Etude de la localisation à grande échelle de la machinerie de transcription de classe III, et de sa relation avec le facteur de transcription TFIIS dans les cellules souches embryonnaires de souris

Lucie Carriere

UNIVERSITE DE PARIS-SUD U.F.R. SCIENTIFIQUE D'ORSAY
Ecole doctorale "Gènes, Génomes, Cellules"

THESE

Présentée par

Lucie Carrière

Pour obtenir le grade de Docteur de l'Université PARIS XI Orsay

Etude de la localisation à grande échelle de la machinerie de transcription de classe III, et de sa relation avec le facteur de transcription TFIIS dans les cellules souches embryonnaires de souris.

soutenue publiquement le 29 septembre 2011

Jury

Dr Irwin Davidson :	Rapporteur
Pr Giorgio Dieci :	Rapporteur
Pr Pierre Capy :	Examinateur
Pr Martin Teichmann :	Examinateur
Dr Michel Werner :	Directeur de thèse

Remerciements

Partie importante sinon essentielle, prendre le temps de remercier tous celles et ceux que nous avons pu rencontrer, côtoyer, qui ont été acteurs tant scientifiques, ou amicaux, qui ont tout autant façonné notre thèse. La thèse n'est pas seulement un défi scientifique, elle constitue surtout une aventure humaine. Quatre ans d'une vie, c'est peu, mais j'en sortirai, comme de toute expérience, mais particulièrement ici, décapée, grandie, ayant acquis de nombreuses compétences, tant dans le domaine de la science, que dans l'apprentissage de l'autre, par la communication, l'écoute attentive, le dépassement.

Un grand merci tout d'abord aux membres de mon jury de thèse, qui ont accepté d'évaluer ce travail, le Dr Irwin Davidson de l'IGBMC à Strasbourg, le professeur Pierre Capy, de l'université Paris XI, qui me connaît un peu au travers de ces rendez-vous annuels. Le Professeur Martin Teichmann, mon professeur en master à Bordeaux (c'est loin déjà !), et le professeur Giorgio Dieci, que j'ai eu le plaisir de rencontrer l'été dernier au Congrès Odd Pols.

Michel, merci d'avoir accepté de m'accueillir dans votre laboratoire à mon arrivée du Canada. Pari à prendre, projet ambitieux, nouvel organisme, reconversion à la bioinformatique, vous seul pouvez mesurer les changements survenus, le chemin parcouru. Merci d'avoir attendu toujours avec patience que je prenne enfin mon projet en main, que je fasse le saut (un gouffre pour moi) vers la bioinformatique. Merci d'y avoir toujours cru, très souvent pour deux…

Merci à mon super formateur Sébastien, post-doc qui m'aura accompagné toute ma première année, je crois que je n'aurais jamais autant avancé qu'en travaillant avec toi. Merci pour ton idée lumineuse, un 14 février 2008, dans le TGV, qui aura permis au projet de décoller.

Merci au laboratoire de Michel Werner, Julie Soutourina pour tous ses conseils, ces discussions sur TFIIS, son bureau toujours ouvert. Yad Ghavi-Helm, avec qui j'ai eu le plaisir de travailler quelques temps, merci de m'avoir transmis le projet TFIIS. Helen Neil-Bernet, avec qui je peux partager avec joie mes maigres connaissances bioinformatiques, Claire Boschiero, Fanny Eyboulet, Camille Cibot, votre bonne humeur, et votre accueil dans ce bon groupe de filles !

Olivier Alibert, à Evry, qui a dû me voir arriver avec quelque inquiétude, il y a plus d'un an maintenant, ne sachant pas ce qu'était Unix, un terminal, un serveur, alors ne parlons même pas des commandes les plus simples d'un bash. Merci d'avoir toujours su me traduire avec une très grande clarté, précision, et surtout une grande patience, les scripts et autres outils de bioinformatiques. Je ne parle pas encore couramment unix ou perl, mais peut-être un jour, qui sait, maintenant que je suis en selle.

Merci à Matthieu Gérard de m'avoir hébergé durant toute ma thèse, d'avoir tout partagé, protocoles, connaissances des cellules ES, paillasse, bureau, réunion, barbecue, plus qu'hébergée…

Merci à ses deux supers collaboratrices, Hélène et Sylvie, pour leur attention, leur aide, leurs histoires, pour être venue me chercher du fond de mon labo, et m'avoir créé une place dans leur bureau, quand soudain tout était déserté, et que l'ordinateur était devenu mon seul compagnon. Merci à Hélène jamais à court d'idées quand il s'agit de bien rigoler et pour sa belle petite Lucie, … Merci à Sylvie pour son attention quasi-maternelle, son écoute, nos quart-d'heure philosophiques…

Merci à Arnaud pour sa force de caractère, sa foi dans la science, son dévouement, toujours prêt à venir pour vous le WE, ou à finir les manips des autres, lorsqu'il est trop tard.

Merci Isabelle, pour ce puits de savoir, bien loin de ne se limiter qu'à la science, pour toutes ses histoires, merci de m'avoir appris que c'est en s'organisant que l'on avance…

Merci Sam d'être passé tous les jours, toute la durée de ma rédaction, scrutant les premiers signes de stress, de tout envoyer f.… en l'air, merci d'avoir toujours été là.

Et Momo, je ne t'ai pas oublié ! Qu'aurait été ma thèse sans ta présence, tes mails, ton enthousiasme communicatif, tes chutes dans le bac à vaisselle, tes découvertes musicales passées en boucle… tes superbes Western, nos bouts de conduite Saclay-Versailles, dis ? T'as jamais eu peur de ma conduite, Hein ? Merci d'être toute donnée, merci pour cette franche amitié.

Merci à tous ceux de l'animalerie, Patrick, quand reviens-tu manger avec nous ? Marine, son éternelle bonne humeur, et inaltérable optimiste, qui aura préféré la liberté à un CDI, Anne-Sophie, pour sa présence, pour ses passages le matin, pour s'enquérir de l'humeur, de la forme, et pour me faire partager ses projets, Sylvain, merci pour tes conseils, comment devenir propriétaire, promis, si un jour la science m'offre un CDI, j'oserai aller voir le banquier ! Et Jean-Charles, toujours là, qui derrière un air faussement bourru, très attentionné, d'une grande gentillesse.

Merci à notre super ingénieur sécurité Stéphanie, son inquiétude réelle pour notre santé, cachée derrière de bonnes enguelades…sans oublier tous les potins, tout ce qui se passe aux bâtiments 142 et 144 !

Chantal, Catherine, merci d'avoir toujours été là pour me faciliter la vie avec toutes ces procédures administratives…moi qui suis décontenancée devant le plus simple formulaire.

Et plus personnellement, merci aux Leroux de m'avoir accueillie chez eux, nourrie, logée, blanchie, et considérée surtout comme une des leurs, merci de m'avoir ouvert tout grand les portes de leur famille. « Oui, oui, le mois prochain, je prends un appart » je ne jurerais plus jamais.

Merci à tous mes amis pour leur présence et leur soutien tout au long de ma thèse, alors que je n'étais pas souvent présente pour eux.

Et, enfin, je ne serais pas là, et je n'en serais pas là sans ma famille, mes parents qui m'ont tout donnée, leur amour, leur confiance, leurs encouragements, leur constante présence derrière moi malgré la distance, et à tous mes frères et sœurs (je ne les citerai pas tous, ma thèse est déjà bien assez longue…).

Bon, je n'y résiste pas…Marie-Adeline, François-Xavier, et Pauline, Mathilde, Jean-Baptiste, Camille et Rémi, Anne-louise, Clarisse, Paul-Antoine, Adrien, et Pierre-Jean. Merci d'être toujours là pour moi.

Demandez et l'on vous donnera ;
Cherchez et vous trouverez ;
Frappez et l'on vous ouvrira.

Saint Luc, 11, 8-9

TABLE DES MATIERES

TABLE DES ILLUSTRATIONS

Abréviations

ADN: Acide DésoxyriboNucléique

ADNc : ADN complémentaire

ARN: Acide RiboNucléique

ARNm: ARN messager

ARNr: ARN ribosomal

ARNt: ARN de transfert

miARN : micro ARN

ncARN : ARN non-codant (non coding RNA)

snARN : petit ARN nucléaire (small nuclear ARN)

ChIP: Immunoprécipitation de chromatine (Chromatin ImmunoPrecipitation)

ChIP-chip : Immunoprécipitation de Chromatine analysée par hybridation sur puce à ADN

ChIP-seq : Immunoprécipitation de chromatine suivi du séquençage Haut-débit

CTD: Domaine carboxy-terminal de Rpb1 (Carboxy Terminal Domain)

CTCF : CCCTC-binding factor

COC : Chromosome-Organizing Clamps

dNTP: désoxyriboNucléoside TriPhosphate

ddNTP: didésoxyriboNucléoside TriPhosphate

DPE: Elément du promoteur aval (Downstream Promoter Element)

DSE: élément de séquence distal (Distal Sequence Element)

DSIF: DRB Sensitivity-Inducing Factor

ELL: eleven-nineteen lysine-rich in leukaemia

ES: cellule souche embryonnaire (Embryonic Stem cell)

ETC : Extra-TFIIIC loci

GA : Genome Analyser

GTF: Facteur Général de Transcription

HA: Human influenza hemagglutinin

HAT: Histone Acetyl-Transférase

HDAC: Histone DéACétylase

ICR: région de contrôle interne (Internal Control Region)

IE: élément intermédiaire (Intermediate Element)

Inr : Initiator

<int> : integer, nombre entier

MEF: fibroblaste embryonnaire murin (Mouse Embryonic Fibroblast)

NELF: Negative ELongation Factor

NGS : Next Generation Sequencing

NTP: Nucléoside TriPhosphate

nt : nucléotide

ORF: Phase ouverte de lecture (Open Reading Frame)

pb : paires de base

PCR: Polymerase Chain Reaction

PBP : PSE-binding Factor

PIC: Complexe de préinitiation (PreInitiation Complex)

PSE: élément de séquence proximal (Proximal Sequence Element)

P-TEFb: Positive Transcription-Elongation Factor-b

PTB : PSE-Transcription Factor

RNAP I, II, III : ARN Polymérase I, II, ou III

RNA-seq : RNA sequencing

RNP : Ribonucléoprotéine

RT-PCR: Reverse Transcription Polymerase Chain Reaction

SBS : Sequencing-by-synthesis

SINE: Short Interspersed repeated DNA elements

SMC : Structural Maintenance of Chromosome

SNAPc : snRNA activating protein complex

SNP : Polymorphisme d'un seul nucléotide, ou « single-nucleotide polymorphism

snARN : petit ARN nucléaire (small nuclear RNA)

snoARN: petit ARN nucléolaire (small nucleolar RNA)

shARN : small hairpin ARN

snRNP : petite particule ribonucléique (small nuclear RNP)

TAF: Facteurs associés à TBP (TBP Associated Factors)

TAP: (Tandem Affinity Purification)

TBP: Protéine de liaison à la boîte TATA (TATA-Binding Protein)

TES : Transcription End Site, coordonnée génomique annotée comme étant la fin de la transcription du gène.

TSS : Transcription Start Site, coordonnée génomique annotée comme le début de la transcription du gène.

Définitions

ChIP : Technique utilisée pour identifier les séquences régulatrices potentielles ou les sites de liaison d'une protéine, en isolant la chromatine par une immunoprécipitation avec un anticorps dirigé contre le facteur étudié.

chrN_random : Les premières tables de gènes comportaient des séquences, auxquelles n'a pu être réattribuée une localisation chromosomique dans la nouvelle version du génome.

Complexe d'initiation : Assemblage de l'ARN polymérase et des facteurs de transcription associés, liés au promoteur.

Enhancer : Séquence située en amont ou en aval des gènes, activant la transcription.

Euchromatine : Domaine de la chromatine généralement « ouvert » regroupant les gènes densément transcrits.

Gènes de classe I, II, III : Gènes respectivement transcrits par les ARN polymérase I, II et III.

Genome table : Table listant pour le génome de référence l'ensemble des chromosomes (ou des contigs) suivi leur taille en pb. Il est à noter que les noms des chromosomes doivent être les mêmes dans cette table et dans le fichier des lectures.

Définition:

<chromosome> <size in bp>

Hétérochromatine : Domaine densément compacté de la chromatine la rendant moins accessible aux facteurs de transcription. Certaines parties du génome sont constituées d'hétérochromatine constitutive, comme les centromères, les télomères, tandis que d'autres régions sont densément empaquetées et réprimées seulement dans certains types cellulaires, ou à un moment T (hétérochromatine facultative). La chromatine constitutive est généralement constituée de séquences répétées.

Isoaccepteur : ARNt avec un anticodon distinct pour un acide aminé particulier

Ilot CpG : Séquence d'au moins 200 pb, avec un nombre de sites CpG plus élevés que le taux attendu en GC. Ces régions peuvent être méthylées indiquant leur répression transcriptionnelle.

Initiator : Séquence consensus des promoteurs de classe II, YYANWYY, où A est le TSS, N, n'importe quel nucléotide, W, une purine, et Y une pyrimidine. Cette séquence aide au recrutement de la machinerie de transcription au promoteur.

Lecture: Séquence issue du séquençage haut-débit, parfois nommé tag dans la littérature.

Méthylation de l'ADN : Modification épigénétique de l'ADN, où un groupement méthyl peut être ajouté ou enlevé sur un carbone de la cytosine.

Next Generation Sequencing : Séquençage à haut débit réalisé sur les plateformes tels que Illumina/Solexa Genome Analyser, Roche/454 Genome Sequencer, et Applied Biosystems SOLiD, ainsi que sur de nouvelles plateformes comme Helicos…

RNA-sequencing : L'ARN isolé est séquencé par NGS, après conversion en ADNc.

Phred score : score utilisé pour caractériser la qualité d'une séquence.

Profondeur de séquençage : Le nombre total de bases séquencées et alignés sur le génome de référence.

Séquençage Sanger : synthèse d'un brin d'ADN complémentaire par une ADN polymérase en présence une amorce spécifique, de dNTP, et de ddNTP fluorescents. L'addition d'un ddNP bloque la synthèse du brin complémentaire. Après électrophorèse sur capillaire, la fluorescence est détectée, indiquant quel est le nucléotide.

Silencer : Séquence d'ADN liant des facteurs de transcription dits répresseurs, qui peuvent influencer négativement la transcription, en empêchant le recrutement de la machinerie de transcription, ou en recrutant des complexes de modification des histones, créant des structures répressives au niveau de la chromatine.

TATA box : Séquence consensus enrichie en résidu Thymine et Adénine, importante pour le recrutement de la machinerie de transcription, au niveau de certains promoteurs.

INTRODUCTION

Plus de 40 ans se sont écoulés depuis la découverte de trois ARN Polymérases (RNAP), dirigeant la transcription des gènes chez les eucaryotes (Roeder and Rutter, 1969). L'ARN Polymérase I transcrit le long précurseur des ARN ribosomaux (ARNr), le 45S chez l'humain, maturé en trois ARNs le 5,8S, le 18S et le 28S, qui seront incorporés dans les ribosomes, particules ribonucléiques, sièges de la traduction. L'ARN Polymérase II transcrit les ARN messagers, et de nombreux petits ARNs non-traduits, comme les microARNs régulant l'expression des ARNm, les petits ARNs nucléolaires (snoARNs, Small Nucleolar RNA), les petits ARN nucléaires (snARNs, small nuclear RNA). L'attention des chercheurs s'est tout naturellement focalisée sur la transcription de classe II, car les messagers transcrits codent pour les protéines, constituant l'essentiel de la masse cellulaire. Pourtant, la transcription réalisée par l'ARN Polymérase III semble finalement receler de plus de complexité qu'initialement pensé. En plus de la transcription des ARN de transfert, la RNAP III est responsable de la synthèse d'un ensemble assez hétérogène de petits ARNs, intervenant à tous les niveaux du métabolisme cellulaire, comme la traduction, l'épissage, la maturation des ARNs, la régulation de la RNAP II, et de nombreux autres dont le rôle n'est pas clairement défini. De plus, l'augmentation de la transcription de classe III est mise en cause dans le processus d'oncogenèse. Enfin, de nombreux liens ont été établis entre la transcription de classe II et de classe III. Le facteur de transcription TFIIS, à l'origine identifié comme impliqué dans la transcription de classe II, joue un rôle dans la transcription de classe III chez la levure *Saccharomyces cerevisiae*. La proximité de la RNAP II des gènes liés par la RNAP III soulève encore de nombreuses questions sur l'influence que ces deux classes de transcription peuvent exercer l'une sur l'autre.

La plupart des transcrits de la RNAP III ont été découverts par des études se basant sur la relative sensibilité de la RNAP III à l'α-amanitine. Ces dernières années, la révolution des approches génomiques a permis l'évaluation systématique des cibles de la RNAP III. La localisation de la RNAP III et de ses facteurs de transcription a été réalisée tout d'abord par des études d'immunoprécipitation suivie d'hybridation sur puces à oligonucléotides (ChIP-chip) chez la levure. La taille et la complexité des génomes des eucaryotes pluricellulaires limitaient ce type d'étude. Pourtant, l'année dernière, plusieurs études indépendantes ont été publiées, décrivant le transcriptome de classe III, ainsi que l'ensemble des sites de liaison de ses facteurs de transcription dans plusieurs lignées humaines. Ces approches ont été possibles grâce au développement de techniques de séquençage à haut-débit, couplées à l'immunoprécipitation de chromatine (ChIP-seq). Cette technique permet d'explorer avec précision la liaison au génome d'une protéine d'intérêt.

Les résultats de ces études ont premièrement confirmé ce qui avait été découvert par des approches génétiques et biochimiques. Elles ont de même souligné la conservation des mécanismes de la transcription de classe III de la levure à l'homme. Sont également conservés des sites mis en évidence chez la levure *S. cerevisiae*, uniquement liés par le facteur de transcription TFIIIC, en l'absence de la

RNAP III et de TFIIIB. Ces sites ont été impliqués dans l'organisation de l'architecture de la chromatine, ajoutant encore un degré de complexité aux rôles de la RNAP III. Enfin, de nouvelles questions surgissent comme celle de la régulation de l'expression des gènes de classe III en fonction du tissu ou du stade de développement.

Notre étude s'est tout d'abord focalisée sur la machinerie de transcription de classe III. Je présenterai donc dans un premier temps les connaissances actuelles de la transcription de classe III, et les rôles de ses facteurs de transcription. Avec le facteur d'élongation de la RNAP II, TFIIS, notre étude s'est orientée vers la transcription de classe II, notamment sur le phénomène de pause au promoteur, caractérisé chez de nombreux organismes. La transcription de classe II et la fonction de TFIIS seront abordées. Un dernier chapitre présente les techniques actuelles appliquées à l'analyse d'une expérience d'immunoprécipitation de chromatine, suivi d'un séquençage haut-débit.

Chapitre I : La transcription de classe III

A. Description des transcrits de classe III

La RNAP III transcrit de nombreux petits ARNs, de fonctions assez hétérogènes, mais tous non-traduits, et pouvant avoir une fonction catalytique. Les ARNs de classe III sont très souvent incorporés dans des complexes ribonucléiques ou RNPs. Ces ARNs sont généralement codés par des familles multigéniques. Ils sont issus de nombreuses unités de transcription répétées à travers le génome. Ces gènes ont été amplifiés dans les génomes eucaryotes par des mécanismes de rétrotransposition.

Les descriptions des gènes transcrits par la RNAP III s'appuient sur les revues de RJ.White (White, 1998) et de G. Dieci (Dieci et al., 2007), ainsi que sur d'autres références spécialisées.

