy (paleopolyploidy) has been extensively studied, both in plants (Cui et al. 2006) and in yeast (Kellis et al. 2004) for example.

Local duplications are usually studied as single gene duplications, and the size of the duplicate segment containing this gene is often not analyzed. Very few reports with analyses of segmental duplications containing several genes are available. These reports concern recent events, and only a few species, and the fate of the genes involved is not determined (Calvacanti et al. 2002; Jiang et al. 2007).

As we do not know the importance of "en bloc" duplication "a priori", we decided to search for paralogous regions, which are not remnants of polyploidization events, and we considered essentially the composition of the segments (similar sets of paralogs in each region). We then verified if these paralogs had resulted from the same duplication event. For this, we used phylogenetic analyses to date the duplications and then used statistical analysis to test the significance of the clusters of paralogous genes. We propose this strategy using the major histocompatibility complex (MHC) region as a pilot region, as during previous investigations of the human MHC region, we observed two segments (Horton et al. 2004) with similar genes composition and which contain members of the same multigenic families (olfactory receptors, MHC class I, proteins containing B30.2 domain and proteins containing IgV-MOG domain). We hypothesise that the gene distribution observed in these regions could be due to "en bloc" duplication rather than to a random process or a polyploidization event; indeed, this organisation could have occurred by chance. To test this hypothesis, we used a combination of mapping, phylogenetic and statistical analyses.

Materials, methods and strategy

Gene identification and description

Olfactory receptors family

Olfactory receptors (ORs) family contains approximately 1,000 members in mammals' genomes; these proteins are in contact with the environment, and indeed, they generally permit the perception of odorant volatile molecules. ORs are classified into 14 subfamilies that have emerged before birds/mammals divergence.

The regions under investigation contain several ORs subfamilies (subfamilies 1, 2, 5, 10, 11, 12, described in HORDE database; http://bioportal.weizmann.ac.il/HORDE). We used HORDE database to search the ORs sequences contained in the distal MHC region. In order not to miss certain sequences, we compared extracted sequences with sequences described by Horton et al. (2004) of this region and also with sequences found in ENSEMBL database (http://www.ensembl.org/). As with many other gene families, the ORs family evolved under the birth and death process, but, nevertheless, links between sequences are clear, and differential loss and gain of genes between species are easily observable.

Proteins with domain B30.2 family

The B30.2 protein family, observed only in vertebrate lineage, contains approximately 50 members. These proteins contain a globular domain, called B30.2, in the C-terminal extremity and different domains in N-terminal extremity. At least three groups are identified: RING finger, immunoglobulin-like and toxins (Henry et al. 1998). These proteins are more and more studied because they are involved in the recognition of some viruses (e.g. TRIM5α/HIV; Yap et al. 2004).

Myelin oligodendrocyte glycoprotein

The myelin oligodendrocyte glycoprotein (MOG) protein contains an IgV domain. It is synthesised late in brain

development and seems to be involved in preservation of myelin integrity around oligodendrocytes. This protein belongs to the IgV-MOG family (Tazi-Ahnini et al. 1997).

Butyrophilins

This family contains seven members divided into three subfamilies (1, 2 and 3). These proteins are composed of an IgV domain, an IgC domain and a B30.2 domain. Butyrophilins are probably the result of an exon shuffling event between a member of IgV (MOG) gene family and a member of B30.2 gene family (Tazi-Ahnini et al. 1997), and we consider that they belong to these two families.

The class I MHC family

These proteins are composed of an IgC1 domain, an $\alpha 1/\alpha 2$ domain and a transmembrane domain. The two domains were used to build phylogenies. We used the data of Hughes et al. (1999) to define family size and found that it contains 13 members.

Other genes

We used ENSEMBL database and the analysis by Horton et al. (2004) to identify genes contained in the regions of interest that are supposed to have occurred via gene duplication. We have also been interested in these genes because they could have duplicated, and thus, we would have to include them in the statistical analyses. We used the non-redundant (NCBI NR) database to extract the sequences; there are 64 genes, of which, 54 are histone genes.

Domains extraction

Domains were used for phylogenies because some proteins have several domains with different evolutionary stories (i.e. butyrophilins) or different rates of evolution (i.e. class I MHC). To extract all domains, we submitted protein sequences against PROSITE database (http://www.expasy.ch/prosite/). In the case of the B30.2 domain, we used the domain defined by Henry et al. (1998).

Phylogenetic analyses

We used the Figenix bioinformatics platform (Gouret et al. 2005) based on an expert system that allows automation of biological annotation pipelines. The phylogenetic pipeline CassiopePhylo+M was used here; it allows the detection of homologous sequences (orthologs and paralogs) for a given gene. It contains 52 steps (sequences retrieval, multiple alignments, tree building, filters, etc.). Consensus phylogenetic trees are obtained by fusion of trees built from three methods: maximum likelihood, maximum parsimony and distance. If tree topologies are not congruent, fusion cannot be achieved or only partially. Orthologs/paralogs inference is made from comparison between this tree and a reference taxonomy tree, available at NCBI web site (http://www.ncbi.nlm.nih.gov/). Robustness of the obtained trees (for the three methods used) was tested with a bootstrap method.

ORs family

These sequences were submitted to Figenix, using ENSEMBL database, and the four species of reference (see below in "Strategy"). We made phylogenies from every member of each subfamily present in the region under investigation, namely, subfamilies 1 (39 members), 2 (139 members), 5 (119 members), 10 (64 members), 11 (22 members) and 12 (three members).

Proteins with domain B30.2 family

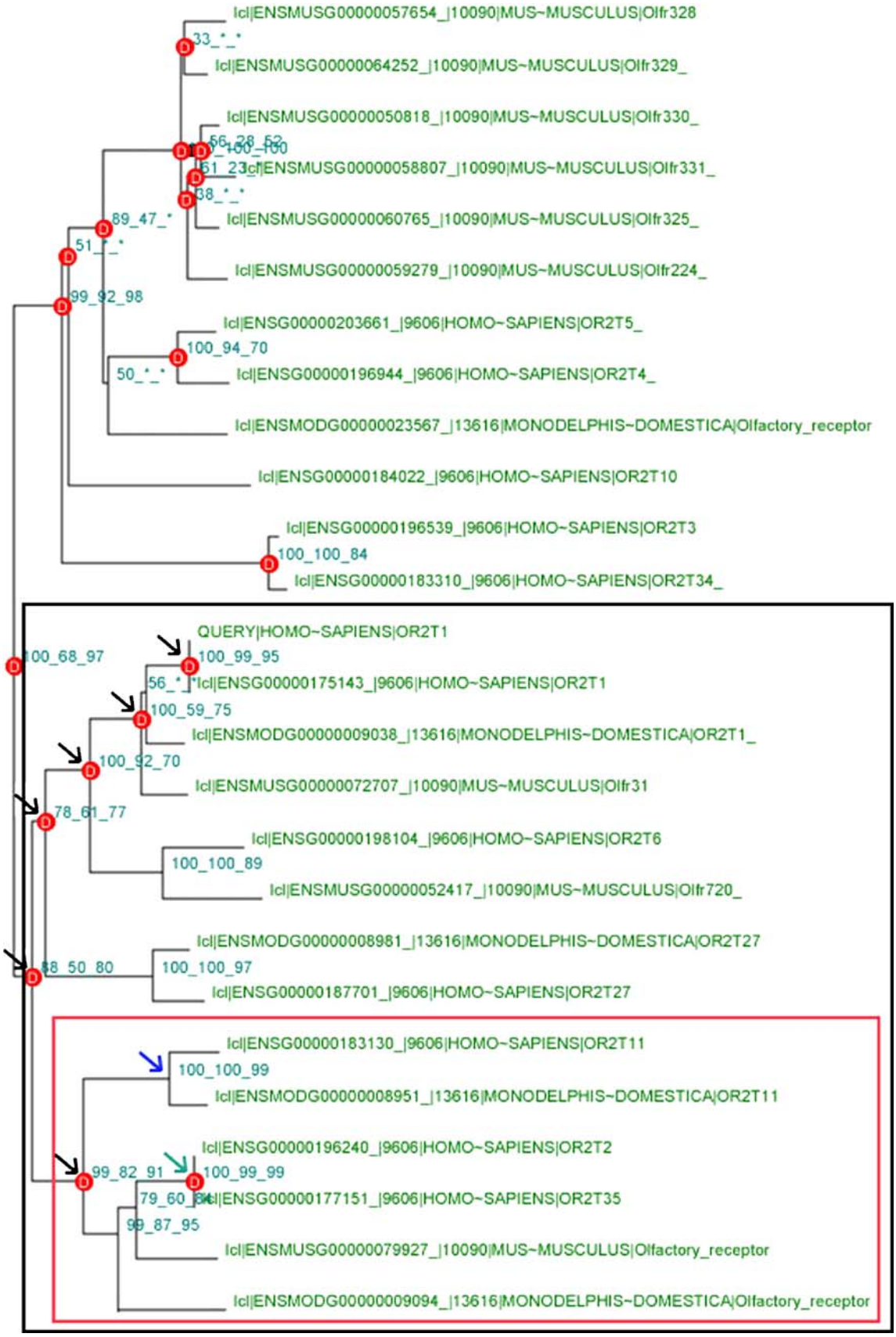
We searched all B30.2 proteins. We used BLASTp (http://www.ncbi.nlm.nih.gov/BLAST/; Altschul et al. 1990) to recover phylogenetically remote sequences so as to be sure we had retrieved all members of this family. We used the TRIM27 B30.2 domain as a query against the NCBI NR database. The B30.2 domain of each protein was submitted to Figenix using ENSEMBL+NR databases and the four species of reference, and we extended the analyses to other vertebrate genomes available in ENSEMBL+NR (see below in "Strategy"). Indeed, links between members of this family were unclear because of birth and death process. We calculate 83 phylogenies with the B30.2 domain of all members.