L'ARN 5S

L'ARN 5S, long d'environ 120 nt, est l'unique ARNr non transcrit par la RNAP I. Les gènes sont souvent regroupés dans des unités répétées, parfois contigus aux gènes des ARNr, comme chez *Saccharomyces cerevisiae*. Chez l'humain, ces gènes sont organisés de façon indépendante, en tandem, ou en simples copies dispersées (Haeusler and Engelke, 2006).

Les ARNt

Les ARNt ont une taille comprise entre 75 et 95 nt, et sont des molécules ubiquitaires. La RNAP III transcrit l'ensemble des ARNt nucléaires. Ces ARNs fonctionnent comme adaptateurs, entre le codon d'un ARNm et l'acide aminé correspondant, à incorporer dans la chaîne polypeptidique naissante. La biogenèse des ARNt inclut la transcription du gène, suivi du clivage de l'extrémité 5' par la RNase P, tandis que l'extrémité 3' est clivée par la RNase Z. A cette extrémité 3', est ajouté le trinucléotide CCA, les introns si présents sont épissés, et plusieurs résidus sont modifiés. Les ARNt sont ensuite exportés du noyau et utilisés lors de la traduction (Phizicky and Hopper, 2010). Tous les ARNt sont caractérisés par une forme en feuille de trèfle, arborant trois boucles et un bras supplémentaire éventuel. Une des boucles porte l'anticodon, site d'appariement de trois nucléotides complémentaire du codon de l'ARNm. Les ARNt sont chargés à leur extrémité 3' CCA de l'acide aminé correspondant (Goodenbour and Pan, 2006). 21 familles d'ARNt isoaccepteurs (ARNt avec un anticodon différent pour un même acide aminé) existent, codant les 20 acides aminés, et un gène d'ARNt codant la sélénocystéine. Une famille d'isoaccepteur peut contenir de un ($ARNt^{Trp}$) à cinq membres ($ARNt^{Leu}$). Chez les bactéries et les

eucaryotes, l'abondance des ARNt isoaccepteurs est corrélée à la préférence des codons parmi les gènes les plus exprimés, comme les gènes des protéines ribosomales. Ce phénomène est appelé le biais d'usage de codon (Goodfellow et al., 2008). Il existe une certaine diversité au sein des gènes d'ARNt isoaccepteurs. Le nombre de gènes d'ARNt ayant le même anticodon, mais une séquence différente dans le reste du gène (gène d'ARNt isodécodeur), est très variable selon les espèces. Le génome de la levure présente assez peu d'isodécodeurs, tandis que ces gènes sont très nombreux chez les mammifères. Les changements de séquences des gènes d'ARNt isodécodeurs sont soumis aux contraintes de structure secondaire et tertiaire de l'ARNt mature. Ces variations peuvent conférer des rôles uniques à ces différents isodécodeurs.

Plusieurs programmes de détection des ARNt existent, comme tRNA-scan_SE (Lowe and Eddy, 1997), ou Aragorn (Laslett and Canback, 2004). Pourtant jusque récemment, l'ensemble des gènes codant les ARNt chez les mammifères était assez mal caractérisé. Ce manque d'information était principalement dû à des obstacles techniques. L'analyse de l'expression des gènes de classe III pris individuellement, se heurte à des problèmes comme la structure secondaire et tertiaire, et la forte similarité de séquence des gènes d'ARNt d'une même famille. Chez la souris, dans une étude menée par Coughlin (Coughlin et al., 2009), une base de données issue des prédictions réalisées par deux programmes Aragorn et tRNA-scan-SE, a été constituée. L'expression de ces gènes a ensuite été vérifiée par puce transcriptomique. Pourtant, en raison de la trop grande similarité de séquences existant au sein des familles de gènes, les auteurs n'ont pu conclure que seul un sous-ensemble de chaque famille était exprimé (Coughlin et al., 2009).

L'ARN U6

Le splicéosome est un complexe ribonucléique (RNP, Ribonucleoprotein), contenu dans le noyau des cellules eucaryotes, responsable de l'épissage des pré-ARNm. Il contient environ 250 protéines et cinq types de snARNs, quatre d'entre eux sont transcrits par la RNAP II, le cinquième, l'ARN U6, est transcrit par la RNAP III. L'ARN U6 s'associe avec le snARN U4

L'ARN de la RNase P

Cet ARN est codé par l'unique gène H1. Cet ARN est retrouvé chez les bactéries, les archébactéries et les eucaryotes. L'ARN H1 est inclus dans le complexe de la RNase P, endoribonucléase qui clive l'extrémité 5' des pré-ARNt, libérant une extrémité 5' phosphate, et 3'-OH (Esakova and Krasilnikov, 2010). Ce complexe est responsable de la maturation d'autres petits ARNs, comme le 4,5S, ou des snoARNs. Le complexe de la RNase P est également impliqué dans la transcription de classe I et III. Ce complexe se lie aux gènes de classe III activement transcrits. De plus, sa dérégulation entraîne une diminution de la transcription de ces deux classes (Esakova and Krasilnikov, 2010).

L'ARN de la RNase MRP

Cet ARN appartient au complexe de la RNase MRP ou Mitochondrial RNA Processing. La RNase MRP a été identifiée à l'origine dans la mitochondrie, où elle génèrerait des amorces ARN, requises pour la réplication du génome mitochondrial. Pourtant, ce complexe est en réalité localisé en majorité au niveau du nucléole, où il intervient lors de la maturation du long précurseur des ARNr. Contrairement à l'ARN H1, dont il serait issu, l'ARN MRP n'est présent que chez les eucaryotes. Ce gène est transcrit par la RNAP III chez les mammifères. Sa transcription dépend de la RNAP II, chez la levure *S. cerevisiae* (Dieci et al., 2009; Esakova and Krasilnikov, 2010).

L'ARN 7SL

En dépit de sa longueur inhabituelle pour un gène de classe III (522 pb chez *S. cerevisiae*), le gène de l'ARN 7SL est transcrit par la RNAP III (Dieci et al., 2002). Cet ARN forme l'échafaudage de la SRP ou Signal Recognition Particle, qui au cours de la traduction guide le polypeptide naissant, vers la membrane du réticulum endoplasmique dans laquelle il sera inséré. SRP reconnait et lie au cours de la traduction, une séquence signal de sécrétion présente sur les polypeptides naissants, destinés à la membrane cytoplasmique. SRP inhibe alors la traduction jusqu'à ce que le polypeptide soit inséré entièrement dans le réticulum endoplasmique.

L'ARN 4,5S

L'ARN 4,5S est un petit ARN nucléaire de 94 nt. Il est présent chez la souris, le rat et le hamster, mais est absent chez l'humain, ou le poulet (Koval and Kramerov, 2009). Sa fonction est inconnue, il a cependant été observé interagissant avec les ARN polyadénylés et d'autres protéines.

ARN Vault

Les complexes Vault sont de très grosses particules RNP, présentes dans le cytoplasme. Leur structure ressemble aux voûtes des cathédrales, d'où le nom qui leur a été attribué. Ils ont été identifiés chez différents eucaryotes, comme les poissons, les amphibiens, les mammifères. Cependant, les protéines constituant le complexe Vault n'ont pas été identifiées chez la levure *S. cerevisiae*, ni chez la drosophile ou *Caenorhabditis elegans.* Ils ne seraient pas non plus présents chez les plantes. Les complexes Vaults sont composés de plusieurs copies de trois protéines et des petits ARNs vault (vARN), de 88 à 141 nt. Ils seraient au nombre de trois chez l'humain, tous retrouvés associés à la particule vault (Mossink et al., 2003). La fonction des RNP est assez mal définie. Ces complexes ont été impliqués dans la résistance à plusieurs drogues. Ils seraient de plus requis dans l'assemblage et /ou le transport des macromolécules. La

forte conservation des protéines constituant ce complexe implique cependant qu'il ait une fonction importante.

ARN Y

Les ARNs Y, d'environ 100 nt, sont des ARNs très structurés. Ils sont conservés parmi l'ensemble des vertébrés. Quatre ARNs Y existent chez l'humain (hY1, hY3, hY4 et hY5). Les ARNs Y ont tout d'abord été identifiés comme composants des particules ribonucléiques Ro (Ro RNP). Les Ro RNP sont des complexes solubles retrouvés chez les vertébrés. Les Ro RNP sont constitués de l'ARN non codant Y, associé aux protéines Ro60, La, et d'autres protéines moins caractérisées. La fonction des Ro RNP n'est pas clairement établie. Les ARNs Y exercent également un rôle en dehors des Ro RNP, notamment au cours de l'initiation de la réplication des chromosomes (Christov et al., 2006).

L'ARN 7SK

L'ARN 7SK est très bien conservé parmi les vertébrés. Comme certains transcrits de classe III, l'ARN U6, ou le SINE B2, l'extrémité 5' de l'ARN contient un groupe méthyl (CH3pppN), contrairement aux autres transcrits de classe III, qui ont habituellement une extrémité libre triphosphate (pppN). Cette modification les protège de la dégradation mais est incompatible avec la traduction (Diribarne and Bensaude, 2009). L'ARN 7SK s'associe tout d'abord avec les protéines HEXIM1/2, LARP7 et BCDIN3. Ce complexe serait recruté au niveau du gène au cours de l'élongation (Prasanth et al., 2010). Ce complexe peut ensuite interagir avec le facteur positif d'élongation b (P-TEFb, Positive transcription Elongation Factor b). Les protéines HEXIM inhibent l'activité kinase de la sous-unité Cdk9 de P-TEFb, mais seulement en présence de l'ARN 7SK. L'inhibition de l'activité kinase et la séquestration de P-TEFb entraînent la répression de la transcription de classe II.

L'ARN BC1

Cet ARN est spécifique des rongeurs, il est plus particulièrement exprimé dans les neurones, et les cellules dendritiques (Martignetti and Brosius, 1995). Son homologue chez l'homme est l'ARN BC200. Il dériverait de la rétrotransposition de l'ARNt de l'alanine.

Les ARNs viraux

Plusieurs virus contiennent des gènes de classe III. Les mieux caractérisés sont les gènes VA I et VA II de l'Adénovirus. Ils sont transcrits au cours de l'infection virale, ils inhibent la protéine kinase (PKR, RNA-activated Protein Kinase), permettant la transcription des ARNm du virus.

Le génome de l'Epstein-Barr virus contient également deux gènes de classe III, EBER I et EBER II, synthétisés au cours de l'infection virale.

Les microARNs

Les microARNs (miARN) sont de petits transcrits, d'une longueur de 19 à 23 nt. Ils agissent comme ARN anti-sens régulant l'expression des gènes, principalement lors de la traduction. Les gènes de microARNs sont parfois organisés en groupe, permettant une co-régulation de l'expression de ces gènes. Un groupe en particulier a été l'objet d'une étude approfondie par Borchert et al. Ce groupe, C19M, présent chez les primates comprend 46 miARNs. Les gènes de ces miARNs résident à l'intérieur de séquences répétées de type Alu. Les éléments Alu présentent un promoteur de classe III, composé des boîtes A et B très bien conservées. Les auteurs ont proposé que l'expression du groupe de miARNs est sous contrôle de ces éléments Alu, indiquant qu'ils seraient transcrits par la RNAP III (Borchert et al., 2006). Cependant, ces données ont été remises en cause par le groupe de Cavaillé. Il a été démontré que la plupart de ces miARNs sont en réalité contenus dans des introns de gènes de classe II, et par là-même sont transcrits par l'ARN polymérase II (Bortolin-Cavaille et al., 2009).

Aucune nouvelle étude n'est encore venue étayer l'hypothèse de l'implication de la RNAP III dans la transcription des miARNs. Les études de la distribution de la RNAP III réalisées chez l'homme, n'identifient pas de miARNs. Si certains ont été inscrits comme gènes potentiels, des études plus approfondies ont révélé qu'ils étaient en réalité des gènes de classe III, mais non annotés comme tels, comme dans le cas du MIR 886, qui recouvre le gène de l'ARN Vault (MIR 886).

Les SINEs ou Short Interspersed repeated DNA Element

Le génome des eucaryotes supérieurs contient de nombreuses séquences répétées, représentant jusqu'à 45% du génome humain ou de la souris (Kramerov and Vassetzky, 2005). La plupart de ces répétitions sont générées par rétrotransposition d'éléments, qui insèrent une copie à de nouvelles localisations dans le génome. Les SINEs sont des rétrotransposons de classe I, car leur réplication implique une étape intermédiaire ARN (la classe II contient les transposons de type ADN). Les classes I et II contiennent des transposons de type autonome et non-autonome. Les transposons autonomes ont un cadre ouvert de lecture (ORF), codant des protéines essentielles pour la transposition. Les transposons non autonomes comme les SINEs ne codent pas les protéines nécessaires. Leur transposition dépend donc de la machinerie codée par les transposons autonomes.

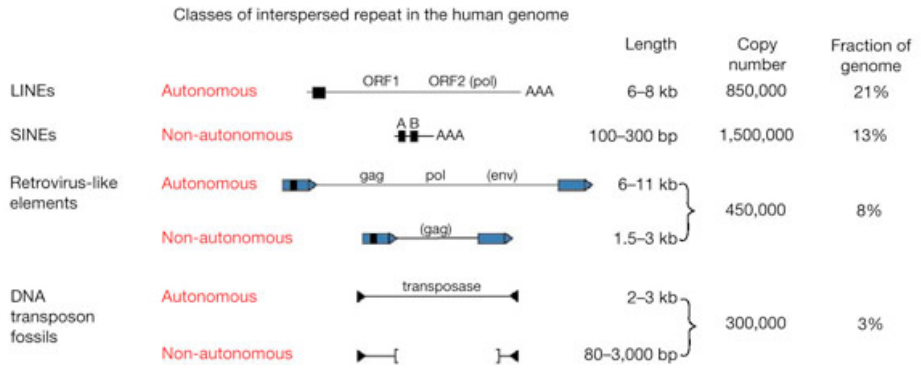


Figure 1. Eléments transposables chez les mammifères (d'après Lander et al., 2001). Ils existent deux classes transposons, les transposons ADN et les rétrotransposons, avec une étape ARN au cours du mécanisme de transposition. Ces deux classes contiennent des éléments autonomes, contenant les éléments nécessaires à leur transposition. Les éléments non-autonomes dépendent des transposons autonomes pour leur insertion dans le génome.

Les SINEs sont de petite taille, entre 50 à 500 nt. Ils sont généralement composés de trois parties, la tête correspondant à la région similaire à l'ARNt, le corps central (région non reliée à l'ARNt), et la queue 3', souvent polyadénylée, ou composée d'une simple répétition. Certains SINEs dévient de ce profil. Leur tête peut dérivée d'un autre ARN cellulaire, leur corps et leur queue 3' peuvent être très courts ou absents. Les SINEs peuvent de plus former des dimères, des trimères, ou plus, entre eux. Le mécanisme d'amplification des SINEs est lié à d'autres éléments rétrostransposables autonomes, les LINEs. Une fois transcrits, les SINEs sont exportés dans le cytoplasme, où ils sont pris en charge par la machinerie de rétrotransposition des LINEs. Leur complément ADN est synthétisé, ils sont réexportés dans le noyau, puis intégrés dans le génome.

SINE	ARN ancestral	Structure	Longueur (nt)	Espèces	Nombre de copies
B1	ARN 7SL	monomérique	135	Rongeurs	$5.6.10^5$
ID	ARNt alanine	monomérique	75	Rongeurs	10^3-10^5
B4 (RSINE2)	ARNt alanine + ARN 7SL	dimérique ID+B1	275	Souris et rat	4.10^5
B2	ARNt alanine	monomérique	185	Muridae	$3.5.10^5$
MIR	ARNt	Monomérique	~270	Vertébrés	1.10^5

Tableau 1. Liste non exhaustive des principaux SINEs présents dans le génome de la souris (d'après Kramerov and Vassetzky, 2005). MIR : Mammalian Interspersed Repeat

Les SINEs sont assez souvent regroupés, bien qu'il existe des éléments isolés. Ils ne sont pas exclusivement présents chez les mammifères. On les retrouve également dans le génome d'insectes, de reptiles, et de plantes. Chez l'humain, la famille SINE la plus connue et ayant le plus de représentants est la famille Alu. Il ne semble pas exister de SINEs chez la levure *S. cerevisiae*, ni chez la drosophile.

Les génomes eucaryotes contiennent un très grand nombre de SINEs, or peu semblent être actifs, la majorité n'est pas transcrite et ne peut donc pas être rétrotransposée. D'où provient une telle exclusion ? La réplication des SINEs implique deux processus : leur transcription et leur rétroposition. La transcription serait le facteur majeur distinguant les SINEs actifs des pseudogènes. Ainsi une copie active de SINE doit être dans un contexte génomique favorable, mais surtout avoir préservé un promoteur de classe III efficace. Les SINEs transposés récemment ne disposent pas de séquences adjacentes favorables à leur transcription. De plus, au cours des réplications successives, ils accumulent des mutations dans leur promoteur, ne permettant plus leur transcription.

Fonctions des SINEs

Considérés longtemps comme des parasites du génome, les SINEs sont pourtant associés à de nombreuses fonctions cellulaires comme la régulation des gènes de classe II, et l'expansion du génome. L'intégration de ces SINEs peut être responsable de la perturbation de l'expression des gènes adjacents, et peut même conduire à de larges réarrangements chromosomiques, par recombinaison homologue inégale. Leur insertion à proximité ou dans le promoteur des gènes de classe II entraîne inévitablement des perturbations de l'expression de ces gènes. Les SINEs peuvent contrôler la transcription des gènes, en tant qu'enhancer ou régulateur *cis*-négatifs. Un SINE B2 contient même un promoteur de classe II, dirigeant l'expression du gène Lama 3 chez l'humain (Ferrigno et al., 2001).

Enfin, les SINEs peuvent avoir une action en *trans*, en effet après un stress cellulaire, comme après un choc thermique, ou une infection virale. Une forte augmentation de la transcription des SINEs est observée en réponse à ce type de stress. D'autres mécanismes pourraient être régulés par les SINEs, comme la nucléation de l'hétérochromatine, la cohésion des chromatides, bien qu'il n'y ait pour le moment pas de démonstration formelle.

B. Organisation des promoteurs des gènes de classe III

Trois types de promoteurs sont reconnus par la RNAP III. Les promoteurs sont typiquement localisés dans la partie interne du gène, exception faite des promoteurs de type III, chez les métazoaires. Ils ont en général une structure discontinue, composée de séquences essentielles séparées par des régions

non-essentielles (Paule and White, 2000). Les séquences en amont des gènes seraient également impliquées dans la régulation de la transcription, cependant elles ne sont ni essentielles, ni conservées entre les différentes espèces (Geiduschek and Kassavetis, 2001).

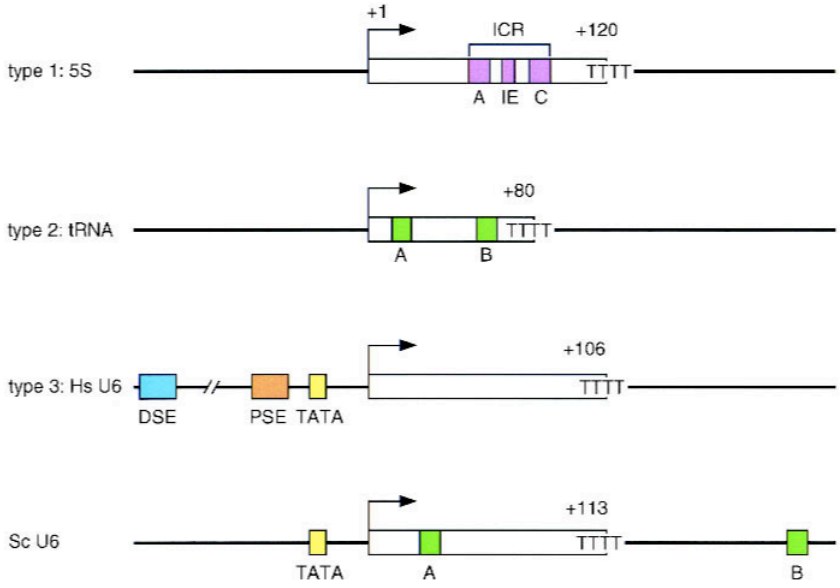


Figure 2. Description des différentes classes de promoteurs des gènes de classe III (d'après Schramm and Hernandez, 2002). Hs : *Homo sapiens*, Sc : *Saccharomyces cerevisiae*. Trois types de promoteurs ont été décrits chez les métazoaires, tandis deux types existent chez *S. cerevisiae*.