Myelin oligodendrocyte glycoprotein

The MOG IgV domain was submitted to Figenix using ENSEMBL+NR and the four species of reference.

Butyrophilins

Even if butyrophilins are composed with domains already studied with other families (IgV: MOG; B30.2: protein B30.2 family), we used the three domains (IgV, IgC, B30.2) of the butyrophilin subfamily 1A member 1 (BTN1A1), as defined by the PROSITE scan tool for phylogenetic analyses. Phylogenies were made using ENSEMBL+NR database and the species of reference. This analysis provides validation with regards to results obtained above, and it also allowed us to perform statistical testing from region 2 that contains the butyrophilin genes to region 1 (see "Statistical significance" for the definition of regions)


lcl|ENSMUSG00000057654_|10090|MUS~MUSCULUS|Olfr328
33_*_*
lcl|ENSMUSG00000064252_|10090|MUS~MUSCULUS|Olfr329_
lcl|ENSMUSG00000050818_|10090|MUS~MUSCULUS|Olfr330_
56_28_52
100_100_100
51_23_
lcl|ENSMUSG00000058807_|10090|MUS~MUSCULUS|Olfr331_
38_*_*
89_47_*
lcl|ENSMUSG00000060765_|10090|MUS~MUSCULUS|Olfr325_
51_*_*
lcl|ENSMUSG00000059279_|10090|MUS~MUSCULUS|Olfr224_
99_92_98
lcl|ENSG00000203661_|9606|HOMO~SAPIENS|OR2T5_
100_94_70
50_*_*
lcl|ENSG00000196944_|9606|HOMO~SAPIENS|OR2T4_
lcl|ENSMODG00000023567_|13616|MONODELPHIS~DOMESTICA|Olfactory_receptor
lcl|ENSG00000184022_|9606|HOMO~SAPIENS|OR2T10
lcl|ENSG00000196539_|9606|HOMO~SAPIENS|OR2T3
100_100_84
lcl|ENSG00000183310_|9606|HOMO~SAPIENS|OR2T34_
QUERY|HOMO~SAPIENS|OR2T1
100_68_97
100_99_95
56_*_*
lcl|ENSG00000175143_|9606|HOMO~SAPIENS|OR2T1
100_59_75
lcl|ENSMODG00000009038_|13616|MONODELPHIS~DOMESTICA|OR2T1_
100_92_70
lcl|ENSMUSG00000072707_|10090|MUS~MUSCULUS|Olfr31
78_61_77
lcl|ENSG00000198104_|9606|HOMO~SAPIENS|OR2T6
100_100_89
lcl|ENSMUSG00000052417_|10090|MUS~MUSCULUS|Olfr720_
lcl|ENSMODG00000008981_|13616|MONODELPHIS~DOMESTICA|OR2T27
68_50_80
100_100_97
lcl|ENSG00000187701_|9606|HOMO~SAPIENS|OR2T27
lcl|ENSG00000183130_|9606|HOMO~SAPIENS|OR2T11
100_100_99
lcl|ENSMODG00000008951_|13616|MONODELPHIS~DOMESTICA|OR2T11
lcl|ENSG00000196240_|9606|HOMO~SAPIENS|OR2T2
99_82_91
100_99_99
79_60_
lcl|ENSG00000177151_|9606|HOMO~SAPIENS|OR2T35
99_87_95
lcl|ENSMUSG00000079927_|10090|MUS~MUSCULUS|Olfactory_receptor
lcl|ENSMODG00000009094_|13616|MONODELPHIS~DOMESTICA|Olfactory_receptor

◀ **Fig. 1** Example of fused phylogenetic tree. The OR2T1 sequence (OR subfamily 2) was submitted as query in figenix. The sub-tree *squared in black* shows six duplications. Five duplications (*black arrows*) concern events that occurred between eutherians/metatherians separation; these five duplications are counted for OR subfamily 2. One duplication (*green arrow*) concerns a human specific duplication; this duplication is not considered in our count. We can observe the birth and death process in the sub-tree *squared in red*; *blue arrow* shows a speciation event, but no ortholog was found in *M. musculus*. The gene could be lost in this lineage after duplication

Class I MHC family

Phylogenies were made with α1/α2 and IgC domains of HLAA, MICA, MICB and HFE proteins. We submitted the sequences against the ENSEMBL database (Fig. 1).

Histones

We used the phylogenetic analysis of Malik and Henikoff (2003). The authors showed that duplications involving the 54 histones occurred before bird/mammal separation.

Other genes

A total of ten phylogenies were made from genes that do not belong to the families described above and that are present in the regions that described above. We used ENSEMBL+NR with the four species of reference (gene ID in Table 1).

Gene mapping

Gene mapping was done using ENSEMBL tool: Graphical view (for example TRIM38 localization in human genome: www.ensembl.org/Homo_sapiens/contigview?l=6:26071254-26093327).

Statistical significance

Implementation of the binomial test

We defined departure and arrival regions as the two regions supposed to have been engendered by duplication. The binomial test is used to determine the probability that the members of one region (departure) gave paralogs in the supposed paralogous region (arrival) by chance. It is extremely important to note here that we will use the duplication time information. It has to be taken into an account that the paralogs have to be issued from duplications that occurred in the same window of times. In some cases, duplicate issued from these duplications have duplicated later. The late duplications occurred in the same region "cis duplication" and remained there, or, alternatively, could have translocated elsewhere in the genome, in which case, we cannot know a priori the ancestral localisation. Therefore, in such cases, a corrector coefficient should be applied. However, we have never witnessed the last case; see after). We applied the binomial test twice; we applied the test once with region 1 [human chromosome 6: 30.4Mb (TRIM39)..28.98Mb (TRIM27)] as the departure region and region 2 [human chromosome 6: 28.1Mb (OR1F12).. 26.07Mb (TRIM38)] (Fig. 2) as arrival region, and on one occasion with region 2 as departure region and region 1 as arrival region.

We tested two alternative hypotheses H0 and H1:

H0: Gene distribution between "departure" region and "arrival" paralogous region results from a random process.
H1: Gene distribution between "departure" region and "arrival" paralogous region does not result from a random process.
p: probability of each gene contained in "departure" region having a paralog (or a group of paralogs in the case of late duplication) in the "arrival" region.
q: probability that the paralogs localise elsewhere in the genome ($q=1-p$).

We hypothesised that the probability of one gene belonging to a region of the genome is proportional to gene density in this region.

We used ENSEMBL approximation of the number of genes in the human genome, 22,741 genes. M is the number of genes in the arrival region, $p=M/22{,}741$.

We note n as the total number of duplications that occurred in the period defined below (see "Strategy") inferred for a family or a group of genes by phylogenetic analyses made from all members of families. We note k as the number of duplications having engendered a paralog in each region.

$$\forall i \in \{0,\ldots,n\} : \; P(X=i) = C_n^i p^i q^{n-i}$$

$P(X=i)$: probability to find i duplications engendering a paralog in each paralogous region under H0 hypothesis, that is, these duplications having engendered duplicates in each region by chance.

$$\alpha = P(X \geq k) = \sum_{i=k}^{n} P(X=i).$$

α is the risk to reject H0 while it is true, if:

$\alpha<5\%$: results are significant
$\alpha<1\%$: results are very significant
$\alpha<1\%0$: results are highly significant

Strategy

Boundaries of the regions and paralogs distribution analysis

To use the binomial test, we had to delimit the regions to be tested. We delimited first after observation of chromosomal

Table 1 Definition of region 1 (chr6, 29.01–30.4Mb) and region 2 (chr6, 26.07–28.14) in human genome

Protein name	Accession number	Localization (Mb)	*n*	*k*
Region 1	chr6, 29.01–30.4 Mb			
TRIM39	CAM25726	30.4	4	1
TRIM26	CAM26027	30.28	8	0
TRIM10-15	CAM26237-CAM26021	30.22–30.23	1	0
TRIM40	CAM26236	30.22	9	0
TRIM31	NP_008959	30.17	10	0
RNF39	CAM26286	30.14	0	0
PPP1R11	CAM26285	30.14	0	0
ZNRD1	CAG33390	30.13	0	0
HCG9	NP_005835	30.05	0	0
HLA F-G-A	BAB63337-CAA43298-CAA64263	29.79–29.9	7	1
MOG	CAM25974	29.73	4	1
GABBR1	CAI17391	29.63	2	0
UBD	CAI18599	29.63	0	0
MAS1L	CAI18459	29.56	0	0
OR10C1	BK004207	29.51	32	0
OR11A1	BK004208	29.5	15	0
OR12D1P-2-3	NG_002196-BK004523-BK004439	29.49–29.47–29.45	3	0
OR5V1	BK004440	29.42	43	0
OR5U1	BK004441	29.32	44	0
OR2U1P-2P	BK004667-BK004666	29.37–29.24	23	0
OR2G1P	NG_004689	29.3	24	0
OR2H1-2-4P-5P	AF042073-AB065965-NG_004685-NG_004662	29.53–29.66–29.29–29.64	2	0
OR2N1P	AF399498	29.2	25	0
OR2J1-2-3-4P	NG_004683-NM_030905-AF399630-AB065683	29.17–29.24–29.18–29.25	1	0
OR2B3-B4P	AF399632-NG_004686	29.16–29.36	3	1
OR2P1P	NG_004693	29.14	26	0
OR2W1	AF399628	29.12	4	1
OR2AD1P	NG_002239	29.1	27	0
RFP	CAM25872	29.01	5	0
Region 2	chr6, 26.07–28.14 Mb			
OR1F12	AB065468	28.14	17	0
OR2B8P	AB065681	28.12	1	0
OR2W6P-4P-2P	AB065475-BK004560-NG_004279	28.01–28.05–28.11	4	1
OR2B2-6-7P	NP_149046-NP_036499-NG_004280	27.9–28.03–28.12	3	1
PRSS16	CAB94769	27.32	0	0
GUSBL1	AAH67351	26.76	0	0
ABT1	NP_037507	26.7	0	0
HMGN4	O00479	26.64	0	0
BTN3A3	CAA17273	26.55		
BTN3A2	CAA17277	26.47		
BTN3A1	NP_008979	26.51		
BTN2A3	NP_076923	28.53	9	2
BTN2A2	AAH17497	26.49		
BTN2A1	CAB71221	26.56		
BTN1A1	AAH96315	26.6		
HFE	CAA70934	26.2	7	1
Hist1 (54 genes)	Malik and Henikoff (2003)	26.39–26.12	0	0
TRIM38	O00635	26.07	11	0

location of the genes belonging to the families present in the MHC region, and we refined them using phylogenies made from all genes contained in these regions. Region 1 starts at 30.4 Mb (TRIM39) and ends at 28.98 Mb (TRIM27); region 2 starts at 28.1 Mb (OR1F12) and ends at 26.07 Mb (TRIM38; Fig. 2 and Table 1).

The significance of results obtained was examined with a probabilistic test that permitted us to discriminate between

two hypotheses: (1) gene distribution is due to a random process and (2) gene distribution is not due to a random process. If hypothesis 1 is rejected, two explanations are possible: (1) the two regions result from a "bloc duplication" and (2) each gene duplicate individually in the same period of time and the duplicates have translocated in the same region independently (convergent evolution). Starting from a determined region, the probability of finding by chance clustered copies of those genes in another delimited region is weak in comparison with the probability of finding copies scattered elsewhere in the genome.

Dating duplication events

Determination of the duplication time is deduced with regard to all phylogenies, and we determined in which period genes in regions of interest had duplicated. We started analyses using all vertebrate species available in ENSEMBL + NR. These first results were not analysable because of the excessively large size of the data; moreover, species too close phylogenetically (that diverge recently as apes and humans or mice and rats) did not provide information about the evolution of our regions. Therefore, we decided to use the more phylogenetic informative species as the species of reference: *Gallus gallus* (chicken, taxID: 9031), *Homo sapiens* (human, taxID: 9606), *Mus musculus* (mouse, taxID: 10090), *Monodelphis domestica* (opossum, taxID: 13616). We chose these species with regards to the first phylogenies, made from members of families present in the regions studied, using all available vertebrates' genomes. Phylogenies from the genes contained in regions under investigation [region 1: human chromosome 6: 30.4Mb (TRIM39)..28.98Mb (TRIM27) and region 2: human chromosome 6: 28.1Mb (OR1F12)..26.07Mb (TRIM38)] (Fig. 2) indicate that the duplication events have engendered a duplicate in each region and occurred after bird/mammal separation. Indeed, when orthologs were found in *G. gallus*'s genome, they systematically constituted an outgroup. Some genes appeared more recently (species specific genes); for example, TRIM48/TRIM49/TRIM51 have emerged in the human genome after rodents/primates separation. Others have an ortholog for at least one human or mouse duplicated gene in *Monodelphis domestica* genome. This indicates that duplications have probably occurred before placental mammal/marsupial separation.

Counting duplication events

We have counted duplication events that occurred in a restricted period: between 173 and 310 million years (Mya; Kumar and Hedges 1998). During the analyses of these phylogenies, we observed several cases (Fig. 3) that gave different results with regards to the statistical analysis that is the increment of the binomial test parameters: n, the total number of duplications that occurred during the period defined above and k, the number of duplications that occurred in this period which engendered paralogs in each region defined above.

Case 1 The ideal case: the phylogenetic tree topology fulfil the three following criteria and no duplication occurred in the mammalian lineage: (1) human paralogs are shared between the two regions; (2) at least one of the human paralogs possesses an ortholog in *M. domestica* genome (one copy can be lost through the birth and death process); (3) there exists a single ortholog in *G. gallus* genome that is placed as an outgroup with regard to the mammalian genes and (4) the duplication event is well supported in the tree. Here, $n = k$.