Les promoteurs de type I et II sont intragéniques. Le gène de l'ARN 5S est l'unique exemple du promoteur de type I, tandis que le promoteur de type II est partagé par les gènes des ARNt, le 7SL, les Short Interspersed Nuclear Element, SINEs). Le promoteur de type I est composé d'une boîte A, d'un élément intermédiaire IE (Intermediate Element), et d'une boîte C, l'ensemble de ces trois éléments constitue l'ICR ou Internal Control Region. Le promoteur de type II est composé des boîtes A et B, chacune d'une dizaine de paires de bases, séparées par d'une quarantaine de paires de bases, environ. Certains gènes d'ARNt chez la levure, comme chez les mammifères possèdent des introns généralement situés entre les deux boîtes. La présence de ces régions entraîne un espacement variable entre les deux boites, ces introns pouvant aller de deux à 400 nucléotides (Coughlin et al., 2009). La position de la boîte A est assez rigide, généralement située 12 à 20 nucléotides en aval du début de la transcription (TSS, Transcription Start Site). La position du TSS est fixée par rapport à la boîte A. La localisation de la boîte B est extrêmement variable, en partie due à la présence d'introns dans certains gènes d'ARNt. La RNAP

III initie préférentiellement la transcription à une purine, précédée de façon optimale d'une pyrimidine. Les boîtes A et B sont très conservées, de la levure à l'homme (Marck et al., 2006). Le consensus de ces deux séquences est :

-boîte A : TGGCNNAGTGG

-boîte B : GGTTCGANNCC

Cette conservation reflète leur importance en tant que promoteur. De plus, ces deux éléments constituent les boucles D et TψC de l'ARNt mature, requises pour leur fonction. La thymine de la boîte B en quatrième position est absolument invariable pour les gènes d'ARNt. Cette thymine est le précurseur du résidu pseudouridine du bras TψC des ARNt. Le promoteur du gène 7SL, bien que de type II, diffère du consensus type ARNt, car les séquences des éléments A et B sont dégénérées. La thymine est par exemple remplacée par une adénine. Cette variation pourrait être impliquée dans le repliement de l'ARN ou la fonction de l'ARN 7SL. Mais elle n'entraîne pas de perturbation de la liaison de TFIIIC (Kaiser and Brow, 1995).

L'organisation du promoteur de type III présente une forte diversité selon l'organisme considéré. Il dirige la transcription des gènes des ARN U6, SRP-7SK, ou l'ARNt de la sélénocystéine (Schramm and Hernandez, 2002). Chez *S. cerevisiae*, le promoteur du gène U6 ressemble au promoteur de type II ; il contient les éléments A et B, bien qu'arrangés différemment ; la boîte B se situe en aval du terminateur. Il possède également une séquence TATA, en amont du TSS. Celle-ci n'est pourtant pas indispensable à la transcription des gènes *in vivo*.

Chez les métazoaires, l'organisation du promoteur des snARN transcrits par la RNAP III tend à ressembler à celle d'un promoteur des gènes de snARN transcrits par la RNAP II (Hernandez, 2001). Les éléments régulateurs sont externes au gène, en amont du TSS. Le promoteur des snARNs de classe II contient seulement l'élément de séquence proximal (PSE, Proximal Sequence Element), entre -60 et -50 du TSS. Le PSE est un élément essentiel du promoteur, nécessaire à la sélection du TSS. Le promoteur des snARNs de classe III consiste en deux éléments le PSE, et la boite TATA, située à une distance fixe en aval du PSE. Cet arrangement est conservé chez les vertébrés et la drosophile, où le PSE est appelé le PSEA (Orioli et al., 2011a). Chez les vertébrés, le promoteur composé du PSE et de la boîte TATA est associé en plus à l'élément de séquence distal, le DSE (Distal Sequence Element), en -250. Cet élément agit en tant qu'activateur de la transcription. Les éléments DSE et PSE sont interchangeables entre les promoteurs des snARNs de classe II et de classe III. La spécificité de transcription par la RNAP III est dictée par la présence de la boîte TATA (Lobo and Hernandez, 1989; Mattaj et al., 1988). La distance séparant les éléments DSE et PSE est conservée pour les gènes de classe II et III. L'élément DSE contient

différents sites de liaisons. Il contient invariablement une séquence octamère, et en aval de cette séquence, est retrouvé l'élément SPH (Sph1 postoctamer Homology).

L'organisation du promoteur du gène de l'ARNt^{sec} diffère de celle des gènes des autres ARN de transfert. Sa transcription est sous la dépendance d'un promoteur de type III portant un DSE et un PSE. Il contient pourtant une boîte A et une boîte B. Cependant, si la transcription de ce gène est stimulée par la boîte B, la boite A est inactive, à cause d'une insertion de 2 pb.

Promoteur	Gènes
Type I	ARN 5S
Type II	ARNt, ARN 7SL, ARN VAI (Adénovirus), ARN 4,5S, SINE, ARN BC1, Vault ARN
Type III	ARN U6, 7SK, gène H1 (RNase P), ARN de la RNase MRP, ARNt sélénocytéine, HY

Tableau 2. Récapitulatif des gènes de classe III chez la souris, classés selon les types de promoteur.

C. La machinerie de transcription des gènes de classe III

L'ARN Polymérase III

Les ARN Polymérases sont des complexes de plusieurs sous-unités, partageant un ensemble de protéines homologues. La RNAP des bactéries est la plus simple ; elle est constituée des sous-unités α, présente sous la forme d'un dimère, ω, ß et ß'. Chez les archae, les sous-unités constituant le cœur de l'enzyme sont conservées. Cependant, l'enzyme contient un total de douze sous-unités ; Rpo1 (homologue de ß'), Rpo2 (homologue de ß), Rpo3 (homologue de α), Rpo4, Rpo5, Rpo6 (homologue de ω), Rpo7, Rpo8, Rpo10, Rpo11 (également homologue de α), Rpo12 et Rpo13. Les RNAP eucaryotes sont constituées de protéines homologues aux sous-unités des archae. La RNAP II est constitué de 12 sous-unités, la RNAP I de 14 sous-unités et la RNAPIII de 17 sous-unités. Chacune des sous-unités de la RNAP II est partagée ou codée par un paralogue chez les deux autres RNAP. Les RNAP I et III possèdent en plus des sous-unités spécifiques (Werner et al., 2009).

RNAP I	RNAP II	RNAP III	
sous-unités homologues ou communes *Saccharomyces cerevisiae*			Sous-unités de *Homo sapiens*
Rpa190 (A190)a	Rpb1	Rpc160 (C160, beta'-like)	RPC1/RPC155
Rpa135 (A135)	Rpb2	Rpc128 (C128, beta-like)	RPC 2
Rpc40 (AC40)	Rpb3	Rpc40 (AC40, alpha-like)	RPAC1/RPA5, RPA39
Rpa14 (A14)	Rpb4	Rpc17 (C17)	RPC9/CGRP-RC
Rpb5 (ABC27)	Rpb5	Rpb5	RPABC1/RPB5, RPB25
Rpb6 (ABC23)	Rpb6	Rpb6 (omega-like)	RPABC2/RPB6, RPB14.4
Rpa43 (A43)	Rpb7	Rpc25 (C25)	RPC8
Rpb8 (ABC14.5)	Rpb8	Rpb8	RPABC3/RPB8, RPB17
Rpa12 (A12.2)	Rpb9	Rpc11	RPC10/RPC11
Rpb10 (ABC10a)	Rpb10	Rpb10	RPABC4/RPB7.0
Rpc19 (AC19)	Rpb11	Rpc19 (AC19, alpha-like)	RPAC2/RPA9, RPA16
Rpb12 (ABC10b)	Rpb12	Rpb12	RPABC5/RPB10, RPB7.6
Sous-unités spécifiques de la RNAP-I ou RNAP-III			
Rpa34 (A34.5)		Rpc31 (C31)	RPC7/RPC32
Rpa49 (A49)		Rpc34 (C34)	RPC6/RPC39
		Rpc37 (C37)	RPC5
		Rpc53 (C53)	RPC4/RPC53
		Rpc82 (C82)	RPC3/RPC62
Nombre de sous-unités			
14	12	17	

Tableau 3. Sous-unités des ARN Polymérases eucaryotes (d'après Werner et al., 2009).
La nomenclature systématique utilisée est celle de SGD, pour les sous-unités de la levure.

Les modèles présentant la structure des ARN Polymérases révèlent la conservation de la forme globale et des principales caractéristiques, comme le sillon, le site actif, les sites de liaison des nucléotides triphosphates, et de l'hybride ADN-ARN, et le tunnel, d'où sort l'ARN naissant. La conservation de ces structures est la conséquence de la conservation entre les ARN Polymérases des mécanismes de base d'élongation. Comme pour la RNAP II, les deux plus grandes sous-unités Rpc1 et Rpc2 forment le site actif de l'enzyme, et le sillon dans lequel s'insère l'ADN.

Chaque RNAP possède un certain nombre de facteurs de transcription, les GTF (General Transcription Factor), requis lors des différentes étapes de la transcription. À l'exception de deux d'entre eux, les GTF diffèrent selon les RNAP, bien qu'ils remplissent des fonctions similaires. TBP est un des GTF commun aux trois RNAP. Le paralogue de TFIIB chez la RNAP III est la sous-unité Brf1 (TFIIB-related factor) présente dans le GTF TFIIIB (Carter and Drouin, 2009; Carter and Drouin, 2010). Les sous-unités spécifiques de la RNAP III seraient en réalité des GTF recrutés de façon permanente aux RNAP. Ce scénario s'appuie sur la comparaison des sous-unités spécifiques de la RNAP III et des GTF de

la RNAP II, mettant en évidence leur homologie. Rpc37/Rpc53 seraient paralogues du GTF TFIIF, tandis que Rpc82/Rpc34 sont paralogues de TFIIE α et TFIIE β respectivement. Cette similarité de structure et de séquence est confirmée par la conservation des fonctions des sous-complexes et des GTF (Carter and Drouin, 2010).

TFIIE est impliqué dans l'initiation de la transcription et la transition de l'initiation à l'élongation (Tanaka et al., 2009). Le sous-complexe Rpc31/Rpc34/Rpc82 situé au niveau de la pince (clamp) aurait également un rôle dans l'initiation de la transcription. Rpc34 est impliqué dans le recrutement de la RNAP III au promoteur, via son interaction directe avec Brf1, et dans la formation du complexe ouvert de transcription (Brun et al., 1997).

Bien que TFIIF et le sous-complexe Rpc53/Rpc37 soient tout deux impliqués dans l'élongation de la transcription, ils exercent des effets opposés. TFIIF, comme ses homologues de la RNAPI, RPA34.5 et RPA49 augmente la processivité (Kuhn et al., 2007). Le sous-complexe Rpc37/Rpc53 ralentit la progression de la polymérase, permettant ainsi la reconnaissance du terminateur (Landrieux et al., 2006). Récemment, Kassavetis et al. ont démontré que ce sous-complexe joue également un rôle dans l'initiation et l'élongation. Rpc37/Rpc53 sont de plus, requis en coopération avec la sous-unité Rpc11, pour la réinitiation de la transcription (Kassavetis et al., 2010).

Le sous-complexe Rpc17/Rpc25, homologue du sous-complexe Rpb4/Rpb7 de la RNAP II, constitue la tige. Il est également essentiel à l'initiation de la transcription, via la reconnaissance de Brf1 par Rpc17. Il interagit de plus avec l'ARN naissant, suggérant un rôle de Rpc17/Rpc25 dans le couplage de la transcription et du repliement de l'ARN naissant (Jasiak et al., 2006).

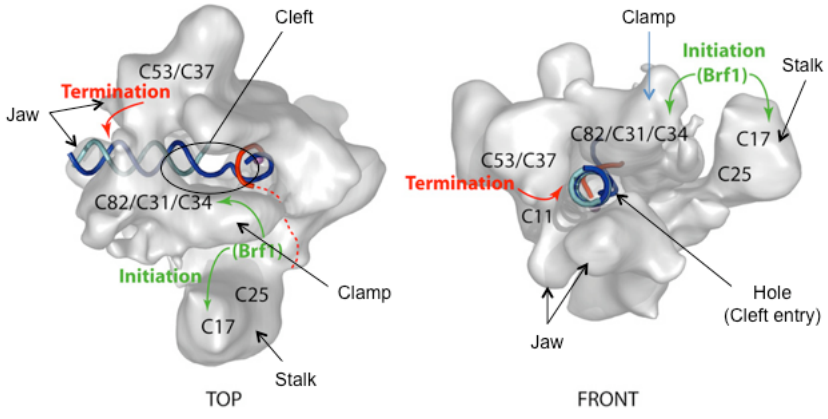


Figure 3. Modèle de la structure de la RNAP III, par cryo-électron microscopie chez *S. cerevisiae* (d'après Fernandez-Tornero et al., 2007).

Les brins codant et non-codant de l'ADN et l'ARN sont représentés en bleu, cyan et rouge respectivement. Le chemin présumé de l'ARN nouvellement synthétisé est représenté par une ligne en

pointillé. Les interactions entre les sous-unités de la RNAP III et Brf1 durant la transcription sont schématisées par des flèches.

Les facteurs généraux de l'ARN polymérase III

L'initiation précise de la transcription par la RNAP III requiert au minimum trois facteurs de transcription : TFIIIA, TFIIIB, TFIIIC.

	M. musculus RNAPIII sous-unités	*S. cerevisiae* RNAPIII sous-unités
TFIIIA	TFIIIA	
TFIIIB	TBP Bdp1 (TFIIIB150) Brf1 (TFIIIB90) ou Brf2 (TFIIIB50)	 Bdp1 (TFIIIB90) Brf1 (TFIIIB70)
TFIIIC	TFIIIC220 TFIIIC102 TFIIIC63 TFIIIC110 TFIIIC90 TFIIIC35	Tfc3 (tau 138) Tfc4 (tau 131/PCF1) Tfc1 (tau 95) Tfc6 (tau 91) Tfc8 (tau 60) Tfc7 (tau 55)

Tableau 4. Sous-unités protéiques des facteurs de transcription chez les mammifères, illustrés par la souris (Mus musculus), et chez la levure *Saccharomyces cerevisiae*.

Le complexe TFIIIC est composé de six sous-unités (Dumay-Odelot et al., 2007; Huang and Maraia, 2001). TFIIIC reconnaît le complexe TFIIIA-ADN, dans le cas du promoteur de type I. L'espacement entre les boites A et B est variable. TFIIIC peut s'adapter à cette distance variable car il est composé de deux sous-domaines reliés par un linker relativement flexible, nommés τA et τB chez la levure. L'architecture des sous-unités de TFIIIC est relativement bien conservée de la levure à l'homme (Orioli et al., 2011a). Le domaine τB est composé des sous-unités Tfc8/TFIIIC90, Tfc6/TFIIIC110 et Tfc3/TFIIIC220. Il se lie à la boîte B, via la sous-unité TFIIIC220, tandis que les deux autres sous-unités constitueraient une plateforme nécessaire à la liaison de Tfc8/TFIIIC220 (Mylona et al., 2006). Le domaine τA est composé des sous-unités Tfc4/TFIIIC102, Tfc1/TFIIIC63 et Tfc7/TFIIIC35. Ce domaine se lie à la boîte A, par l'intermédiaire de TFIIIC63 et de TFIIIC35. TFIIIC90 forme un pont entre ces deux domaines. La liaison de TFIIIC au domaine A est partie intégrante du mécanisme de mise en place du complexe de préinitiation (PIC), tandis que la liaison de TFIIIC à la boîte B doit être considérée comme un mécanisme d'activation. Le domaine τB n'est pas responsable du recrutement de TFIIIB, sa liaison est indépendante de la distance par rapport au TSS (Orioli et al., 2011a).

TFIIIC peut se lier à la chromatine *in vitro*. Le complexe TFIIIC humain porte une activité Histone Acetyl-transferase (HAT) pour trois de ses sous-unités, TFIIIC90 et TFIIIC110, TFIIIC220 (Geiduschek and Kassavetis, 2001; Hsieh et al., 1999; Kundu et al., 1999). Ce type d'activité n'a pas été démontré chez la levure.

Sous-unités *H. sapiens*	Commentaires
TFIIIC220	TFIIIC220 lie l'ADN, au niveau de la boîte B, en coopération avec TFIIIC110.
TFIIIC102	TFIIIC102 s'associe avec Brf1, TBP et TFIIIC63. Sous-unité de TFIIIC la plus conservée.
TFIIIC63	Se lie à la boîte A des promoteurs de type II. Avec TFIIIC35, elle forme un sous-complexe. Se lie à Brf1, TBP, TFIIIC102 et Rpc62.
TFIIIC110	Se lie au terminateur. TFIIIC110 et TFIIIC220 forment un sous complexe, et se lient en coopération, à la boîte B. TFIIIC110 possède une activité HAT.
TFIIIC90	TFIIIC90 forme le pont reliant les deux domaines τA et τB. Se lie à TBP. TFIIIC90 possède une activité HAT.
TFIIIC35	Se lie à la boîte A des promoteurs de type II.

Tableau 5. Description des rôles des six sous-unités du complexe TFIIIC (d'après Schramm and Hernandez, 2002).

TFIIIA est constitué d'un unique polypeptide de 38 kDa, contenant neuf domaines à doigt de zinc (Geiduschek and Kassavetis, 2001). Le rôle essentiel de TFIIIA est le recrutement de TFIIIC au promoteur des gènes d'ARN 5S.

La protéine responsable du recrutement de la RNAP III au promoteur est TFIIIB, car elle contacte l'ARN polymérase directement. TFIIIB a été défini initialement chez la levure *S. cerevisiae*. TFIIIB est composé de trois sous-unités la TATA-box protéine TBP, une sous-unité TFIIB-related factor BRF (Hernandez, 1993) et la sous-unité nommée B'' ou Bdp1 (B double prime) (Kassavetis et al., 1995 ; Ruth et al., 1996). L'homologue humain de B'' présente une forte similarité avec le facteur de la levure au niveau du domaine SANT, essentiel à la transcription chez la levure.

TBP était considéré comme un facteur de transcription universel, puisqu'il est impliqué dans les trois classes de transcription (Hernandez, 1993). Cependant, plusieurs paralogues de TBP ont été identifiés, TRF1, TRF2, TRF3. Leur découverte indique que la transcription de certains gènes pourrait être indépendant de TBP. TRF1, exprimé de façon ubiquitaire chez *Drosophila melanogaster* est spécifique des insectes. TRF1 est impliqué dans la transcription des gènes de classe III (Isogai et al., 2007). TRF1 est également responsable de la transcription de snARNs de classe II. TRF2, retrouvé chez plusieurs espèces de métazoaires, comme *D. melanogaster* ou les mammifères est plutôt impliqué dans la transcription de

classe II. (Reina and Hernandez, 2007). TRF3, également nommé TBP2, est également retrouvé chez divers métazoaires, de l'humain au poisson. TRF3 est impliqué dans la différentiation des myoblastes de souris, dans l'hématopoïèse chez le Xénope (Muller et al., 2010). Son activité serait restreinte aux gènes de classe II, pourtant le fort degré de conservation entre son domaine C-terminal et celui de TBP suggère qu'il puisse être impliqué dans la transcription des gènes de classe III.

Brf1 possède une structure similaire à TFIIB. Son extrémité N-terminale contient un domaine à Zinc de liaison à l'ADN. Ce complexe TFIIIB est impliqué dans la transcription de tous les types de gènes de classe III chez la levure. Chez l'humain, deux formes différentes de TFIIIB ont été identifiées (Teichmann and Seifart, 1995). La forme hTFIIIBα (h, human) est recrutée aux promoteurs de type III, tandis que la forme hTFIIIBβ est recrutée aux promoteurs de type intragénique. TFIIIBβ contient l'isoforme Brf1. TFIIIBα contient un isoforme de BRF, nommée TFIIIB50 ou Brf2 (Teichmann et al., 2000). Brf2 diffère essentiellement de Brf1 au niveau du domaine C-terminal. TBP et Bdp1 sont requis pour la transcription des promoteurs de type I, II et III.

Le promoteur de type III requiert des facteurs spécifiques, sans équivalents chez la levure. Le PSE est lié par le complexe SNAPc (snRNA Activating Protein complex), appelé également PTF (PSE Transcription Factor) ou PBP (PSE-Binding Factor) (Waldschmidt et al., 1991; Yoon et al., 1995). Le complexe SNAPc se lie indistinctement aux éléments PSE des gènes de classe II et III (Hernandez, 2001). SNAPc est un complexe composé de cinq sous-unités (Henry et al., 1998), SNAP19, SNAP43, SNAP45, SNAP50 et SNAP190. La cinquième sous-unité SNAP19 n'a pas été identifié dans le complexe PTF (Yoon et al., 1995). Le DSE est lié par les protéines Oct1 à la séquence octamère (Murphy et al., 1992) et Staf ou SPF (SPH1 binding Factor) au motif SPH (Schaub et al., 2000), facilitant l'interaction de SNAPc au PSE (Hernandez, 2001). Oct1 est le membre fondateur de la famille des protéines à domaine POU. Cette famille est caractérisée par ce domaine POU, structure bipartite consistant en un domaine spécifique POU, et un homéodomaine POU (POU_H), reliés par un domaine flexible (Hernandez, 2001).

D. La transcription de classe III

Les trois phases principales de la transcription sont l'initiation, l'élongation et la terminaison, chaque étape est régulée. Durant l'initiation, un complexe compétent de l'ARN polymérase et de ses facteurs de transcription est formé au promoteur, la matrice ADN est alignée au niveau du site catalytique de l'ARN polymérase. Au site catalytique, les nucléotides sont appariés et ajoutés de façon processive lors de l'élongation. Enfin, le complexe de transcription et le transcrit sont relargués de la matrice au cours de la terminaison de la transcription.

Initiation

Recrutement des facteurs de transcription aux promoteurs-formation du PIC

La première étape de la formation des complexes d'initiation de la transcription (PIC), de la RNAP III sur les promoteurs de type II consiste dans la reconnaissance des sites A et B par TFIIIC. Une fois le complexe TFIIIC-ADN formé, TFIIIB est recruté et se fixe stablement environ 25 bases en amont du TSS. La connexion entre TFIIIB et TFIIIC est due principalement à l'interaction entre Bdp1 et TFIIIC102. TBP et Bdp1 induisent une courbure à l'ADN. La RNAP III est ensuite recrutée, via des interactions entre TFIIIB et la RNAP III. Les principaux contacts entre la RNAP III et TFIIIB s'effectuent via les sous-unités Rpc17, Rpc34 d'une part et Brf1 d'autre part (Geiduschek and Kassavetis, 2001). Brf1 contacte également les sous unités Rpc32/Rpc39/Rpc62. Cette interaction est observée chez *S. cerevisiae,* et chez les mammifères (Kenneth et al., 2008).

TFIIIB joue également un rôle en aval du recrutement de la RNAP III, dans l'initiation de la transcription, guidant l'ouverture de la bulle de transcription (Grove et al., 2002; Kassavetis et al., 2003; Kassavetis et al., 2001). TFIIIB peut, *in vitro,* être recruté indépendamment de TFIIIC, en amont de certains gènes de *S. cerevisiae* possédant une boîte TATA, comme le gène de l'ARN U6, ou certains gènes d'ARNt (Dieci et al., 2000). Mais l'assemblage de TFIIIB en amont du TSS des gènes de classe III est entièrement dépendant de TFIIIC *in vivo* (Burnol et al., 1993). La plupart des promoteurs de type I et II ne possèdent pas de séquence TATA, c'est pourquoi TFIIIB doit être recruté par TFIIIC (Kassavetis et al., 2003).

Dans le cas du promoteur de type I, l'ICR est tout d'abord reconnu et lié par le polypeptide TFIIIA, servant de plateforme au recrutement de TFIIIC. Le recrutement de TFIIIB et de la RNAP III suit ensuite une voie identique à la formation du PIC sur les promoteurs de type II.

Le gène U6 chez la levure dépend de TFIIIC, tandis que chez les mammifères, TFIIIC n'est pas nécessaire, la forme TFIIIBα est recrutée. Oct1 et SNAPc se lient de façon coopérative au niveau de leur site respectif DSE et PSE. Cette liaison requiert un contact direct protéine-protéine entre Oct1 et SNAP190. Cette interaction, malgré la distance séparant leurs sites de liaison est possible grâce à la structure particulière de la chromatine ; un nucléosome positionné entre le DSE et le PSE rapproche les deux sites et permet le contact entre les deux protéines (Zhao et al., 2001). SNAPc recrute alors TFIIIB, qui recrute à son tour la RNAP III. Malgré la similarité partielle des protéines Brf1 et Brf2, le recrutement de la RNAP III aux promoteurs de type III requiert les mêmes interactions entre Brf2 et Rpc32/Rpc39/Rpc62 (Kenneth et al., 2008).

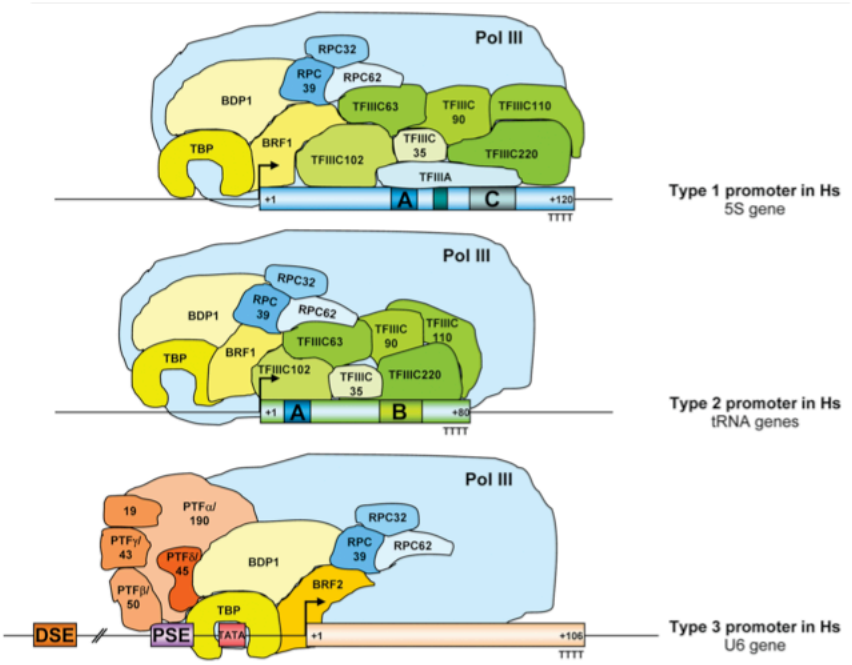


Figure 4. Recrutement des facteurs de transcription aux promoteurs des gènes de classe III (d'après Teichmann et al., 2010). Hs : *Homo sapiens*

Les sous-unités du complexe PSE-binding transcription factor (PTF) ont été indiquées selon la nomenclature PTF, leur masse moléculaire correspond aux sous-unités du complexe SNAPc. SNAPc 19 n'a pas été reporté dans le complexe PTF.

Initiation de la transcription

In vitro, TFIIIB une fois recruté au promoteur par TFIIIC ou SNAPc est capable de diriger plusieurs cycles de réinitiation de la transcription, et cela même en l'absence de TFIIIC (Paule and White, 2000), démontrant que l'interaction entre la RNAP III et TFIIIC, TFIIIA et/ou SNAPc n'est pas essentielle à l'initiation. Ainsi, TFIIIA, TFIIIC et SNAPc sont considérés comme des facteurs de recrutements de TFIIIB aux différents promoteurs, au même titre que TFIIB dirige le recrutement de la RNAP III (Kassavetis et al., 1990). Le positionnement de TBP dans TFIIIB est dû au recrutement de TFIIIB par TFIIIC, et est spécifié par la position de la boite A (Joazeiro et al., 1996). La RNAP III identifie le site d'initiation de la transcription sur la base de la distance imposée par la fixation de TBP à l'ADN. Cet espacement n'est pas complètement fixe, car la RNAP III peut se déplacer sur l'ADN jusqu'à localiser le signal -1/+1 pyrimidine/purine, spécifiant l'extrémité 5' de l'ARN (Grove et al., 2002).

Le recrutement de la RNAP III par TFIIIB induit une dénaturation de l'ADN autour du TSS de la transcription sur une matrice linéaire ou surenroulée ; cette bulle s'étend sur environ 22 pb autour du nucléotide d'initiation (Kassavetis et al., 1992; Kassavetis et al., 1990).

Au cours de la transition de l'initiation en élongation, deux ou trois transcrits abortifs sont produits avant que la synthèse du transcrit, en entier, ne démarre. Cette phase est rapide, et ne se traduit pas par une pause particulière de la RNAP III (Bhargava and Kassavetis, 1999). La RNAP III se détache de TFIIIB après avoir transcrit une chaîne d'environ six nucléotides au minimum (Kassavetis et al., 1992).

Elongation

L'étape d'échappée du promoteur n'est pas limitante. Une fois la transcription amorcée, la RNAP III progresse le long du gène, en parallèle de la bulle de transcription. L'élongation n'est pourtant pas uniforme, de nombreux sites de pause émaillent l'unité de transcription, comme les nucléosomes, ou des dommages à l'ADN. Les ARN Polymérases en élongation rencontrent souvent divers obstacles provoquant des pauses, et parfois même la terminaison de la transcription. Un mécanisme permettant le passage de ces blocages est le clivage du transcrit naissant dans le complexe ternaire, permettant d'aligner de nouveau l'extrémité 3' du transcrit dans le site actif de l'enzyme. L'activité hydrolytique 3'-> 5' de l'ARN est ubiquitaire, et est une fonction intrinsèque de l'ARN polymérase elle-même. Pourtant, il existe souvent un facteur associé à la RNAP, stimulant cette activité. La RNAP III, comme la RNAP I n'emploie pas de facteur d'élongation particulier. La reconnaissance de ces sites de pause est effectuée par la sous-unité Rpc11, pour la RNAP III. Cette sous-unité est l'homologue fonctionnel de TFIIS, facteur d'élongation de la RNAP II (Chedin et al., 1998; Landrieux et al., 2006).

Terminaison

Une fois l'élongation achevée, la reconnaissance des signaux de terminaison permet à la RNAP III de re-larguer le nouveau transcrit et de ré-initier un nouveau cycle de transcription. Ces sites sont généralement une succession de thymidines, au nombre de quatre chez les mammifères, à cinq ou six thymidines chez la levure. Le terminateur est généralement situé en aval de la séquence de l'ARN mature, impliquant une étape post-transcriptionnelle de clivage de l'extrémité 3' du pré-ARN. Cependant, le terminateur d'un grand nombre de gènes d'ARNt ne ressemble pas à une séquence canonique, et serait situé très en aval. *In vitro*, ces sites peuvent soit assurer une terminaison de la transcription efficace, ou être permissifs. Ils laisseraient passer la RNAP III au travers. Dans ce cas-là, la terminaison aurait lieu plus en aval, au niveau d'un autre site de terminaison. Cette « fuite » génère par conséquent une séquence 3' du pré-ARN beaucoup plus longue. Cette séquence est clivée lors de la maturation de l'ARN. Orioli suggère que cette séquence pourrait donner naissance à une nouvelle espèce d'ARN (Orioli et al., 2011b). La partie 3' des SINEs contient normalement un terminateur. Cependant, certains SINEs ne possèdent pas cette séquence signal. Leur transcription se poursuit jusqu'à ce que la RNAP III rencontre un terminateur. Ce mécanisme peut également être la source de nouveaux transcrits.

La terminaison est également une propriété intrinsèque de la RNAP III, ne nécessitant généralement pas l'intervention de facteurs externes. Comme précisé précédemment, cette étape requiert le sous-complexe Rpc53/Rpc37, qui diminuant la processivité de la RNAP III, ne lui permet pas de passer au travers des séquences Poly(dT) (Landrieux et al., 2006).

Réinitiation

Deux stratégies peuvent être adoptées lors de la réinitiation de la transcription (Dieci and Sentenac, 2003). Dans le cas de la réinitiation assistée par le PIC, une ou plusieurs étapes du cycle de transcription peuvent être contournées. Les facteurs de transcription recrutés au moment de l'initiation du premier tour de transcription restent associés au promoteur, durant les cycles successifs de transcription. Ces complexes contiennent les facteurs nécessaires au recrutement de l'ARN polymérase. Dans le cas des réinitiations très efficaces, l'ARN polymérase reste stablement associée à la matrice, et réinitie très rapidement de nouveaux tours de transcription. La deuxième stratégie de réinititation est basée sur les protéines. Dans ce cas, ces facteurs doivent être recyclés pour diriger de nouveaux tours de transcription.

Tous les facteurs de transcription de classe III restent liés à l'ADN durant la transcription facilitant la ré-initiation de la transcription du même gène (Ferrari et al., 2004). La RNAP III réinitie la transcription sur la même matrice, dans un processus couplé à la terminaison. Au moment de la terminaison, l'ARN est libéré, tandis que la RNAP III reste associée au gène. L'initiation pourrait être facilitée par l'établissement d'un contact spécifique entre la RNAP III présente au terminateur et un facteur du complexe de préinitiation. Ce contact pourrait impliquer la formation d'une boucle facilitée par la petite taille des gènes de classe III (Dieci and Sentenac, 1996). Ce processus de réinitiation est également observé pour des gènes plus long comme *SCR1* chez la levure (ARN 7SL) (Dieci et al., 2002). Chez l'homme, il existe des facteurs putatifs facilitant la réinitiation, comme la protéine La, ou NF1. La phosphoprotéine La peut intervenir durant la terminaison et le ré-initiation. Cette protéine peut lier l'oligo(U) (signal de terminaison) en 3' de l'ARN naissant, et le protéger de la dégradation. Il pourrait de plus faciliter la maturation de l'ARN et son assemblage dans les complexes RNP (Maraia, 2001).

E. Régulation de la RNAP III, et son implication dans le cancer

Le cycle cellulaire implique une coordination finement régulée de la transcription des ARN Polymérases. La RNAP III transcrit l'unique ARNr, non transcrit par la RNAP I, le 5S. Elle transcrit également l'ARN MRP, impliqué dans la maturation des ARNr. La RNAP III transcrit quant à elle, des ARN essentiels à la traduction des ARNm. D'autres ARN comme le 7SK et les SINEs sont impliqués dans la régulation de la RNAP II elle-même. Le couplage de la transcription par les trois ARN

polymérases des ARNs nécessaires à la machinerie de traduction suggère une co-régulation très fine de ce processus au cours du cycle cellulaire. De nombreux régulateurs sont impliqués dans le contrôle de la transcription de classe I et III, comme Ras, Erk, PTEN, Myc, p53, ARF et RB. Le changement d'un de ces régulateurs peut être suffisant pour déréguler la transcription des deux classes, dans les cellules transformées. De plus, certains de ces changements coopèrent, augmentant la dérégulation de la transcription.

L'augmentation de l'abondance des ARNs de classe III a été observée dans un grand nombre de cellules transformées, comme des cancers ovariens ou différents carcinomes de l'œsophage, du poumon, par exemple (White, 2004). Cette augmentation générale des transcrits de classe III est issue d'au moins trois grandes causes : i) la surexpression des facteurs de transcription de la RNAP III, ii) la diminution de la répression de ces facteurs de transcription, et enfin iii) de l'activation directe par des oncogènes dérégulés dans les cellules transformées (White, 2005).

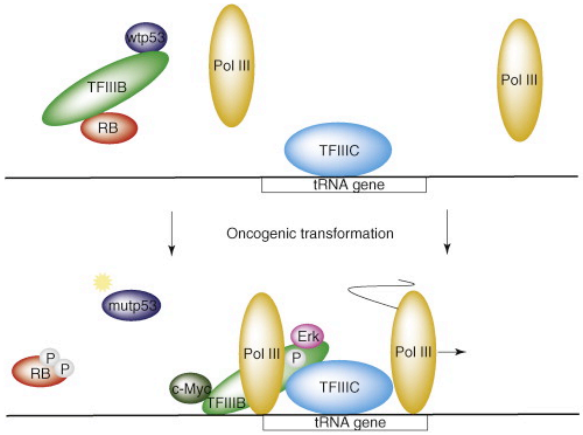


Figure 5. Illustration schématisant certains changements de la machinerie de classe III accompagnant la transformation oncogénique (d'après White, 2008).

RB et p53 répriment la transcription de classe III, via des interactions avec TFIIIB, inhibant son interaction avec TFIIIC et la RNAP III. L'hyperphosphorylation de RB, et des mutations altérant l'expression de p53, entraînent la perte de cette régulation. La transformation est accompagnée de l'augmentation de l'activité de c-Myc, de Erk, tous deux stimulant le recrutement de TFIIIB. TFIIIB peut également être phosphorylé et activité via la voie PI3K.

Phosphorylation : P

La régulation de la transcription de la RNAPIII est liée aux mécanismes de contrôle du cycle cellulaire. Des suppresseurs de tumeur tels pRB (protéine du Rétinoblastome) (Larminie et al., 1997; Scott

et al., 2001; Sutcliffe et al., 2000), p53 (Cairns and White, 1998), PTEN (Woiwode et al., 2008) répriment la transcription de classe III. Cette inhibition s'exerce principalement via TFIIIB. Les suppresseurs de tumeur inhibent l'accès de TFIIIB au promoteur, via une interaction directe. La fonction de ces suppresseurs de tumeurs est souvent compromise dans un grand nombre de cancer, le contrôle s'exerçant sur TFIIIB est alors perdu, expliquant l'augmentation générale de la synthèse des transcrits de classe III.

L'activité de TFIIIB est stimulée par phosphorylation par la kinase Erk de la voie MAPK (Mitogen Activated protein Kinase). Brf1 est lié, phosphorylé et activé par Erk, augmentant l'occupation des gènes de classe III, par TFIIIB (Felton-Edkins et al., 2003). Or Erk est anormalement active dans environ 30% des cancers (Felton-Edkins et al., 2003). La forme oncogénique de Ras de la voie MAPK augmente l'expression de Erk, stimulant ainsi la transcription de classe III. En parallèle, Ras induit l'expression de TBP, sous-unité de TFIIIB. La déplétion de PTEN, un suppresseur de tumeur dont l'activité phosphatase inhibe la voie PI2K entraîne l'augmentation de la transcription de classe I et III (Woiwode et al., 2008).