Case 2 The phylogenetic tree topology fills the four previous criteria and duplication(s) that occurred in the mammalian lineage: we would increment n by 1 and k would equal 1 divided by the total

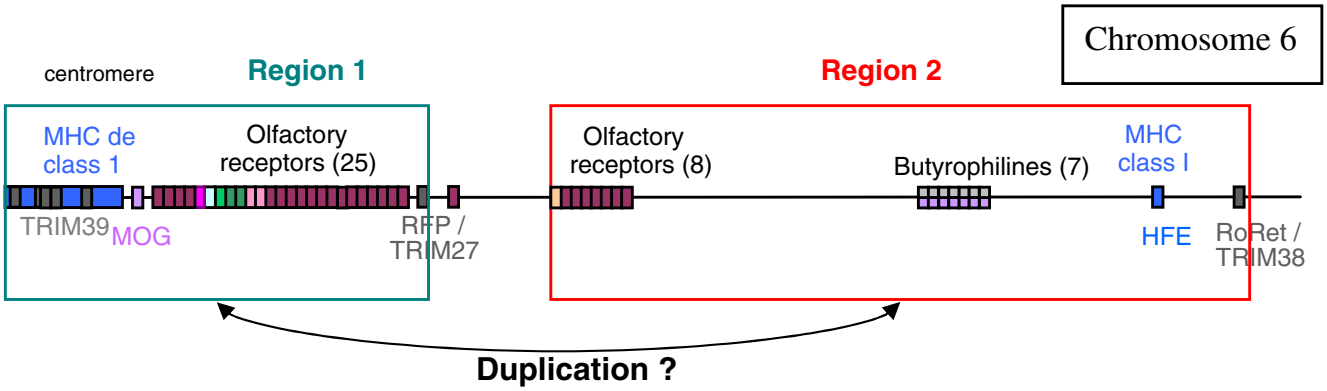
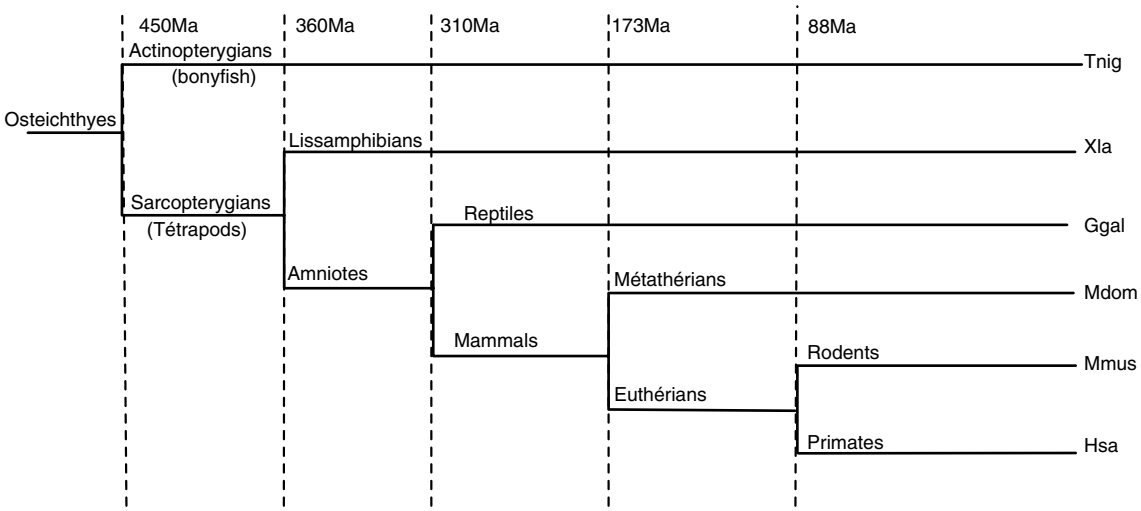
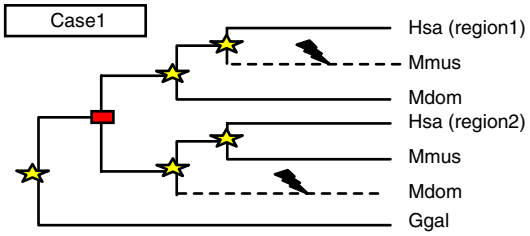
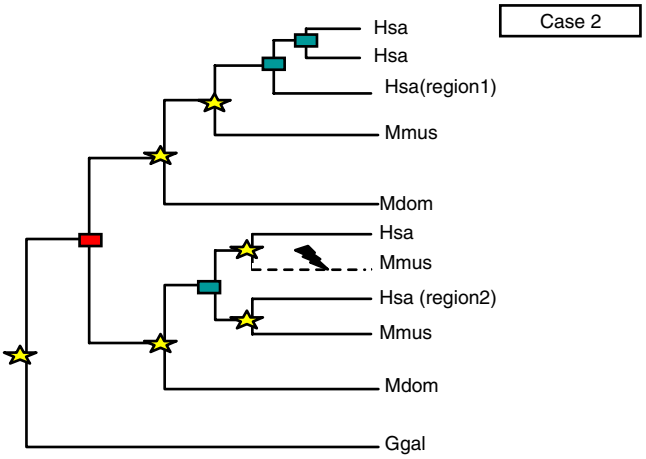
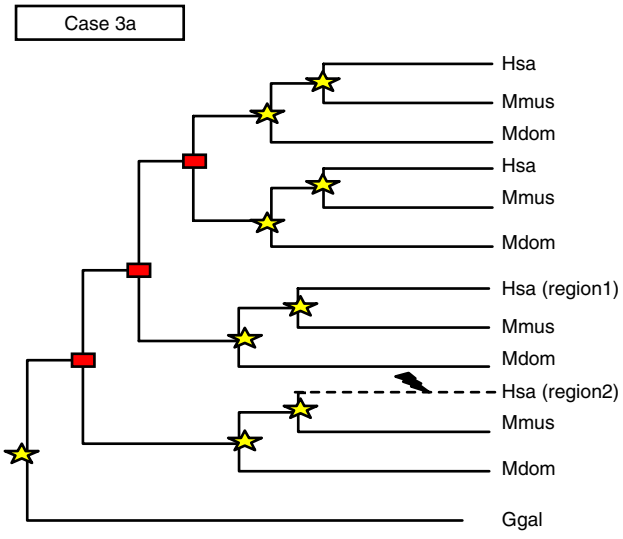


Fig. 2 Simple description of human MHC distal region of chromosome 6. Region 1 (*green*) and region 2 (*red*) contain members of the same multigenic families: olfactory receptors (subfamilies 1 in *beige*, 2 in *purple*, 5 in *light pink*, 10 in *white*, 11 in *pink*, 12 in *green*), MHC class I in *blue*, MOG in *light purple*, B30.2 in *grey* and butyrophilins in *light purple* and *grey*









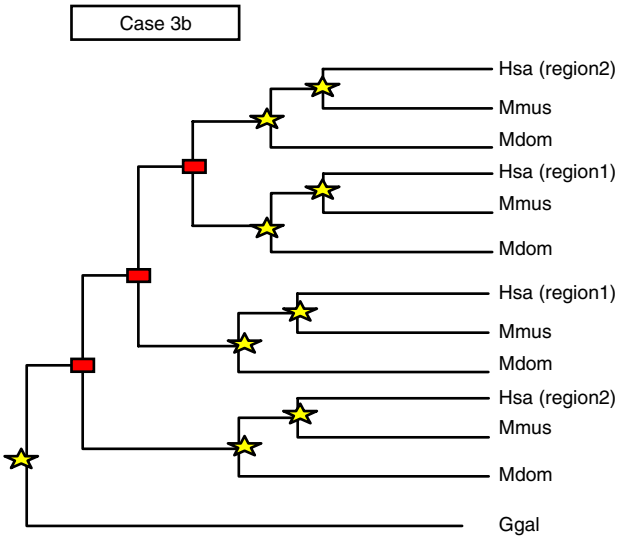


Fig. 3 Schematic representation of the different cases found in the phylogenetic trees as described in strategy ("Materials, methods and strategy"). At the *top*, we present the phylogenetic tree of vertebrate species. *Hsa H. sapiens*, *Mmus M. musculus*, *Mdom M. domestica*, *Ggal Gallus gallus*, *Xla Xenopus laevis*, *Tnig T. nigroviridis*. Following Xla and Tnig are noted "other vertebrates": *yellow stars* represent speciation events, *red squares* represent duplication events that occurred between amniotes and mammals radiations (310–173 Mya), *blue squares* represent duplication events that occurred out of the period of interest, *dotted lines and sparks* represent gene lost or rapid evolution

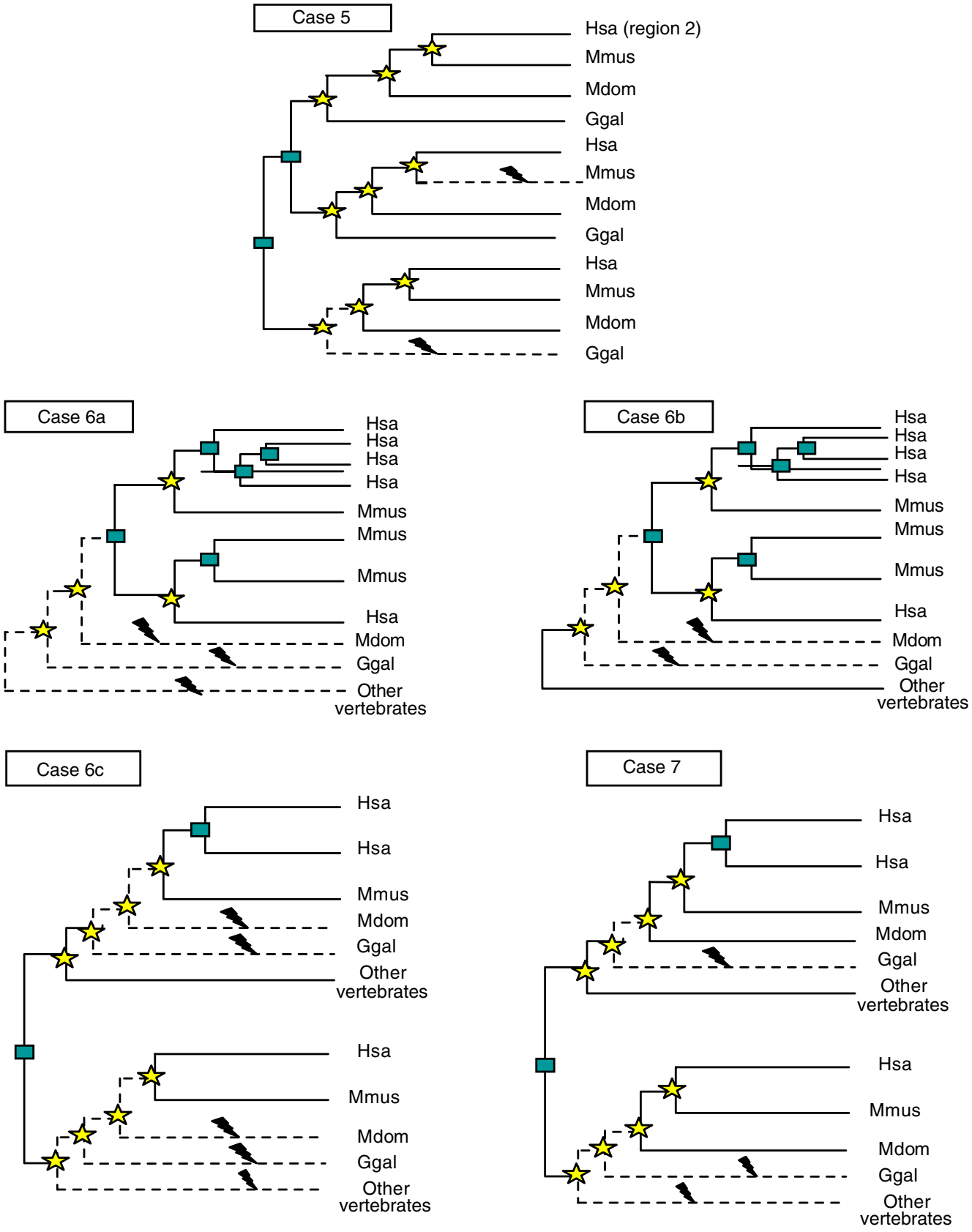


Fig. 3 (continued)

number of duplications that occurred since the first duplication included. Indeed, we could not know which duplication engendered the gene localised into the region under investigation.

Case 3 Criteria (2) and (3) are filled and phylogenetic links between members of a family are unclear, but an ortholog in *M. domestica* genome exists for each human paralog; we count a single duplication for each member of the family −1 (for example, three duplications engender four genes). Therefore, if m is the number of member engendered during the period of interest, $n=m-1$. For k parameter, two sub-cases exist: (3a) if one or more members localise in only one of the two regions even if a copy that localise in the other region could be lost, $k=0$; (3b) one or more members are localised in the two regions, $n \geq k \geq 1$, it would be determined for each case depending on the tree topology.

Case 4 It is a particular case of case 3: phylogenies do not give us clear relations, but groups of paralogs can still be highlighted. For example, for the B30.2 protein family, six members belong to one sub-tree in all phylogenies, but links between them change. Therefore, for this group, we counted five duplications ($n=5$; k is determined in "Results" below; Table 1).

Case 5 Criterion (3) is not filled; one or more orthologs are found in *G. gallus* genome for at least one human paralog. The duplication occurred before the amniotes radiation. Therefore, $n=k=0$.

Case 6 No ortholog exists in *G. gallus* and *M. domestica* genomes: we used phylogenetically distant species from amniotes lineage as *Xenopus laevis*, *Tetraodon nigroviridis* or all available vertebrate genomes.

6a No more orthologs are found in the remote species (likely due to too rapid evolution): the duplication is suggested to have occurred after the mammalian radiation.

6b One ortholog is found in one ore more distant species genomes located as out group(s) in the tree: we deduced that the duplication occurred after the mammalian radiation.

6c Orthologs are found for one or more mammalian paralogs in distant species genomes: this suggests that the duplication occurred before the amniotes radiation. In these three cases, n and k are not incremented.

Case 7 Criterion (3) is not filled: no ortholog exists in *G. gallus*; we used phylogenetically distant species from amniotes lineage as *X. laevis*, *T. nigroviridis* or all available vertebrate genomes. We should conclude that the duplication occurred in the period of interest but, as in the case of some members of the B30.2 proteins family, such as TRIM16/16L, TRIM62, TRIM46, at least one ortholog was found in the distant species genomes for one or more mammalian paralogs, which suggests that the duplication occurred before the amniotes radiation. In this case, $n=k=0$.

Results

We observed two chromosomal segments on human chromosome 6, involving the distal region of MHC, in which we can find clusters of members of the same multigenic families [ORs, B30.2, IgV (MOG), class I MHC]. We hypothesised that this gene distribution was due to "en bloc" duplication. We present (in Fig. 4a) an ideal case of "en bloc" duplication where no mutation (gene loss, inversion, neo duplication, etc.) occurred since the duplication; however, if the duplication is ancient, some of these events have probably occurred.

If "en bloc" duplication did occur, we expect to find that the duplication time is the same for all the couples of genes contained in the studied regions and that the results are statistically supported. To test this hypothesis, we calculated phylogenies using sequences of all proteins coded by genes present in the two chromosomal segments under investigation. As a result of these phylogenetic analyses, we could determine a period during which some genes from these regions appeared as a result of duplication, giving rise to a paralog in each region. In Fig. 4b, we present ideal phylogenies where duplication occurred between A1/A2 separation and A1c speciation. All genes (1, 2, 3, 4, 1′, 2′, 3′ and 4′) contained in the duplicated segments (a) and (b) possess an ortholog in every current species (A1a, A1b and A1c), and each paralog pair possesses one ortholog in the current species that have diverged from ancestral species A1 before the duplication (A2a).