Le proto-oncogène c-Myc est un activateur de la croissance et de la prolifération cellulaire. c-Myc régule la production des ARNs de classe III, comme les ARNt, le SINE B2 ou l'ARN 7SL (Gomez-Roman et al., 2003; Goodfellow et al., 2006). Les gènes de classe III n'ont pas de boîtes E, utilisées par c-Myc pour activer la transcription de classe I et II. c-Myc peut être recruté par une interaction directe entre son domaine de transactivation et TFIIIB (Gomez-Roman et al., 2003). c-Myc peut également être recruté par TFIIIC, comme il a été identifié comme partenaire de TFIIIC lors d'un criblage protéomique (Koch et al., 2007). Une fois lié au promoteur des gènes d'ARNt et d'ARN 5S, c-Myc recrute les complexes TRRAP et GCN5, une histone acétyl-transférase. L'histone H3 est alors acétylée ; cette modification de la structure de la chromatine conduirait à l'assemblage du complexe de transcription TFIIIB suivi du recrutement de la RNAP III (Kenneth et al., 2007). L'induction de la transformation des cellules par c-Myc pourrait être compromise lorsque la transcription de classe III est perturbée. Lorsque l'activité de Brf1 est inhibée, bloquant seulement l'induction de la transcription, tout en maintenant un taux normal de transcription, la transformation par c-Myc est atténuée (Johnson et al., 2008).

La surexpression de la transcription de classe III est observée dans un certain nombre de cancer, mais son implication directe en tant que facteur déclenchant n'a jamais été démontrée clairement. Aucune mutation récurrente d'un de ses facteurs de transcription n'a été associée directement à un type de tumeur. Récemment, il a été montré que la sous-unité Brf1 peut avoir une activité oncogénique. Lorsque Brf1 est surexprimé, son induction est suffisante pour immortaliser des fibroblastes embryonnaires de souris (MEF) (Marshall et al., 2008). En outre, l'induction d'un gène dont la transcription est dépendante de Brf1, l'ARNt méthionine initiateur ($ARNt^{METi}$), est à elle seule capable de transformer les MEFs. L'injection des cellules induites provoque l'apparition de tumeurs lorsqu'elles sont injectées chez la

souris. En plus de l'augmentation globale de la synthèse des protéines, la surexpression de Brf1 ou de l'ARNt^{METi} augmente l'expression des régulateurs du cycle cellulaire c-Myc, de la cycline D1, de FGF2 (Fibroblast Growth Factor 2), et du VEGF (Vascular Endothelial Growth Factor) et des facteurs suppresseurs de l'apoptose Survivin, et Bax-inhibitor 1. Cependant, de quelle manière c-Myc est impliqué dans la transformation des cellules par Brf1 ou cet ARNt reste à déterminer. Cette étude suggère une boucle de régulation positive, où la stimulation de la transcription de classe III par c-Myc de l'ARNt initiateur induirait en retour la traduction de c-Myc.

La répression rapide de la transcription par la RNAP III assure la survie de la cellule suite à un stress. Maf1, protéine conservée de la levure à l'humain, réprime la RNAP III en réponse à des dommages à l'ADN, des stress oxydatifs, un traitement à la rapamycine (inhibiteur de la voie TOR, Target of Rapamycin), ou encore le blocage de la voie de sécrétion. En condition normale de croissance, Maf1 est phosphorylée et présente dans le cytoplasme. Suite à un stress, Maf1 est déphosphorylé, et importé dans le noyau. Maf1 se lie alors avec le sous-complexe Rpc82/ Rpc34/ Rpc31, provoquant un réarrangement de la RNAP III, prévenant la liaison de TFIIIB à la polymérase. Maf1 peut également se lier à la RNAP III en élongation, empêchant la réinitiation de la transcription (Vannini et al., 2010).

Les études de ChIP-seq (Barski et al., 2010; Moqtaderi et al., 2010; Oler et al., 2010; Raha et al., 2010) ont également mis en évidence la présence d'autres facteurs tels que FOS, JUN et ETS1. Bien que la co-localisation ne prouve pas l'interaction fonctionnelle entre ces protéines et la machinerie de classe III, ces facteurs pourraient influencer la transcription. Ils pourraient de même participer à l'augmentation de la transcription de classe III, observée dans de nombreux cancers. L'implication de ces différents facteurs dans le contrôle de la transcription de classe II et de classe III permettrait de réguler finement la croissance et la division cellulaire.

Enfin, récemment, deux isoformes de la RNAP III ont été identifiées chez les mammifères (Haurie et al., 2010). Elles diffèrent par la sous-unité RPC32. RPC32 est la seule sous-unité spécifique de la RNAP III, ne présentant aucune homologie de séquence ou de structure avec une sous-unité ou un facteur de transcription des autres ARN Polymérases. La RNAP IIIß contient la sous-unité RPC32ß, et est ubiquitaire, tandis que la RNAP IIIα contenant la sous-unité RPC32α n'est détectée que dans des lignées cellulaires de leucémie ou dans plusieurs types de lymphomes. La suppression par ARN interférence de l'expression de RPC32α réduit la formation de colonie des cellules HeLa, dans des essais soft-agar. RPC32α est régulée négativement durant la différentiation et est activée lors de la transformation en cellules tumorales.

F. Expression différentielle des gènes de classe III

Un domaine encore assez peu exploré et pourtant mis en lumière dans de récentes études (Moqtaderi et al., 2010; Oler et al., 2010), est la régulation temporelle et spatiale de l'expression des gènes de classe III. Alors qu'il a longtemps été considéré que les transcrits de classe III étaient ubiquitaires, de nombreux contre-exemples émergent. L'identification par ChIP-seq des gènes de classe III a également révélé que 26% des loci des gènes d'ARNt sont différentiellement occupés selon le type cellulaire (Moqtaderi et al., 2010; Oler et al., 2010). L'expression des ARNt varierait jusqu'à dix fois parmi les tissus humains (Dittmar et al., 2006). De plus, bien qu'aucune des familles n'ait d'expression restreinte à un tissu, ou à un stade de développement, l'abondance relative des ARNt est très variable, le cerveau présentant le plus fort taux d'expression (Coughlin et al., 2009).

Une telle régulation différentielle impliquerait l'existence de séquences spécifiques agissant en *cis*, et l'intervention de nouveaux facteurs agissant en *trans*. De même, des isoformes de la RNAP III ou de ses facteurs de transcription pourraient exister dans les différents tissus, et ce à l'image de la sous-unité RPC32 (Haurie et al., 2010). Enfin, la variation de l'expression des gènes d'ARNt pourrait directement découler de l'abondance relative des facteurs de transcription TFIIIB et TFIIIC.

Le gène BC1 chez la souris (ou de son équivalent BC200 chez l'humain), qui, par rétrotransposition a été placé en aval d'éléments régulateurs spécifiques aux neurones, entraînant une expression normalement restreinte au tissu neuronal (Martignetti and Brosius, 1995). L'analyse des séquences des gènes d'ARNt a révélé l'existence de très nombreuses espèces d'ARNt qui, bien qu'ayant le même anticodon, sont très dissemblables dans le reste de leur séquence (Goodenbour and Pan, 2006). Cette diversité de séquences pourrait constituer un nouveau niveau de régulation de la traduction des ARNm. Les éléments A et B des ARNt sont très conservés, pourtant chez l'humain, certaines variations ont été notées dans les positions normalement les plus conservées, car intervenant dans la structure tertiaire de l'ARNt. Ces différences de séquences pourraient être impliquées dans l'expression différentielle de ces ARNt (Goodenbour and Pan, 2006). Certains gènes de classe III sont quant à eux localisés dans des introns de gènes dont l'expression est spécifique à un tissu ou à un moment particulier, cette localisation pourrait entraîner pour les gènes de classe III le même profil d'expression.

Biais de codon

Le contrôle de la transcription de ces gènes d'ARNt pourrait jouer un rôle lors du développement, ou dans la spécificité tissulaire. L'abondance des ARNt pourrait également avoir un impact dans le contrôle de la traduction des gènes les plus exprimés, via le biais d'usage de codon. Chez les bactéries et

la levure, l'abondance relative des ARNt isoaccepteurs influence considérablement l'expression des gènes. Le biais d'usage des codons existe au sein même des tissus différenciés, un tel biais pourrait être relié à l'expression différentielle des gènes d'ARNt selon les tissus (Dittmar et al., 2006). Plotkin observe l'existence d'un biais de codon au sein des tissus différenciés, pour des gènes paralogues. Il propose alors, que ce biais serait relié à l'expression différentielle des gènes d'ARNt (Plotkin et al., 2004).

Expression différentielle des SINEs

Au cours des premières étapes de l'embryogénèse, les SINEs sont très fortement exprimés. Leur expression décroît ensuite très rapidement au cours du développement. Cette forte expression est corrélée avec la déméthylation de l'ADN lors de l'embryogénèse précoce (Jaenisch, 1997). De même, les SINEs sont très exprimés dans les cellules tumorales, et très peu dans les tissus normaux et différenciés. Enfin, l'expression des SINEs augmente en réponse à un stress cellulaire, comme un choc thermique. Le rôle des SINEs dans la régulation de l'expression de certains gènes de classe II, en réponse à un choc thermique, ou au cours de la croissance, impliquerait une expression coordonnée de ces SINEs à ceux des gènes de classe II, ou en réponse à un stimulus.

Après insertion, les SINEs sont localisés dans un nouvel environnement chromatinien ; il est donc difficile d'isoler des séquences flanquantes communes à une famille de SINE, dirigeant l'expression tissu- ou stade spécifique (Kobayashi and Anzai, 1998; Roy et al., 2000). L'expression différentielle sera une conséquence du contexte dans lequel sera inséré le SINE, plutôt qu'un élément propre au SINE.

G. Chromatine

La chromatine consiste en sous-unités répétées nommées nucléosomes, composées de 147 pb d'ADN enroulés autour d'un octamère d'histones contenant deux de chaque histone H2A, H2B, H3, et H4. Entre chaque nucléosome, l'histone H1, dit linker, se fixe à l'ADN. Les nucléosomes sont ensuite compactés en fibres de 30 nm de diamètre, puis subissent encore d'autres niveaux de compaction. Les multiples interactions entre l'ADN et les histones font du nucléosome un des complexes ADN-protéine les plus stables. Cependant, la modulation de la structure de cette chromatine est essentielle pour des processus comme la réplication, et la transcription. Ce complexe est également un des plus dynamiques, de nombreuses modifications post-traductionnelles des histones entraînent des variations de la compaction de la chromatine, permettant une régulation fine de l'accessibilité à la chromatine, et de la transcription des gènes. La chromatine est remodelée avant et après l'initiation de la transcription, et durant l'élongation du transcrit (Studitsky et al., 2004).

La répression de la transcription associée à l'assemblage de l'ADN en chromatine résulte d'un défaut d'accessibilité des facteurs de transcription à leur site de liaison respectif, et d'autres effets comme le blocage de l'élongation (Li et al., 2007). Dans la plupart des cas, ce manque d'accessibilité est contourné par l'action d'activateurs spécifiques, capables de lier la chromatine, puis de recruter des facteurs de remodelage, ou de modifications des histones.

La susceptibilité à la répression chromatinienne des gènes de classe III est très variable (Paule and White, 2000). La transcription des gènes de l'ARN 5S est inhibée lorsque leur promoteur est occupé par un nucléosome, bloquant la liaison de TFIIIA. Cependant lorsque ces histones sont acétylées, TFIIIA est capable d'accéder à ses sites de liaison avec une affinité comparable à une matrice d'ADN nue. La forme du nucléosome change suffisamment provoquant un relâchement de l'enroulement de l'ADN. La perturbation de la structure de la chromatine entraîne une augmentation de la transcription de ces gènes, ceci découlant peut-être du déplacement des nucléosomes le long de l'ADN.

Les gènes d'ARNt sont assez résistants à la répression nucléosomale. Contrairement aux gènes 5S, la présence de l'histone H1 ne provoque pas de perte d'accessibilité des facteurs de transcription au promoteur. Cependant, il est important de noter que l'environnement chromatinien influence l'activité des gènes d'ARNt ; si un gène d'ARNt est inséré dans un domaine de chromatine inactive, comme à côté du locus silencieux HMR chez la levure, sa transcription sera réprimée.

La structure de la chromatine a également été examinée *in vivo* pour le gène U6, chez la levure. La transcription d'un gène U6 couvert par les nucléosomes régulièrement espacés, peut être restauré par l'addition de TFIIIC, mais non par TFIIIB ou TBP. Cette restauration dépend de la présence de la boîte B. Cette capacité de TFIIIC à se lier à la boîte B et à permettre la transcription du gène U6 est mis en exergue par une expérience où la transcription du gène U6, en l'absence de nucléosome ne requiert ni la séquence B ni TFIIIC (Schramm et al., 2000). Ces résultats suggèrent que TFIIIC, en plus de l'assemblage des complexes de transcription, serait capable d'inhiber la formation des nucléosomes au promoteur d'au moins certains gènes de classe III. TFIIIC peut également entrer en compétition avec les histones, *in vitro*, alors même que le gène est déjà incorporé dans la chromatine. La susceptibilité du gène 5S pourrait ainsi provenir de l'absence de boîte B au promoteur. Chez l'humain, TFIIIC recruterait p300 au promoteur des gènes d'ARNt. La fonction HAT de p300 est essentielle à l'activation de la transcription pour contrer la répression exercée par les nucléosomes. En plus de l'acétylation des histones, p300 stabiliserait TFIIIC au promoteur. p300 est également recruté au promoteur des gènes U6, *in vivo* (Mertens and Roeder, 2008). Pourtant, d'autres études montrent que contrairement à ce qui avait été rapporté précédemment (Kundu et al., 1999), TFIIIC seul, lié au promoteur ne serait pas capable de surmonter la répression de la chromatine.

Les SINEs potentiellement actifs dans la cellule sont très nombreux, ce qui pourrait entraîner de graves effets délétères si tous s'exprimaient. Les SINEs B2 et Alu sont très susceptibles à la répression nucléosomales. Le contexte chromatinien d'un SINE peut être un facteur important de sa régulation tissu- ou développement-spécifique (Kramerov and Vassetzky, 2005).

In vitro, la méthylation de l'ADN peut réprimer l'expression des gènes de classe III. Or les SINEs sont particulièrement riches en îlots CpG. Il a été ainsi observé que des éléments Alu fortement méthylés sont réprimés dans des cellules différenciées. Cependant, cette méthylation peut également n'être que la conséquence de la répression de l'expression de ces SINEs, ajoutant un verrou supplémentaire, et non la cause de cette répression.

Chez la levure, en phase exponentielle de croissance, l'ensemble des gènes d'ARNt est occupé par la RNAP III (Harismendy et al., 2003; Moqtaderi and Struhl, 2004 ; Roberts et al., 2003). Chez les mammifères, la régulation de l'expression des ARNt est variable selon le type cellulaire (Moqtaderi et al., 2010; Oler et al., 2010). Un environnement chromatinien permissif des gènes semble être déterminant pour la transcription de classe III, comme il l'est pour les gènes de classe II. Les analyses réalisées par ChIP-seq ont mis en évidence que les gènes actifs d'ARNt présentent des marques d'euchromatine comme l'acétylation de H3, ou la triméthylation de la lysine 4 de H3 (H3K4me3), contrairement aux gènes inactifs, qui présenteraient plutôt des marques spécifiques de l'hétérochromatine comme H3K27me3, et H3K9me3 (Barski et al., 2010; Moqtaderi et al., 2010; Oler et al., 2010).

L'acétylation de l'histone H3 corrèlerait avec le modèle de l'induction de la transcription par c-Myc, via le recrutement de l'acétyltransférase GCN5 (Kenneth et al., 2007). L'ensemble des modifications des histones étudiées corrèle avec l'activité de la RNAP III, comme cela a été démontré pour l'activité de la RNAP II. Certaines différences notables ont pourtant été observées. Les nucléosomes sont en amont du gène, tandis que la partie codante en est totalement dépourvue. Bien que la RNAP III puisse transcrire à travers les nucléosomes *in vitro* (Studitsky et al., 1997), le fort taux de transcription des gènes de classe III, et la continuelle présence des facteurs de transcription au corps du gène entraîne une déplétion des nucléosome en aval du TSS. Les nucléosomes seraient plutôt déplacés en amont de l'unité de transcription.

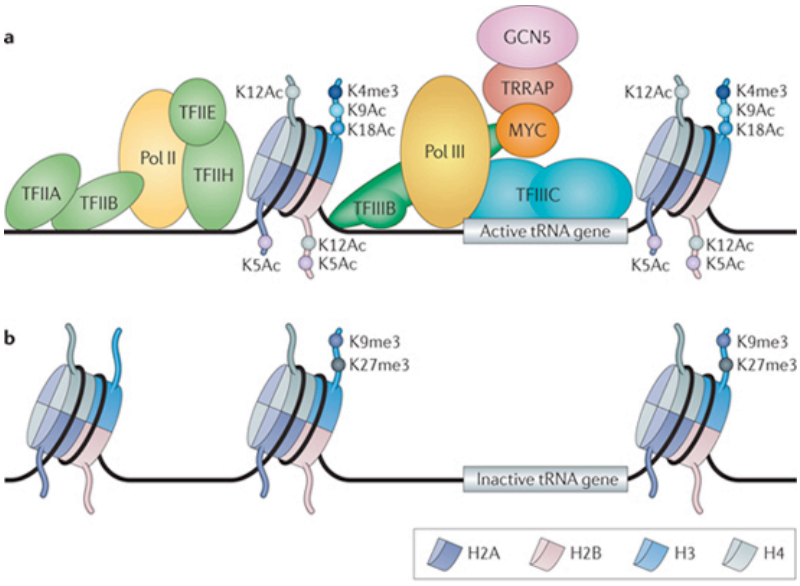


Figure 6. Schéma représentant l'environnement chromatinien et le recrutement des facteurs aux gènes d'ARN de transfert (A) actifs, (B) non transcrits, dans des cellules humaines (d'après White, 2011).

(A) les gènes d'ARNt actifs sont occupés par les facteurs de transcription TFIIIC, TFIIIB, l'ARN polymérase III, et souvent par MYC, qui recrute via le cofacteur TRRAP l'histone acétyl-transférase GCN5. L'environnement chromatinien présente des modifications associées à l'euchromatine, comme H3K4me3, H2AK5Ac, H2BK12Ac, H3K9Ac, H3K18Ac and H4K12Ac. L'ARN polymérase II peut souvent être détectée environ 200pb en amont chez l'humain, ainsi que certains de ses facteurs de transcription.

(B) les gènes d'ARNt inactifs ne sont pas occupés par la machinerie de transcription de classe III, ni par la machinerie de classe II en amont. Les histones à proximité de ces gènes sont peu acétylées et présentent des modifications caractéristiques de l'hétérochromatine, H3K9me3, H3K27me3.

Chapitre II : Rôle insulateur de la machinerie de classe III.