Having determined the period, we were then able, through phylogenetic analyses with sequences of members of the families described in "Materials, methods and strategy", to look for duplications that have occurred in this period and to count the number of paralogs issued from such duplication localised in the studied regions and elsewhere in the genome. With this enumeration, we could test our hypothesis, and using a statistical test (binomial test) we tested if the distribution of the paralogs in two clusters is significantly different from what could be observed in a random model.

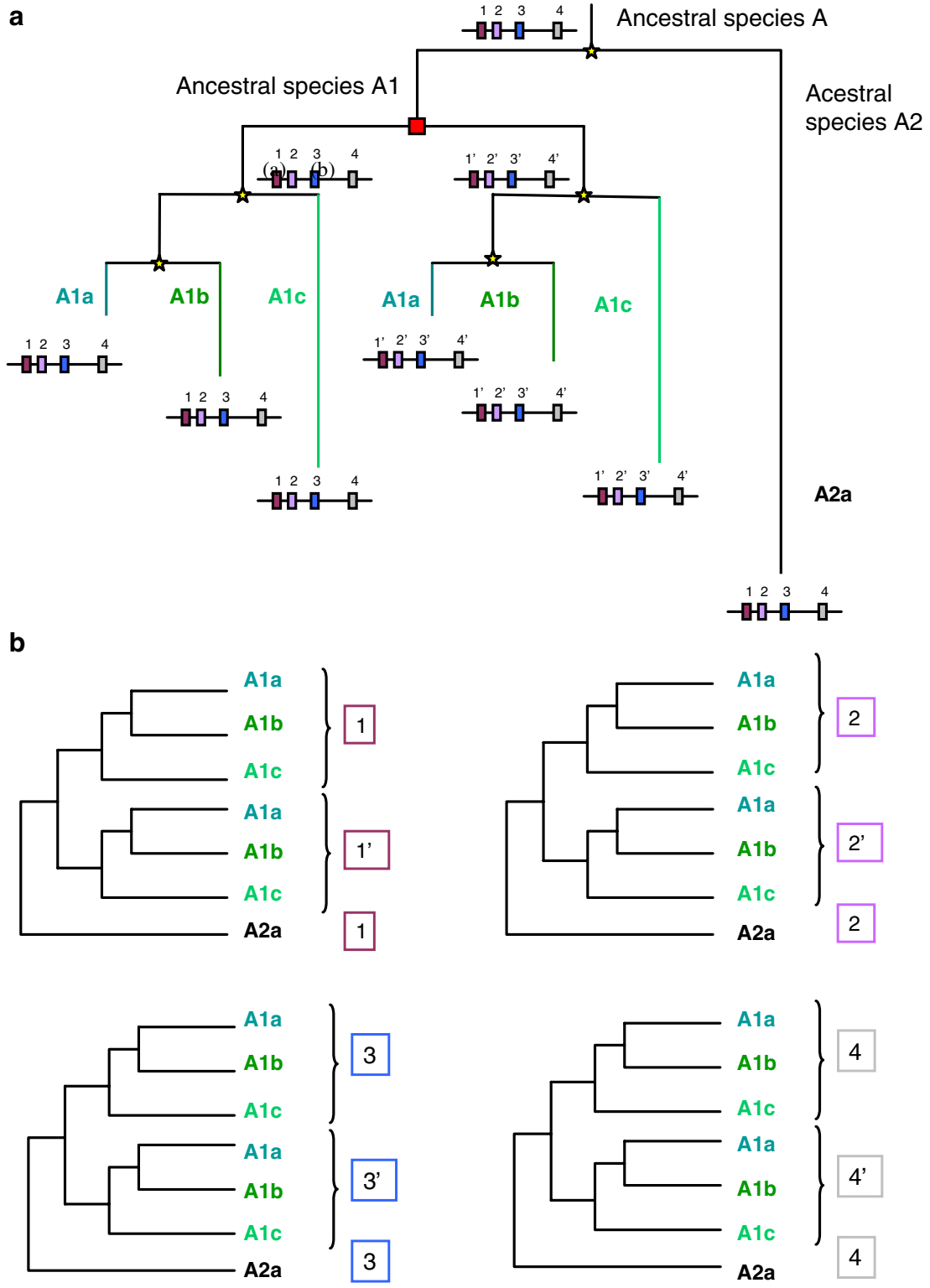


Fig. 4 "En bloc" duplication concerns a genomic segment containing several genes that has been duplicated. **A** Here, an ancestral species A contains in its genome a genomic segment with four genes (1, 2, 3 and 4). First, a speciation event (*yellow stars*) occurred, and species A diverged to species A1 and A2. In A1 lineage, an "en bloc" duplication occurred (*red square*); species A1 contains two paralogous segments (*a*) and (*b*). Genes contained in these segments are two–two paralogs (1–1′; 2–2′; 3–3′, 4–4′). Two speciation events follow the duplication and species A1a, A1b and A1c arose. These three species contains the two segments in their genomes. **B** Attempted phylogenetic results. Research of orthologs of the paralogs in given species allows dating the duplication; indeed, if genes contained in segment have duplicated in the same period, it is a first argument to prove "en bloc" duplication. Here, duplication has occurred after A1/A2 speciation and before A1a/A1b/A1c speciation. In species A2a, descending from species A2, we observed one copy of each gene. In phylogenies, these genes localise as outgroup with regard to other species

Fig. 5 OR2B4P phylogeny, consensus tree between parsimony and NJ tree. Proteins coding by genes contained in region 1 are *squared in red*, and those contained in region 2 are *squared in blue*. We can observe three duplications (*1*, *2* and *3*) that occurred between eutherians/metatherians, separation and birds/mammals separation (presence of a *G. gallus* sequence as outgroup). Duplication 1 correspond to a *cis* duplication that occurred likely after the "en bloc" duplication, duplication 2 gave a member in the region 1, k is incremented by 1, and duplication 3 led to a paralog in region 1 but paralog in region 2 had been probably lost; so, as described in case 3a ("Materials, methods and strategy"), k is not incremented. Therefore, n=3, but k=1

Enumeration of duplication events by phylogenetic analyses

Olfactory receptors (ORs) family

ORs family is composed by several subfamilies as described previously

At least 35 duplication events occurred in the period 173–310 Mya in ORs subfamily 2. Phylogenies from OR2B2 and OR2B3 (Fig. 5) show that relationships between these sequences are stable, and these relationships are found in all phylogenies where this paralog group appears. Genes OR2B2 (region 2)/OR2B3 (region 1; n=3; k=1) and OR2W1 (region1)/OR2W6P (region 2; n=4; k=1) come from duplication events in the period defined above, so two duplications engendered paralogs in each region (1 and 2) occurring in the defined period. The results for all members are available in Table 1. For this subfamily, n=35 and k=2.

Phylogenetic studies performed on the ORs subfamily 5 highlighted 44 duplications between bird/mammal separation and placental mammal/marsupial separation. For ORs subfamily 11, 15 duplications occurred in this period. For ORs subfamily 10, 32 duplications occurred, and for ORs subfamily 1, 17 duplications have been highlighted (Table 1).

Table 2 Duplication number inferred to members of families and the name of their paralog(s)

Region 1: paralog protein names	Duplication numbers	Region 2:paralog protein names
HLAA/F/G	7	HFE
OR2B3-B4P	3	OR2B2-6
OR2W1	4	OR2W6P-4P-2P
MOG	4	BTN
RFP/TRIM39	5	BTN

On the left: genes contained in region 1 and on the right those contained in region 2. First line: MHC class I members, second and third: ORs subfamily 2 members, fourth: IgV MOG family members and fifth: B30.2 family members. For butyrophilins, inferred number of duplication is 5(B30.2) + 4(IgV MOG)=10.
BTN butyrophilins

Proteins with a B30.2 domain

B30.2 protein family phylogenies performed from every member show a total of 18 duplications in the period of interest (Tables 1 and 2). Here, n=18 of which, as described in case 4 ("Materials, methods and strategy"), five duplications are attributed to a group of six members (TRIM39, RFP, BTN, TRIM11, ERMAP, MEFV; k=1).

Myelin oligodendrocyte glycoprotein

Phylogenies led from the IgV domain of MOG (region 1) have highlighted four duplications in period of interest. A single duplication concerns butyrophilins (region 2). Therefore, n=4 and k=1.

Butyrophilins

Phylogenies show evidence of an ortholog of subfamily 2 in *M. domestica* genome. Relationships between subfamily 2 and the others are not clear. Indeed, subfamilies 1 and 3 are always linked in phylogenies; they seem to have duplicated in mammals' genome, and, indeed, they possess orthologs in *M. musculus* and *H. sapiens* genomes but not in *M. domestica* genome. Orthologs in *M. domestica* genome have probably been lost and duplication occurred before eutherian/metatherian separation. We count a single duplication which has engendered ancestral subfamily of subfamilies 1 or 3 and subfamily 2 for this family, although this duplication could have occurred after the period of interest. As butyrophilins belong to IgV (MOG) and B30.2 families, we added duplications occurring in B30.2 and IgV (MOG) families and, in total, nine duplications (Tables 1 and 2). *G. gallus* genome contains a gene having the same domains; however, phylogenies derived from human and *G. gallus* sequences suggest that exon shuffling events have occurred independently in these two lineages. Here, n=9 and k=2 [one giving MOG gene and one giving TRIM39 (region 1) as the exon shuffling occurred after the "en bloc" duplication].

MHC class I family

MHC class I family has duplicated seven times (Tables 1 and 2) during the period determined above. The HFE (region 2) comes from HLA gene (region 1) duplication (n=7 and k = 1).

Other genes

Phylogenies from the other genes contained in these regions have highlighted that these genes are in single (unique) copy or they have duplicated out of the period of interest.

Statistical significance

The binomial test was used to test whether the observed distribution of paralogous genes could be explained by:

1. a random distribution
2. non-random distribution.

The non-random distribution could be explained either by block duplication or by localisation of the paralog in the same chromosomal region (convergent evolution). The probability, from a delimited chromosomal region, of finding by random a cluster of copies in another chromosomal region, suggested to have come from duplication, is weaker than finding copies randomly distributed on the genome

We tested first whether duplication had occurred from region 1 to region 2, counting how many paralogs of chromosome 6 region 1 genes are located in chromosome 6 region 2 in comparison with paralogs located elsewhere in the genome. Secondly, we tested in the interval from region 2 to region 1 ("Materials, methods and strategy").

From region 1 to region 2

Region 1 contains 36 genes; at least five genes have a paralog in region 2 that contains 66 genes.

p is the probability that these paralogs localise in region 2, q is the probability that these paralogs localise elsewhere in genome.

$$p = 66/22,741 = 2.9 \times 10^{-3} \quad q = 1 - p = 0.9971$$

n is the total number of duplications that occurred in the defined time period and k the number of duplications having engendered a paralog in region 2 (see "Materials, methods and strategy").

$$n = 313 \quad k = 5$$

α represents the risk to reject H0 (random hypothesis materials and methods):

α=0.0003

The result is highly significant because alpha is lower than 0.001. The hypothesis that genes distribution between this two regions occurred by chance can be rejected.

From region 2 to region 1

p is the probability that the paralogs localise in region 1, and q is the probability that these paralogs localise elsewhere in genome.

$$p = 36/22,741 = 1.58 \times 10^{-3} \quad q = 1 - p = 0.99842$$

n is the total number of duplication and k the number of duplication having engendered a paralog in region 2.

$n = 52 \qquad \kappa = 5$

α represents the risk to reject H0

$\alpha = 2.97 \times 10^{-10}$

The H0 hypothesis can be rejected with high significance because the alpha risk is inferior to 0.001.

Localisation of the orthologous genes in non-human mammalians genomes

We have analyzed the localization of the orthologous genes from the human regions 1 and 2 in *M. musculus* and *M. domestica* genomes. Figure 6c shows the organisation that we can observe today in *H. sapiens*, *M. musculus* and *M. domestica* genomes. The distribution of paralogs in these species is conserved between mammals. These results reinforce the hypothesis of an ancestral duplication at the origin of the two observed regions in human and other mammals, and it suggests that this duplication occurred before the divergence of mammals. From these results, we propose a hypothetical ancestral architecture of the MHC region in the amniotes ancestor before the duplication (Fig. 6a) and in the mammalians ancestor after the duplication (Fig. 6b).

During the analysis of orthologous regions in other vertebrate genomes, we noticed that some orthologs have been lost differentially between vertebrate species or the presence of more recent paralogs within a given lineage (birth and death process), so phylogenies as clear as