A. Organisation du génome

L'organisation cellulaire d'un génome fonctionne selon trois modes hiérarchiques : (i) l'organisation spatiale et temporelle des processus comme la transcription, la maturation des ARNs, la réplication et la réparation de l'ADN, (ii) l'organisation de la chromatine en domaines, et (iii) l'arrangement spatial des chromosomes et des gènes au sein du noyau (Misteli, 2007). Chacun de ces niveaux interdépendants représente un niveau potentiel de régulation de l'expression des gènes. Le lien établi entre l'activité des gènes et la structure de la chromatine provient de l'observation que les gènes actifs sont situés dans l'euchromatine, largement décondensée, tandis que les régions dont l'expression est réprimée, comme les séquences répétées, sont dans l'hétérochromatine condensée. Le modèle communément admis postulait que l'organisation dense de la chromatine serait à l'origine du mécanisme de régulation, en bloquant l'accès des protéines, les excluant de la chromatine condensée. Cette vision est un peu simpliste et ne permet pas d'expliquer le manque de corrélation observée entre l'activité des gènes et la condensation de la chromatine, ou l'accessibilité des protéines de l'hétérochromatine à leur site, au sein de régions très densément condensées (Misteli, 2007). La relation structure-fonction serait alors à l'échelle du nucléosome. L'accessibilité à la chromatine serait modulée via les interactions entre la séquence d'ADN et les marques épigénétiques, comme les modifications d'histones, et la méthylation de l'ADN, modifications dynamiques établies au cours de la vie de l'organisme (Bernstein et al., 2007).

En plus de la structure de la chromatine, le positionnement spatial des segments génomiques à l'intérieur du noyau influence l'expression du génome. Les chromosomes sont généralement organisés de façon radiale ; les chromosomes pauvres en gènes situés à la périphérie, tandis que les chromosomes denses en gènes sont localisés au centre du noyau. Le positionnement des chromosomes est dynamique durant l'interphase, suggérant qu'elle varierait en réponse à des stimuli extérieurs (Haeusler and Engelke, 2004). L'organisation en territoires distincts de la transcription présente l'avantage évident de regrouper et de favoriser les interactions des facteurs requis pour des processus comme la transcription, ou la répression. Cette organisation permet également la régulation coordonnée de l'expression de gènes. L'établissement de boucles intra-chromosomiques et de contacts inter-chromosomiques pourrait être stochastique. Pourtant des liens ont été établis entre ces structures et des fonctions biologiques importantes (Phillips and Corces, 2009). Les boucles de chromatine rapprochent des éléments éloignés, permettant la régulation coordonnée de ces sites, tandis leur séparation permet de les réguler indépendamment. Ce type

de structure a été impliqué à de nombreux niveaux de l'organisation chromatinienne, et peut être établi sur de longues distances, créant des interactions entre des éléments régulateurs, comme les enhancers et les silencers et les promoteurs des gènes. Les boucles peuvent contribuer à placer un gène dans un environnement nucléaire distinct, ou à l'isoler de son environnement pour le réguler différemment des gènes adjacents (Misteli, 2007). Les boucles de chromatine s'établissant au niveau local sont probablement impliquées dans la maintenance des propriétés comme l'expression spécifique de gènes voisins ou de domaines d'hétérochromatine et d'euchromatine. Les enhancers sont des éléments activant les promoteurs, à distance, et de façon indépendante de leur orientation. Les silencers sont des régulateurs négatifs, composés de sites de liaison de divers facteurs, dont la fonction est d'établir un état hétérochromatinien. Comme les enhancers, les silencers agissent indépendamment de leur orientation et de la distance les séparant des promoteurs, qu'ils répriment (Valenzuela et al., 2008). Les insulateurs ou régions frontières sont des séquences ADN, qui protègent un locus de l'influence de la chromatine voisine. Ils empêchent les interactions inappropriées entre des gènes voisins, et inhibent les communications à distance entre enhancers, ou silencers et les promoteurs. Deux types d'insulateurs ont été définis, les « enhancers blockers » qui bloquent l'activité d'un enhancer lorsque situés entre cet enhancer et le promoteur régulé, réprimant la transcription de ce gène. L'élément barrière est une séquence bloquant la propagation des marques d'hétérochromatine, séparant physiquement deux domaines organisés différemment au niveau chromatinien (Amouyal, 2010). Ces éléments insulateurs sont retrouvés chez la plupart des eucaryotes, soulignant leur importance fondamentale dans le contrôle de l'expression des gènes.

La formation de l'hétérochromatine est initiée par la liaison de complexes protéiques à des séquences répétées ou spécifiques, les silencers. Ces complexes de silencing recrutent la machinerie d'ARN interférence (Zaratiegui et al., 2007) ou des facteurs de transcription. L'hétérochromatine est capable de se propager de proche en proche, par le recrutement de complexes de répression, entraînant la déacétylation des histones, et leur méthylation, suivi de l'association avec des protéines non-histones, qui à leur tour recrutent des protéines qui modifient le nucléosome suivant (Grewal and Moazed, 2003). Chez la levure *S. cerevisiae*, la répression est sous contrôle des protéines SIR, tandis que chez *S. pombe*, l'initiation du silencing utilise la machinerie d'ARN interférence, et implique l'association de la protéine SWI6p et des histones (Valenzuela et al., 2008).

B. Les éléments insulateurs dépendants de la machinerie de classe III.

Les gènes d'ARN de transfert ont été impliqués dans un grand nombre de processus, outre de la production d'ARNt, où ils exercent un effet dit de position. Depuis longtemps, la répression des gènes de

classe II par les gènes d'ARNt voisins était connu ; l'intégration d'un gène d'ARNt en amont d'un promoteur de classe II, entraîne la répression de celui-ci (Hull et al., 1994). Ils peuvent également représenter des obstacles à la progression de la fourche de réplication (McFarlane and Whitehall, 2009). Finalement, ils ont une fonction d'élément barrière, en délimitant deux domaines fonctionnels de chromatine, bloquant ainsi la propagation de l'hétérochromatine.

Les ARN de transfert

Chez *S. cerevisiae*, la chromatine condensée est retrouvée au locus déterminant le type conjuguant (« mating-type », MAT), dans les régions télomériques, et autour du centromère chez *S. pombe*. La levure *S. cerevisiae* haploïde sauvage possède deux gènes *MATa* et *MATα*, non exprimés au locus *HRM* et *HML* respectivement (Hidden MAT Right/Left), et une copie active issue de la transposition d'un de ces deux gènes au locus MAT. Les copies aux loci HM sont inactives, à cause de la structure répressive de la chromatine. Les deux séquences HRM-E et HRM-I encadrant le gène *MATa* sont requises pour la formation de l'hétérochromatine à ce locus (Donze et al., 1999). Or, au locus HRM-I, se trouve un gène d'ARNt (Donze and Kamakaka, 2001). La délétion de ce gène, ou l'inhibition de sa transcription entraîne la propagation de l'hétérochromatine hors des frontières HMR.

D'autres gènes d'ARNt seraient des éléments barrières chez *S. cerevisiae*, comme *TRT2*, un gène d'ARNt inhibant la propagation de l'hétérochromatine aux gènes adjacents de *STE6*, réprimé dans le type cellulaire MATα (Simms et al., 2008), ou l'ARNt^{Gln} bloquant la propagation de l'hétérochromatine associée au locus des gènes d'ARNr (Biswas et al., 2009).

Le rôle barrière des gènes d'ARNt ne semble pas restreint à la levure *S. cerevisiae*. Scott et al (Scott et al., 2006) démontrent qu'un ARNt^{Ala} localisé au centromère du chromosome 1 de *S. pombe* exerce également cette fonction. Des mutations au promoteur de ce gène, qui compromettent l'assemblage de la machinerie de classe III résultent en une mauvaise ségrégation des chromosomes, lors de la division méiotique, due à un dysfonctionnement de l'activité de cet élément barrière. Enfin, la présence de cet ARNt^{Ala} empêche la propagation de la modification H3K9me2, marque de l'hétérochromatine. L'activité transcriptionnelle de ces gènes semble requise pour exercer la fonction insulatrice. Pourtant, la mutation de TFIIIC entraîne une perte de la fonction insulatrice aux sites HMR, contrairement à la mutation de la RNAP III (Donze and Kamakaka, 2001). Ainsi, l'assemblage d'un complexe de transcription de classe III complet ne serait pas toujours la règle.

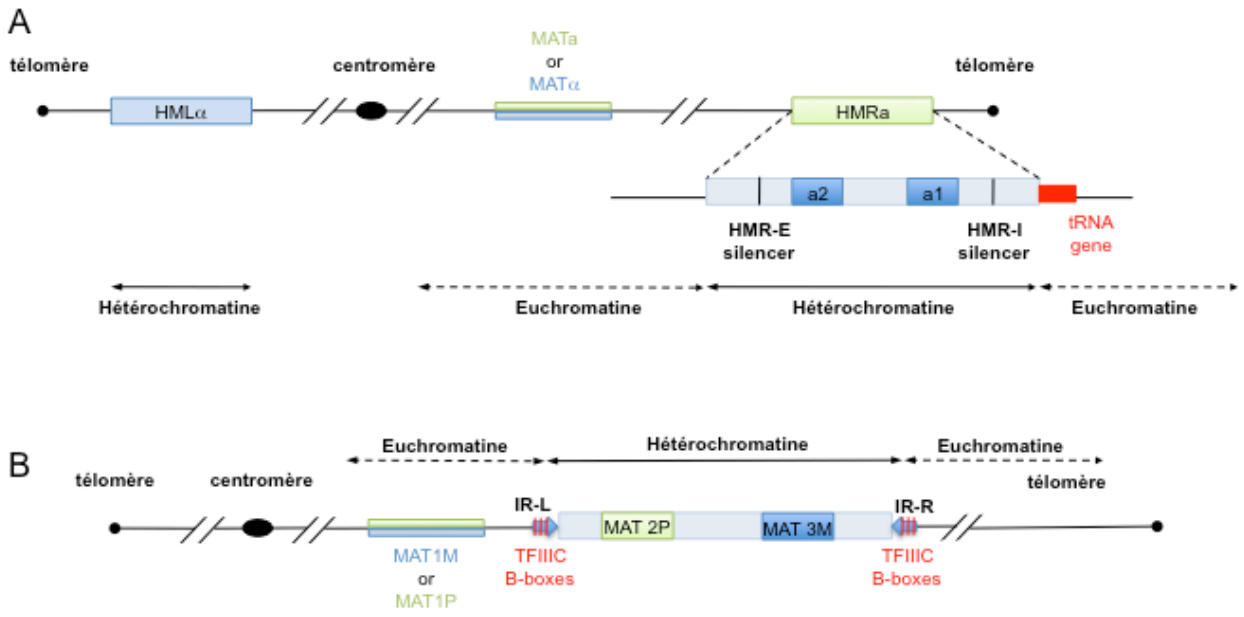


Figure 7. Insulation chez les levures *S. cerevisiae* et *S. pombe* au locus *MAT* (d'après Amouyal, 2010).

La chromatine condensée est présente (A) au locus HM chez la levure *S. cerevisiae*. Au locus *HMR*, deux copies du gène *a*, *a1* et *a2* sont présentes. (B) Chez *S. pombe*, le type cellulaire P ou M est spécifié par les gènes *MAT2P* et *MAT3M*, respectivement. Comme les loci *HMR* et *HMLα*, ce locus est réprimé par l'hétérochromatine.

TFIIIC

Chez *S. pombe*, le changement de type sexuel P ou M, est spécifié par la transposition au locus Mat1 d'un des deux gènes *MAT3P* et *MAT2M*, équivalents des sites *HRMa* et *HMRα*. Ces deux loci sont également réprimés par l'hétérochromatine, dont la propagation est limitée par deux répétitions inversées IR-R et L (Right et Left). Lorsque ces deux séquences sont éliminées, l'hétérochromatine envahit les régions adjacentes, provoquant la répression des gènes voisins (Noma et al., 2001). Le profil de modification des histones, au niveau du locus *MAT1* entre les deux IR, présente une transition abrupte des marques associées à l'euchromatine, et des modifications caractéristiques de l'hétérochromatine. Ceci suggère la présence d'un élément barrière au niveau des IR. Contrairement à *S. cerevisiae*, il n'y a pas d'interruption de la structure des nucléosomes. L'élément barrière serait constitué par cinq boîtes B, où seul TFIIIC serait recruté. Ni TFIIIB, ni la RNAP III n'ont jamais été observés à ces sites (Noma et al., 2006). De plus, ces sites sont souvent localisés en amont de gènes de classe II, suggérant qu'ils puissent les réguler. Contrairement aux promoteurs classiques des gènes de classe III, la boîte A est absente,

suggérant que TFIIIC adopterait une conformation différente lorsqu'il se lierait à ces boîtes B. Ce changement de conformation pourrait expliquer l'absence de recrutement du reste de la machinerie de classe III. L'analyse du génome de *S. pombe* a révélé qu'il existait plusieurs sites similaires, dénommés COC ou Chromosome Organizing Clamp.

Il est intéressant de noter, cependant, que les gènes d'ARNt ne possèdent pas tous cette fonction barrière. Ceci suggère que les séquences flanquantes pourraient également avoir un rôle. Ceci est également vrai pour les sites liés uniquement par TFIIIC, (ETC, Extra-TFIIIC loci), la localisation ectopique d'un des ces sites entraîne la diminution de son activité barrière, mettant en évidence l'importance des régions adjacentes (Valenzuela et al., 2009).

C. Mécanisme d'insulation

Le mécanisme précis par lequel les gènes d'ARNt ou les boîtes B exercent leur fonction barrière n'est pas clairement élucidé. Pourtant plusieurs modèles ont été proposés, sans exclusion mutuelle entre eux.

Le recrutement de la machinerie de classe III au promoteur crée l'assemblage d'un très large complexe, protégeant environ 150 pb d'ADN de la digestion par la DNase I (Chedin et al., 1998). Des mutations du promoteur du gène d'ARNt au locus *HMR* ou dans la machinerie de transcription entraîne la perte de la fonction barrière de ce gène, suggérant que l'intégrité de la machinerie de classe III est nécessaire à l'activité de cette barrière. Ce gène est activement transcrit, la machinerie recrutée est pleinement fonctionnelle. Une autre caractéristique des gènes d'ARNt est la réinitiation de type hyper-processive (Dieci and Sentenac, 1996). L'occupation persistante par le complexe de la RNAP III peut contribuer à la fonction insulatrice des gènes d'ARNt (Simms et al., 2004).

L'assemblage de la machinerie de transcription aux gènes de classe III entraîne une redistribution des nucléosomes, les gènes d'ARNt activement transcrits sont alors dépourvus de nucléosomes. Ceci provoque une rupture dans la structure de la chromatine (Mavrich et al., 2008). Cet intervalle crée par le gène transcrit, éliminerait le substrat nécessaire à la propagation de l'hétérochromatine, le nucléosome. Le mécanisme proposé est que l'assemblage de ce complexe de transcription agirait comme un bloc physique empêchant la propagation de l'hétérochromatine (Donze and Kamakaka, 2001). Cependant, des mutations restaurant la présence de nucléosomes au gène d'ARNt du locus *HMR*, entraînent seulement une perte mineure de la fonction barrière. Ceci suggère que l'absence de nucléosome ne serait pas nécessaire à l'établissement de cette fonction (Oki and Kamakaka, 2005).

Il existe deux autres modèles pouvant tout aussi bien s'appliquer aux sites barrières des gènes d'ARNt qu'aux sites liés par TFIIIC. L'organisation de la chromatine en des structures complexes, également dénommées architecture chromosomique pourrait avoir un rôle dans l'établissement des fonctions insulatrices. L'inhibition de la transcription de la RNAP II est associée à une relocalisation « sous-nucléaire » particulière, du locus *MAT*, aux télomères et centromères. Cependant aucun lien causal n'a été mis en évidence entre cette localisation et la répression de la transcription (Haeusler and Engelke, 2004). Les gènes d'ARNt sont localisés dans le nucléole, et cette association est largement dépendante de leur activité transcriptionnelle. Cet arrangement permet de concentrer la transcription de la RNAP III, et de la maturation des ARNt. De plus, il y aurait la possibilité de coordonner la bioynthèse des ARN impliqués dans la traduction, les ARNt, et l'ARN5S, et les ARNr (Bertrand et al., 1998). Etant donné le grand nombre de gènes d'ARNt dispersés sur l'ensemble des chromosomes, le regroupement des gènes au nucléole affectera la position de l'ensemble de l'ADN. La perturbation de la localisation au sein du nucléole des gènes d'ARNt entraîne la perte de la répression des gènes voisins de classe II (Haeusler et al., 2008; Kendall et al., 2000). La position des ARNt n'est cependant pas rigide, et est soumise aux contraintes exercées sur la chromatine environnante.

La fonction insulatrice peut être liée aux protéines recrutées par un des facteurs de transcription de classe III. La modification du profil des histones entraîne un changement dans la structure de la chromatine. Bien que les séquences ADN des barrières soient assez hétérogènes, l'activité intrinsèque (TFIIIC chez l'humain) ou le recrutement d'une activité histone-acétylase corrèle avec cette fonction dans de nombreux organismes (Donze and Kamakaka, 2002; Hsieh et al., 1999; Kundu et al., 1999 ; Lunyak et al., 2007 ; Mertens and Roeder, 2008). La fonction barrière du gène d'ARNt au locus HMR de *S. cerevisiae* est une combinaison de l'interruption de la chaine des nucléosomes, observée classiquement aux gènes d'ARNt, mais est également due au rôle HAT des protéines telles que Ada 2, Eaf3 et Sas2 (Oki and Kamakaka, 2005).

Chez *S. cerevisiae*, la fonction insulatrice de TFIIIC aux sites ETC utilise des histones acétylases et les remodeleurs de chromatine tels que le complexe RSC ; la fonction barrière passerait par le remodelage des histones (Valenzuela et al., 2009).

La structuration de la chromatine établissant la fonction insulatrice de ces gènes requiert des protéines comme la cohésine. Les silencers HMR-E et HMR-I interagissent ensemble formant une boucle, la nucléation de l'hétérochromatine initiée au locus HMR-E peut alors se propager via l'interaction avec HMR-I permettant l'établissement de l'hétérochromatine, au locus *MAT*. Cette interaction requiert les protéines Sir, et la cohésine qui permet l'appariement des deux domaines silencers (Valenzuela et al.,

2008). Le gène d'ARNt au locus HMR est requis pour l'établissement de la cohésion (Dubey and Gartenberg, 2007), et lorsque les deux sous-unités Smc1 et Smc3 de la cohésine sont mutées, la fonction barrière du locus HMR est compromise (Donze et al., 1999). Enfin, l'activité insulatrice du gène d'ARNt au locus ADNr requiert également la cohésine (Biswas et al., 2009). La condensine se lie aux gènes d'ARNt chez *S. cerevisiae* et *S. pombe*, avec une préférence pour les boîtes B liées par TFIIIC. De plus TFIIIC aurait un rôle dans l'étape de chargement de la condensine aux chromosomes. La condensine permettrait le regroupement des gènes d'ARNt (D'Ambrosio et al., 2008; Haeusler et al., 2008).

Enfin, l'ancrage des sites barrières à des structures comme les pores nucléaires formerait un obstacle à la propagation de l'hétérochromatine (Misteli, 2007). Les sites HMR sont localisés à la périphérie nucléaire, à proximité des télomères. Cet ancrage à la périphérie, peut-être dû aux séquences HMR, serait un des évènements responsable de la répression de ce locus (Valenzuela et al., 2008). Les sites COC chez *S. pombe* sont localisés à la périphérie du noyau. Cette localisation dépend de la liaison de TFIIIC aux boîtes B. Elle entraînerait la formation d'une structure particulière de la chromatine, à l'origine de la fonction insulatrice de TFIIIC. Ces sites seraient des centres organisateurs de la chromatine (d'où le nom, Chromosome-Organizing Clamps).

D. Conservation de ces sites

Ces sites COC ne sont pas sans rappeler les sites identifiées par ChIP-chip chez *S. cerevisiae*, dénommés ETC, ou Extra TFIIIC loci (Harismendy et al., 2003; Moqtaderi and Struhl, 2004; Roberts et al., 2003). En plus des gènes de classe III, TFIIIC se lie à plusieurs sites indépendamment de la liaison de la RNAP III. Ces régions sont parfaitement conservées chez les autres espèces *Saccharomyces*, suggérant que ces sites liant TFIIIC représentent une nouvelle classe d'éléments insulateur. Récemment des études similaires réalisées chez l'homme, par une approche de ChIP-seq, ont démontré l'existence de tels sites (Moqtaderi et al., 2010; Oler et al., 2010).

De plus chez *S. cerevisiae*, la fonction barrière a été confirmée pour certains sites ETC ; la liaison de TFIIIC au site *ETC6*, sans recrutement de TFIIIB, stabilise l'élément barrière, en bloquant la propagation de la répression effectuée par les protéines Sir (Simms et al., 2008). L'occupation de TFIIIC aux ETC est dynamique et est directement en compétition avec la liaison des protéines Sir (Valenzuela et al., 2009). Une étude très récente montre que la liaison de TFIIIC à la boîte B (ETC6) localisée dans le promoteur du gène *TFC6*, chez *S. cerevisiae*, régule négativement la transcription de ce gène (Kleinschmidt et al., 2011). Cette observation suggère que ces boîtes B isolées peuvent agir en tant

qu'élément *cis*, dans les promoteurs de classe II (Orioli et al., 2011a). Le rôle de TFIIIC n'est établi clairement, pour l'instant, que chez la levure.

Il est plus qu'improbable que les gènes d'ARNt aient évolué pour fonctionner comme des répresseurs de la transcription de classe II. On peut cependant imaginer que les propriétés de répression des gènes de classe III soient largement utilisées pour réguler l'environnement chromosomique des gènes de classe II adjacents. De plus les génomes eucaryotes contiennent de très nombreux éléments répétés, les SINEs, où sont recrutés les facteurs de transcription de classe III, et la RNAP III elle-même. Les SINEs sont plus particulièrement situés dans les zones riches en GC, généralement denses en gènes de classe II. Il a été suggéré que les SINEs étaient concentrés près des gènes car ils auraient un rôle dans le contrôle de la structure de la chromatine et réguleraient les gènes de classe II.

Chez les mammifères, les SINEs représentent un réservoir immense de sites insulateurs potentiels, étant issus de gènes de classe III, ils contiennent des boîtes B. Un élément Alu flanquant le gène de la kératine K18, chez l'homme, permet l'expression d'un transgène chez la souris, indépendamment de sa localisation génomique, rappelant la définition d'un élément barrière (Willoughby et al., 2000). La mutation de la boîte B de cet Alu abolit la protection contre l'effet de position. Chez la souris, un SINE B1 est également retrouvé au promoteur du gène de la kératine. La localisation similaire de l'élément Alu et du SINE B1 pourrait refléter une sélection fonctionnelle.

Une étude informatique a révélé que plus de 40% des promoteurs humains et murins de gènes de classe II renferment un SINE de type Alu, ou B1/B2. De plus ces SINEs sont préférentiellement situés entre -1000 et -200 avant le TSS. Les auteurs suggèrent que ces SINEs pourraient protéger les promoteurs de la répression chromatinienne (Usmanova et Tomilin, 2008). Les SINE B2, en plus de leur promoteur de classe III, peuvent pourvoir un promoteur de classe II, situé hors de la région similaire à l'ARNt (Ferrigno et al., 2001). Au locus murin de l'hormone de croissance, des transcrits recouvrants issus de transcription de classe II et de classe III sont générés au niveau d'un SINE B2. Ici, la transcription de ce SINE B2 est activée au cours du développement, et de façon tissu-spécifique. Cette transcription corrèle avec le réarrangement chromosomique de ce locus, et la localisation d'un domaine d'hétérochromatine, vers un compartiment euchromatinien. Ce SINE B2 constitue une barrière, empêchant la propagation de l'hétérochromatine. La transcription de classe II et III est ici nécessaire à cette fonction insulatrice (Lunyak et al., 2007). Enfin, les SINEs sont des sites de liaison de la cohésine chez l'humain (Hakimi et al., 2002).

Chapitre III : La transcription de classe II

A. L'ARN Polymérase II

La synthèse des ARNm est réalisée chez les eucaryotes par la RNAP II. Durant la transcription des ARNm, la RNAP II s'associe transitoirement avec différents facteurs de transcription, TFIIB, -D, -E, -F, -H. Leurs rôles sont multiples et distribués au cours des nombreuses étapes la transcription, de la reconnaissance du promoteur, et de l'ouverture de la bulle de transcription lors de l'initiation. Ils interagissent successivement avec les co-activateurs, transmettant les signaux de régulation de la transcription, avec les facteurs d'élongation, pour assurer l'élongation correcte des longs gènes, et enfin avec les facteurs multiprotéiques intervenants au moment de la terminaison, et de la maturation de l'extrémité 3' du transcrit nouvellement synthétisé.

Structure

La structure crystallographique de la RNAP II a été résolue dans l'équipe de Kornberg (Cramer et al., 2000; Cramer et al., 2001; Gnatt et al., 2001). La RNAP II est un complexe d'une masse supérieure à 0,5 MDa, composé de douze sous-unités très conservées chez les eucaryotes (Cramer et al., 2001). Les sous-unités de la RNAP II peuvent être classées en sous-unités du domaine cœur, homologues des sous-unités bactériennes (Rpb1, 2, 3 et 11), des sous-unités partagés par les RNAPI et III (Rpb5, 6, 8, 10, et 12) et des sous-unités spécifiques, et non essentielles lors de l'élongation (Rpb4, 7 et 9). La RNAP II peut se dissocier en un cœur catalytique de dix sous-unités, et un hétérodimère Rbp4 et Rpb7 (Cramer et al., 2001).

La RNAP II est composée de quatre éléments mobiles, le cœur, la pince (clamp), le module shelf, et la mâchoire (jaw). L'élément cœur est constitué par les sous-unités Rpb3, 10, 11, 12 et des régions Rpb1 et 2 formant le cœur catalytique. Au centre de l'enzyme, se trouve le sillon (cleft) par lequel s'introduit l'ADN. Le sillon est formé par les deux plus grosses sous-unités Rpb1 et Rpb2, il est chargé positivement. Un des côtés du sillon est formé par la pince (clamp) mobile qui adopte une structure ouverte ou fermée selon qu'elle interagisse ou non avec l'ADN. Le cœur catalytique est localisé au fond du sillon, au niveau du mur (wall). Un pore situé au-dessous du cœur catalytique s'élargit sur l'extérieur, créant un entonnoir (funnel) inversé. Le rebord du pore est constitué par une boucle de la sous-unité Rpb1, qui lie un ion magnésium Mg^{2+} (Armache et al., 2003; Brueckner et al., 2009). La structure minimale d'un complexe en élongation révèle un hybride ADN-ARN de 8 à 9 pb, localisé au cœur catalytique, et la

pince est dans un état fermé (Gnatt et al., 2001). L'hétérodimère Rpb4/Rpb7 se situe à la base de la pince, en dehors de la surface de l'ARN polymérase (Armache et al., 2005; Bushnell and Kornberg, 2003), Rpb7 servirait de « cale » pour bloquer la pince en position fermée. Le couvercle (lid) permettrait de séparer l'ADN de l'ARN nouvellement synthétisé, et de le guider vers la sortie.

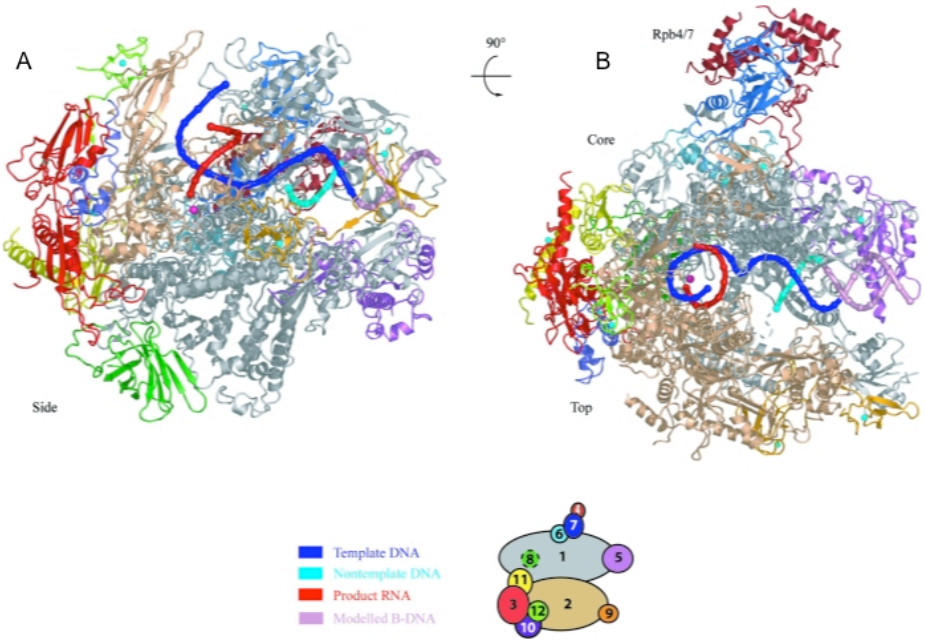


Figure 8. Structure du complexe en élongation de l'ARN polymérase II (d'après Kettenberger et al., 2004).

(A) Vue de côté de l'ARN polymérase II. Les sous-unités de l'ARN polymérase II sont représentées par des couleurs différentes (se référer au schéma situé entre les deux panneaux: les sous-unités sont numérotées d'après la nomenclature usuelle de Rpb1 à Rpb12). (B) Vue du dessus de l'ARN polymérase II. Les sous-unités de Pol II sont représentées par des couleurs différentes. L'ion Magnésium est représenté en violet.

Domaine Carboxy-terminal de Rpb1

Au niveau de la face amont, sous la pince, dépasse le domaine non-structuré CTD (Carboxy-Terminal Domain) de la sous-unité Rpb1. La polymérase est régulée tout au long du cycle de transcription par la phosphorylation et la déphosphorylation de son domaine CTD. Le domaine CTD agit comme une plateforme servant de surface de liaison aux facteurs impliqués dans la transcription, la maturation de l'ARNm, et la modification des histones (Buratowski, 2009). Le CTD consiste en de multiples répétitions

d'un heptapeptide Tyr_1-Ser_2-Pro_3-Thr_4-Ser_5-Pro_6-Ser_7. Cette région est hyperphosphorylée durant la transcription sur la sérine 5. Au cours de la formation du complexe de préinitiation (PIC), le Médiateur, co-activateur de la transcription, lie la RNAP II non phosphorylée, lors de l'incorporation de la polymérase, le médiateur stimule la fonction CTD-kinase de TFIIH sur la sérine 5 (Max et al., 2007). Cette phosphorylation perturbe la liaison du médiateur à la RNAP II et permet l'initiation de la transcription. Le médiateur reste cependant lié au promoteur après l'initiation, facilitant ainsi les cycles successifs de recrutement de la polymérase et de réinitiation de transcription (Yudkovsky et al., 2000). La phosphorylation de la sérine 5 est associée au promoteur, mais décroît très rapidement en aval. Cette phosphorylation est suivie du recrutement de plusieurs complexes de modification de l'ARNm, tel le complexe de coiffage de l'ARNm.

Les enzymes de modifications des histones peuvent reconnaître les modifications du CTD pour distinguer les régions promoteur-proximal et les régions distales (Hampsey and Reinberg, 2003). Les gènes actifs portent plusieurs modifications : H3K4me3 près des promoteurs, H3K79me2 juste après le promoteur, sur environ 5kb du gène transcrit, et H3K36me3, en aval des promoteurs, au niveau de la région transcrite.

TFIIH peut également phosphoryler la sérine 7 du CTD (Akhtar et al., 2009). Cette modification est associée spécifiquement à la transcription des snARNs, car elle permettrait le recrutement du complexe Intégrateur de maturation de l'extrémité 3' de ces ARNs (Egloff et al., 2007).

Au fur et à mesure de l'élongation, la phosphorylation de la sérine diminue, tandis qu'apparaît la phosphorylation de la sérine 2 (Komarnitsky et al., 2000). Chez les mammifères, la kinase Cdk9 du complexe d'élongation PTEF-b est responsable de la phosphorylation de la sérine 2. Chez la levure, les kinases sont Ctk1 et Bur1 (Buratowski, 2009). La phosphorylation du CTD ne semble pas déterminante en elle-même pour l'initiation de la transcription et l'élongation. Elle jouerait plutôt un rôle dans le couplage de l'élongation de la RNAP II et les étapes post-transcriptionnelles comme le coiffage, l'épissage, la polyadénylation du pré-ARNm, et la modification de la chromatine (Wade and Struhl, 2008).

La transcription

La régulation de l'expression des gènes est fondamentale pour la croissance normale le développement et la survie d'un organisme. L'expression des gènes est principalement régulée au moment de la transcription. De nombreux facteurs régulent la transcription en influençant la capacité de la RNAP II d'accéder, de lier et de transcrire le gène, en réponse au stimulus approprié.

Des études récentes remettent en cause le schéma, un peu simplifié de la transcription, postulant

que la régulation de la transcription s'exerce de façon prédominante au moment du recrutement de la RNAP II. L'élongation est en réalité plus complexe et se découple en étapes d'échappée du promoteur ou « promoter escape », de pause au promoteur, « proximal-promoter pausing » et d'élongation productive. Chacune de ces étapes est définie par des différences marquées de stabilité et de comportement du complexe de transcription, et des facteurs associés. Chez la levure *S. cerevisiae*, des analyses à grande échelle montrent que le niveau d'association de la RNAP II semble égal tout au long du gène, indiquant une transition rapide entre l'initiation et l'élongation (Radonjic et al., 2005; Wade and Struhl, 2008). Cependant, au moment de la phase stationnaire, la RNAP II s'associe à de nombreux promoteurs inactifs, en particulier des gènes dont la transcription est activée très rapidement au moment de la sortie de la phase stationnaire Ainsi, bien qu'il semble être moins répandu, le phénomène de pause au promoteur est retrouvé chez la levure. La distribution de la RNAP II chez les bactéries ressemble plus à celle de la drosophile et des mammifères que de la levure avec une accumulation importante de l'ARN polymérase pausée dans la partie proximale des gènes (Wade and Struhl, 2008).

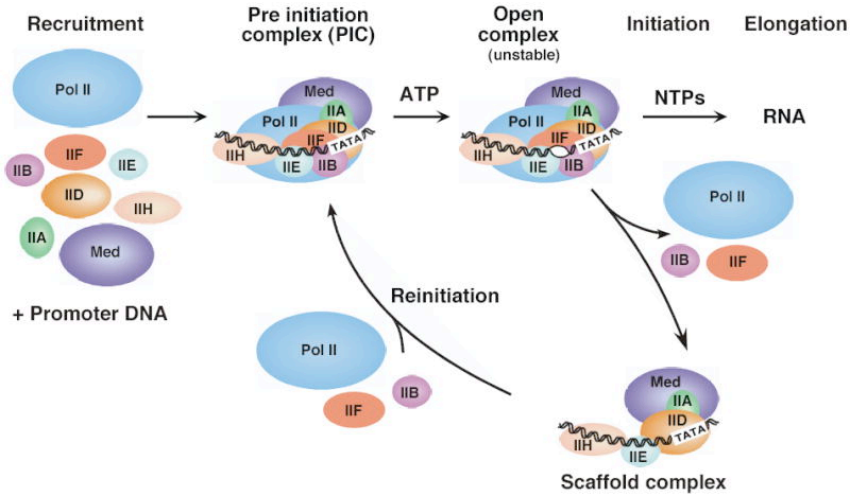


Figure 9. Mode de recrutement de l'ARN polymérase II au cours de l'initiation et de la réinitiation (d'après Hahn, 2004). Se référer au texte pour la description des différentes étapes.
Med : Médiateur

L'initiation de transcription par l'ARN polymérase II

La transcription des gènes commence avec la formation du complexe de pré-initiation le PIC au promoteur. Ce PIC contient RNAPII et les facteurs généraux de transcription, TFIIB, -D, -E, -F, et –H. Le facteur de transcription TFIIA servirait à stabiliser les interactions. La reconnaissance du promoteur par la machinerie de transcription est essentielle pour le positionnement et l'assemblage corrects de la

machinerie de transcription. Un promoteur inclut une combinaison des éléments suivants : une séquence TATA (TBP-binding site), un élément Initiateur Inr, le BRE (TFIIB-recognition element), et le DPE (downstream promoter element). La plupart des promoteurs comprenne un ou plusieurs de ces éléments, mais aucun n'est indispensable à la fonction du promoteur. Les éléments du promoteur servent de sites de liaison aux facteurs de transcription et permettent d'orienter la machinerie de transcription afin de diriger la transcription unidirectionnelle. Il existe pourtant un phénomène assez largement répandu, chez la levure, comme chez l'homme, la transcription divergente. La majorité des promoteurs engagés dans la transcription canonique dans la direction « sens », est également engagée en transcription « antisens », produisant des transcrits cryptiques instables (CUTs), d'environ 700 nt (Churchman and Weissman, 2011; Min et al., 2010; Seila et al., 2008; Seila et al., 2009). Cependant la plupart des promoteurs montre tout de même une forte préférence pour la transcription « sens ». De plus, il n'y a pas de corrélation entre le niveau de transcription directionnelle et de transcrits « antisens » (Churchman and Weissman, 2011). La transcription « anti-sens » n'aurait pas rôle à promouvoir la directionalité de la transcription.

La liaison de TBP à la boîte TATA provoque une courbure de l'ADN. Le recrutement de TBP au promoteur est le premier événement de l'initiation. L'élément BRE reconnu par TFIIB est déterminant pour l'orientation de la transcription. Les protéines TAFs (Transcription Associated Factor) de TFIID lient les séquences DPE et Inr.

Bien qu'initialement, la séquence TATA était considérée comme un élément contenu dans tous les promoteurs, il apparaît que ce type de promoteurs est en réalité minoritaire. Aux promoteurs dépourvus de la séquence TATA, la liaison de TBP à l'ADN serait non-spécifique, mais serait facilitée, de même que la courbure à l'ADN via l'interaction avec les autres facteurs de transcription.

TFIID sert ensuite de plateforme pour le recrutement successif des autres facteurs de transcription. TFIIA et TFIIB sont recrutés, leur liaison stabilise le complexe TBP-ADN. Un complexe formé par TFIIF et la RNAP II arrive ensuite. Enfin, TFIIE puis TFIIH viennent s'associer à la structure préexistante, et l'activité hélicase de TFIIH catalyse l'ouverture de la bulle de transcription. Le complexe de transcription est alors appelé complexe ouvert. La RNAP II est la seule ARN polymérase à recruter pour cela une activité hélicase ATP-dépendante. La sélection du TSS dépendrait de la liaison de TFIIB à la polymérase et au promoteur (Hahn, 2004).

Il faut cependant mettre un bémol à cette vue statique de la mise en place du PIC. L'expression et le recrutement différentiels de la famille des protéines TBP et des TAFs durant l'ontogénèse indiquent qu'il y aurait bien plus d'un modèle dans la reconnaissance des promoteurs de classe II, très variables dans leur composition (Cler et al., 2009; D'Alessio et al., 2009; Muller et al., 2010). D'autre part, plusieurs modes de recrutement menant à la constitution du PIC peuvent coexister, avec différents intermédiaires, *in vivo*. Le médiateur joue un rôle critique dans le recrutement de TFIIE et TFIIH. Le Médiateur pourrait soit

recruter la RNAP II, puis le complexe composé de TFIIH-TFIIE, ou alors former tout d'abord un complexe avec TFIIE-TFIIH, recrutant ensuite la RNAP II (Esnault et al., 2008).