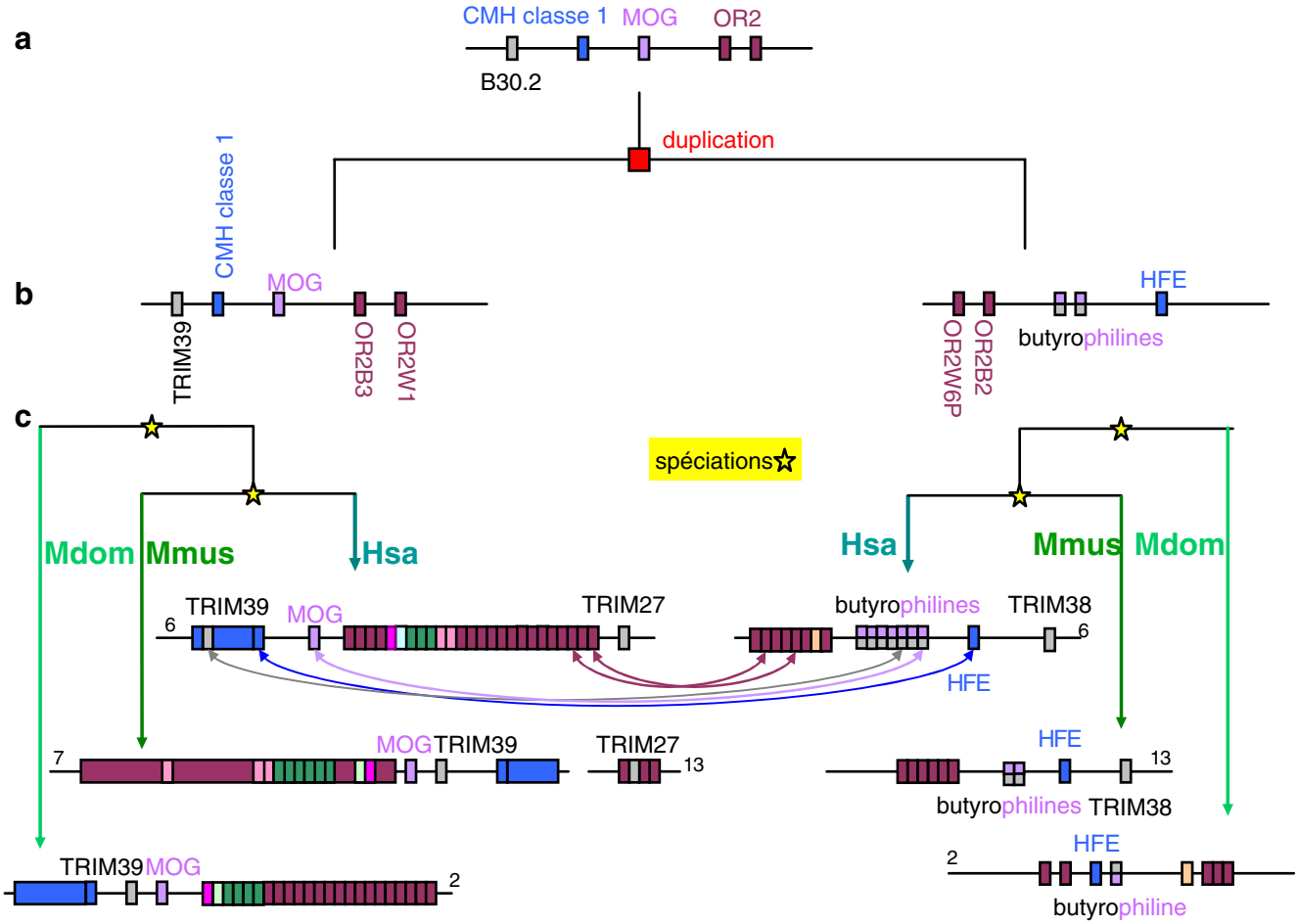


Fig. 6 Hypothesis of MHC distal region evolution in *H. sapiens* (*Hsa*), *M. musculus* (*Mmus*) and *M. domestica* (*Mdom*). **a** Unique hypothetical sequence in an ancestral species. **b** In the same ancestral species as in **a**, duplication event occurred and gave two hypothetic paralogous segments. **c** Sequences observed today in Hsa (number 6: chromosome 6), Mmus (numbers 7 and 13: chromosome 7 and 13) and Mdom (number 2: chromosome 2) obtained after speciation events. Colour code is the same as in previous figure. Presence of segments with same colour in different species highlights presence of orthologs. *Arrows* between the two Hsa segments represent paralogs relationships; it is a summary of duplications that occurred between birds/mammals separation and eutherians/methaterians separation. *Purple arrows* represent ORs gene subfamily 2 paralogy links, *grey arrows* represent B30.2 genes family paralogy links, the *light purple arrow* represents a MOG gene family paralogy link, and the *blue arrow* represents the MHC class I gene family paralogy link

presented in Fig. 4b are not common. Nevertheless, we can do analogies between this figure and the results we obtained. Ancestral species A could be the common ancestor of amniotes. Ancestral species A1, and A2 could be common ancestor of mammals and birds, respectively. A1a, A1b and A1c are *H. sapiens*, *M. musculus* and *M. domestica*, respectively; A2a is *G. gallus.* However, the *G. gallus* MHC region is located on a micro-chromosome that undergoes many rearrangements and does not permit us to reconstruct the region of interest.

Discussion

As a result of phylogenetic and genes distribution studies about the two chromosomal regions [Chr6, 30.4 Mb (TRIM39)—28.98 Mb (TRIM27); Chr6, 28.1 Mb (OR1F12)—26,07Mb (TRIM38)], we can reject random distribution hypothesis of duplicated genes under investigation. We suggest that a co-duplication (location and period) or "en bloc" duplication happened, and therefore, we used genomic data from *M. domestica* (opossum) and *M. musculus* (mouse) genomes and we observed that orthologs of human genes are organised in clusters with a conserved synteny between these species. "En bloc" duplication probably occurred before mammals/marsupials separation. Another possibility is that independent duplication occurred from the same region in the ancestral mammalian genome and that each duplicate localised in the same region forming, a new cluster.

Most analyses of paralogous regions today are done on regions engendered by polyploidization (see for example Dehal and Boore 2005). In these studies, "en bloc" duplications are analysed as stigmata of polyploidization followed by rearrangements, but discrimination between regional and whole genome duplications is confused. Few reports focused on regional duplications, and the studies were specific to a restricted scope. They focused mainly on primate lineage (Jiang et al. 2007) corresponding to recent events compared to the ones analysed in the present reports. It has to be noted that several studies have reports primate specific bloc duplication within the MHC. Here, the duplication block include MHC class I gene and pseudogene and surrounding non-related sequences whose function is uncertain (see for a review Dawkins et al. 1999).

In our study, the duplication probably concerned a regional duplication because no polyploidization events are known to have occurred in mammalian lineage (Hillier et al. 2004). We showed "en bloc" duplication for one region, and it would be interesting to make this kind of analyses on the whole genome. Concerning this point, it should be noted that Dehal and Boore (2005) highlighted 80 blocks of duplications that occurred after the bony fish–tetrapods split. Phylogenetic analyses did not provide precise dates because they used only four species, human/mouse, Fugu and Ciona that are remote species, so they do not allow dating precisely duplication events. It would be important to reanalyse these data using the protocol proposed in this article to show that these segments evolved indeed by regional duplication.

In our study, we had the opportunity to analyse the genes retained after local duplication. We observed that only environmental multigenic families' genes (i.e. ORs) were present in the segments we studied. The genes with such functions are known not to be gene-dosage-sensitive. Therefore, these observations agree with the gene dosage hypothesis (Papp et al. 2003). Moreover, for the OR family, it was shown that in a cell (neuron for ORs), a single paralog is expressed (Reed 2000). Therefore, there is no isotype competition.

Furthermore, we note that some genes in the duplicated regions could have undergone positive selection [i.e. ORs (Krautwurst et al. 1998) or MHC class I (Hughes and Nei 1989)]. The fact that these genes evolved under positive selection has allowed higher probability of the fixation of these genes (Lynch et al. 2001), and this could permit fixation of the entire segments. This mechanism is well described for point mutations (see for example Kim and Stephan 2002), a phenomenon known as the hitchhiking effect. Our results led us to introduce the hitchhiking concept applied to regional duplications. The impact of such mechanism could be very important because it permits fixation of duplicates that would have been lost if not linked to a gene under positive selection.

Acknowledgements We thank Philippe Monget for discussion, Olivier Chabrol for his help with bioinformatics, Anne Grimaldi and Sophie Roetynck for having initiated this work. This work was supported by the ANR program no. ANR-07-BLAN-0054-01.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7:R43

Cui L, Wall PK, Leebens-Mack JH et al (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738–749

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:e314

Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, Martinez P, Kulski J (1999) Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. Immunol Rev 167:275–304

Force A, Lynch M, Bryan FB, Pickett M, Yan Y, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 16:805–814

Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EG (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. BMC Bioinformatics 6:198

Henry J, Mather IH, McDermott MF, Pontarotti P (1998) B30.2-like domain proteins: update and new insights into a rapidly expanding family of proteins. Mol Biol Evol 15:1696–1705

Hillier LW, Miller W, Birney E et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716

Horton R, Wilming L, Rand V et al (2004) Gene map of the extended human MHC. Nat Rev Genet 5:889–899

Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci 86:958–962

Hughes AL, Yeager M, Elshof AET, Chorney MJ (1999) A new taxonomy of mammalian MHC class I molecules. Immunol 20:22–26

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet 39:1361–1368

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777

Krautwurst D, Yau KW, Reed R (1998) Identification of ligands for olfactory receptors by functional expression of a receptor library. Cell 95:917–926

Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392:917–920

Lundin LG (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16:1–19

Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genomics 3:35–44

Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. Genetics 159:1789–1804

Malik HS, Henikoff S (2003) Phylogenomics of the nucleosome. Nat Struct Biol 10:882–891

McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. Nat Genet 31:204–205

Otto SP, Yong P (2002) The evolution of gene duplicates. Adv Genet 46:451–483

Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197

Reed RR (2000) Regulating olfactory receptor expression: controlling globally, acting locally. Nat Neurosci 7:638–639

Rubin GM, Yandell MD, Wortman JR et al (2000) Comparative genomics of the eukaryotes. Science 287:2204–2215

Tazi-Ahnini R, Henry J, Offer C, Bouissou-Bouchouata C, Mather IH, Pontarotti P (1997) Cloning, localization, and structure of new members of the butyrophilin gene family in the juxta-telomeric region of the major histocompatibility complex. Immunogenetics 47:55–63

Yap MW, Nisole S, Lynch C, Stoye JP (2004) Trim5 protein restricts both HIV-1 and murine leukemia virus. Proc Natl Acad Sci 101: 10786–10791