Cette diversité de facteurs et de modes de constitutions du PIC étend encore le répertoire des mécanismes de régulation de l'expression des gènes de classe II.

Échappée du promoteur, promoter-escape, promoter-clearance : première étape de l'élongation

Les ARN polymérases en élongation productive peuvent transcrire le gène en entier sans se détacher de la matrice, ou relarguer le transcrit. Cependant, la RNAP II doit auparavant subir une maturation fonctionnelle et structurelle, pour passer de la phase d'initiation à la phase d'élongation. L'étape d'échappée du promoteur regroupe l'ensemble de ces processus, durant lesquels la RNAP II rompt les contacts avec les éléments du promoteur, et certains des facteurs de transcription. Le processus est considéré comme complété lorsque l'ARN naissant est associé stablement avec le complexe de transcription (Saunders et al., 2006). Le complexe de transcription, qui au départ est nommé ITC ou Initially Transcribing Complex, est maintenant appelé Early Elongation Complex ou EEC. Ce complexe stable présente toujours des caractéristiques différentes du complexe d'élongation mature. Au cours de cette étape, l'ITC subit plusieurs initiations abortives, où la RNAP II synthétise et relargue continuellement de petits ARNs, sans se détacher de la matrice ADN. La transition de l'ITC à l'ETC s'effectue une fois que la RNAP II a ajouté les huit premiers nucléotides, et que le site actif de la RNAP II est transloqué à la neuvième position (Saunders et al., 2006). La bulle de transcription « s'effondre », produisant une bulle de transcription d'une taille caractéristique de l'élongation.

La pause proximale au promoteur, « promoter-proximal pausing » : deuxième étape de l'élongation.

La pause au promoteur est un phénomène où l'ARN polymérase II pause au niveau de la région 5' de l'unité de transcription, et n'entre en élongation productive que sous l'impulsion de signaux appropriés. La pause au promoteur constitue une étape importante de régulation de l'expression des gènes *in vivo*, et fonctionne comme un point de contrôle avant l'élongation productive. La RNAP II peut échapper rapidement à la pause pour entrer en élongation, ce niveau de régulation permet de contrôler finement l'expression des gènes. De plus, cette étape peut être limitante et régulée même après l'induction de l'expression. Le phénomène de pause proximale a été identifié initialement pour les gènes de choc thermique Hsp70 chez la drosophile (Rougvie and Lis, 1988), et avait été mis en évidence au promoteur des proto-oncogènes *MYC* et *FOS* (Lis, 1998), chez les mammifères. Cette étape de pause semble aujourd'hui être une caractéristique de la majorité des gènes induits, ou exprimés constitutivement, chez la

drosophile et les mammifères (Guenther et al., 2007; Min et al., 2010; Muse et al., 2007; Rahl et al., 2010; Zeitlinger et al., 2007). La RNAP II est pausée, après avoir synthétisé un transcrit de 25 à 40 nucléotides environ.

Les facteurs impliqués dans la rétention de la polymérase au promoteur comprennent le facteur induisant la sensibilité au DRB (DSIF, DRB-sensitivity-inducing factor), et le Negative Elongation Factor (NELF) (Wu et al., 2003). DSIF est contitué des sous-unité Spt4 et Spt5, conservées de la levure à l'homme. NELF comprend quatre sous-unités, NELF-A, B, C/D et –E, est retrouvé chez la drosophile et les mammifères. La position de la RNAP II pausée corrèle avec celles de DSIF et NELF, *in vivo* (Rahl et al., 2010). Les mécanismes permettant à NELF et DSIF de retenir la RNAP II au promoteur ne sont pas connus. Pourtant, NELF lie l'ARN, il est possible que NELF exerce son activité de rétention en se liant à l'ARN (Fujinaga et al., 2004).

Ce phénomène semble affecter l'ensemble des gènes transcrits par la RNAP II, suggérant que ce mécanisme est important. Il a été proposé que la sortie de pause par la RNAP II dicte la cinétique d'activation des gènes lors de la différentiation des cellules souches. Les gènes pausés seraient les premiers activés. La pause permettrait de synchroniser l'activation des gènes lors du développement (Espinosa, 2010). Enfin, la pause de la RNAP II contrerait la répression nucléosomale, créant un domaine compétent pour la transcription, et ainsi permettant l'activation précise en réponse à un stimuli (Gilchrist et al., 2010).

L'élongation productive : troisième étape de l'élongation

L'échappé du promoteur et la pause de la RNAP II sont les deux étapes généralement limitantes de l'élongation. Plusieurs facteurs sont requis pour la transition de l'étape de pause au promoteur en élongation productive. Les effets négatifs de DSIF et NELF sont levés probablement sous l'action du facteur d'élongation P-TEFb (Positive Transcription Elongation Factor-b). Ce complexe comprend une cycline-dépendante kinase Cdk9, et une cycline partenaire cyclin-T (Peterlin and Price, 2006). P-TEFb phosphoryle DSIF, NELF et la sérine 2 du CTD. L'enzyme responsable d'ajout de la coiffe au messager aurait également un effet négatif sur DSIF et NELF. L'activité de TFIIS est aussi importante pour le passage efficace de la pause (Adelman et al., 2005).

Une fois la transition en élongation effectuée, NELF se dissocie du complexe de transcription, tandis que DSIF reste associé (Rahl et al., 2010). Le recrutement de P-TEFb aux gènes est une étape limitante et contribue à la régulation des gènes *in vivo* (Saunders et al., 2006).

Une classe de facteurs stimule également l'élongation et le passage au travers des sites de pause ; TFIIF, ELL (eleven-nineteen lysine-rich in leukemia), et Elongin (Saunders et al., 2006). ELL et l'élongin seraient recrutés lors des pauses au cours de la transcription, le long du corps du gène, alors que TFIIF est

un facteur de trancription général, recruté au cours de l'initiation. TFIIF, seul ne peut pas stimuler la transition de la RNAP II pausée, en l'élongation (Cheng and Price, 2007).

La terminaison de la transcription

La terminaison de la transcription par l'ARN polymérase II est une étape essentielle permettant d'éviter les interférences avec les gènes en aval. Le mécanisme de la terminaison est relié à celui de la maturation de l'extrémité 3' des transcrits. Deux modèles ont été proposés pour expliquer ce lien. Le modèle "anti-terminateur" suggère que les séquences de poly-adénylations de l'ARN entraînent une modification des facteurs associés à l'ARN polymérase, la rendant moins processive, conduisant ainsi à l'arrêt de la transcription. Un second modèle appelé "torpille" propose que le clivage du transcrit par la machinerie de poly-adénylation génère une nouvelle extrémité 5' qui serait le substrat d'une nucléase dont l'activité entraîne la dissociation de l'ARN polymérase II (Buratowski, 2005). Par ailleurs, un mode de terminaison alternatif a récemment été mis en évidence dans le cas des ARN non-codants. Celui-ci fait intervenir le complexe Ndr1, qui reconnait des sites spécifiques à l'extrémité 3' de l'ARN. Ces ARNs sont généralement plus ou moins rapidement dégradés par l'exosome et TRAMP (Lykke-Andersen et al., 2011).

La réinitiation de la transcription

Après l'initiation de la transcription par l'ARN polymérase II *in vitro*, un certain nombre de facteurs généraux de la transcription peuvent rester au niveau du promoteur dans un complexe appelé le "scaffold" (plateforme de réinitiation, ou échafaudage). Ce complexe, uniquement identifié *in vitro*, serait formé du Médiateur et des facteurs généraux de la transcription à l'exception de TFIIB et TFIIF. Ce complexe existe probablement dans le cas des gènes les plus transcrits et permettrait d'éviter l'étape relativement lente du recrutement des facteurs généraux de transcription pour les cycles de transcription ultérieurs. Le complexe "scaffold" pourrait alors recruter rapidement les facteurs généraux manquants et permettre la réinitiation de la transcription (Hahn, 2004). La réinitiation de la transcription de classe II dépend, de même que pour la transcription de classe III, d'un couplage entre les machineries d'initiation et de terminaison, facilite le recyclage de la RNAP II sur les mêmes matrices (Dieci and Sentenac, 2003).

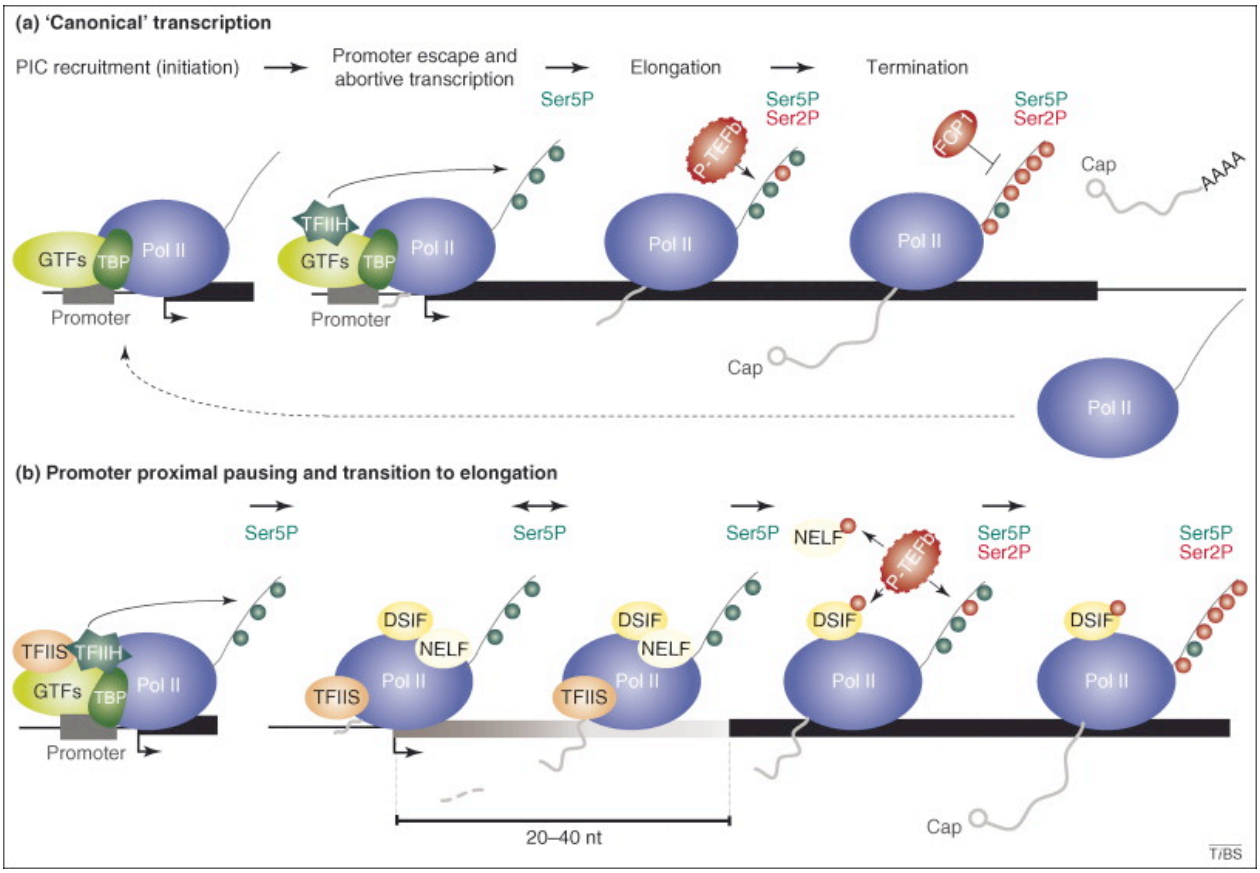


Figure 10. Schéma récapitulant l'ensemble des étapes de la transcription par l'ARN polymérase II (d'après Koch et al., 2008).

Les différentes étapes de la transcription dite « canonique » sont relativement bien caractérisées, tandis que les évènements de pause aux promoteurs sont moins bien connus. Toutes les étapes de la transcription sont représentées de la gauche vers la droite. (a) la transcription est initiée par le recrutement du PIC, contenant les facteurs de transcription, et la RNAP II (Pol II), au promoteur par la TBP. La kinase Cdk7, sous-unité de TFIIH phosphoryle la sérine 5 du CTD, facilitant l'échappée du promoteur. L'élongation est initiée par la phosphorylation de la sérine 2 par la kinase CDK9 du facteur d'élongation P-TEFb. Une fois au site de terminaison, la sérine 2 du CTD est déphosphorylée. L'ARN coiffé et polyadénylé est relargué, la RNAP II peut être recyclée. La phosphorylation de la sérine 5 décroît, tandis que la phosphorylation de la sérine 2 augmente le long du gène.

(b) La RNAP II est engagée entre transcription au promoteur proximal, produisant des transcrits entre 50 et 100 nucléotides. Cependant, seule la sérine 5 du CTD est phosphorylée. Ce processus de pause est régulé par les protéines DSIF et NELF, TFIIS. La transition nécessite le recrutement de P-TEFb, qui

phosphoryle la sérine 2 du CTD, et les facteurs NELF et DSIF. NELF se dissocierait, et l'élongation productive pourrait être amorcée.

Les modifications d'histones

Les histones composant les nucléosomes possèdent un domaine C-terminal globulaire, tandis que leur extrémité N-terminale n'est pas structurée. Cette extrémité est la cible de nombreuses modifications, la méthylation de l'arginine (R), la méthylation, l'acétylation, l'ubiquitination, la sumoylation, des lysines (K), la phosphorylation des sérines (S), et thréonines (T). Bien que la signification de ces modifications ne soit pas encore complètement comprise, certains profils caractéristiques sont associés à l'activation ou à la répression de la transcription.

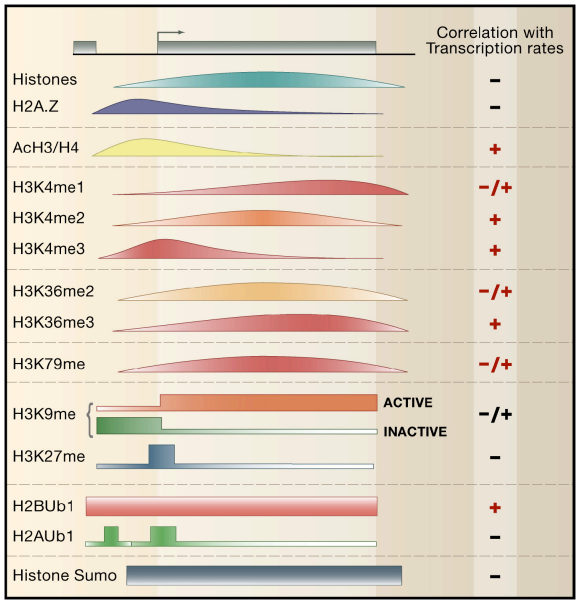


Figure 11. Profils des distributions des modifications d'histones autour des gènes (d'après Li et al., 2007).

La distribution des histones et de leurs modifications est cartographiée sur un exemple de gène de classe II, par rapport à son promoteur, l'ORF, et l'extrémité 3'. Les courbes représentent les profils qui ont été déterminées par des études sur l'ensemble du génome. Les carrés indiquent lorsque les études ont été réalisées sur un faible nombre de gène. A l'exception des données de K9 et K27, les études proviennent de la levure. Il faut en plus ajouter la marque H3K79me2, associée aux gènes transcrits, et présente au niveau du TSS, et jusqu'à 5 kb en aval.

L'acétylation et la déacétylation corrèlent avec l'accessibilité de la chromatine, et la transcription, tandis que la méthylation de la lysine exerce divers effets selon le nombre de groupement méthyl et le résidu ciblé. La triméthylation (me3) de H3K9 et H3K27 est associée à la répression, elle est retrouvée au niveau des gènes inactifs du génome. Les enhancers actifs présentent les modifications H3K4me1, et H3K27ac (Creyghton et al., 2010). Les gènes actifs possèdent un fort enrichissement en H3K4me3, qui marque le TSS. Cette marque est associée aux gènes liés par la RNAP II à leur promoteur, regroupant les gènes exprimés comme les gènes dits « non-productifs », ou « potentiellement actifs », car liés par la RNAP II pausée au promoteur, et n'étant pas entrée en élongation productive (Rahl et al., 2010). Une combinaison d'acétylation et de méthylation des H3K4, H3K36, et H3K79 est associée aux gènes transcriptionnellement actifs. La plupart des gènes liés par la RNAP II initient la transcription, mais seuls les gènes présentant les modifications H3K36me3 et H3K79me2 sont en élongation et produisent un transcrit mature. Ces deux marques ont observées au niveau du corps du gène, H3K79me2 est identifié dans la zone du TSS, jusqu'à 5 kb en aval, tandis qu'H3K36me3 est associé tout au long du gène transcrit. Contrairement à la marque de méthylation de ces résidus qui peut être identifiée à de faible niveau dans la chromatine inactive, l'acétylation est une marque exclusivement des zones actives (Kim et al., 2009).

B. Le facteur de transcription TFIIS

La synthèse du transcrit n'est pas un processus linéaire. Les pauses au cours de la transcription arrivent fréquemment le long du corps du gène. Ces sites de pauses semblent distribués de façon égale, après les 700 premiers nucléotides (Churchman and Weissman, 2011). Lorsque la RNAP II rencontre un obstacle, dont la nature n'est pas clairement définie, elle arrête l'élongation et recule (backtracking). Ce mouvement résulte en un mauvais alignement de l'extrémité 3' de l'ARN en cours de synthèse, avec le site catalytique de la polymérase. Pour reprendre l'élongation, la RNAP II hydrolyse le transcrit pour créer une nouvelle extrémité 3'. Ainsi, l'extrémité est de nouveau en phase avec le site actif, et de nouveaux nucléotides peuvent être ajoutés.

Ce phénomène de clivage a été initialement décrit chez *Escherichia coli* (Krummel and Chamberlin, 1989). Il a été proposé que cette activité hydrolytique était intrinsèque à la polymérase elle-même. Identifiée par la suite chez l'humain, cette activité est une fonction conservée au cours de l'évolution et est partagée par les ARN polymérases (Izban and Luse, 1992; Wang and Hawley, 1993).

Les facteurs jouant un rôle dans l'élongation peuvent être classés en plusieurs catégories ; ceux supprimant les pauses transitoires de la polymérase comme ELL, ou Elongin, et ceux capables de réactiver l'ARN polymérase arrêtée durant la transcription. cette dernière catégorie regroupe les facteurs

procaryotiques GreA et GreB, et le facteur eucaryote TFIIS. Une autre catégorie comprend les facteurs stimulant l'élongation via le remodelage de la chromatine comme le complexe SWI/SNF.

Les protéines TFIIS et Gre stimulent l'élongation en permettant à la RNAP II de passer au travers des sites de blocage, comme les nucléosomes ou des séquences ADN, provoquant la pause de la RNAP II (Kireeva et al., 2005). Certaines de ces séquences ADN ont été caractérisées, et présentent un enrichissement en nucléotides A-T, bien que ce ne soit pas la règle (Hawley et al., 1993). Les nucléosomes induisent une pause et le recul de la RNAP II au cours la transcription, *in vitro* (Kireeva et al., 2005), et *in vivo* (Churchman and Weissman, 2011)

Structure de TFIIS

a

b

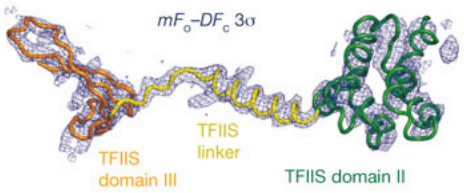


Figure 12. Architecture de TFIIS (d'après Kettenberger et al., 2003 et Cheung and Cramer, 2011).

Structure et organisation en domaines de TFIIS. (a) Le diagramme montre la répartition des trois domaines de TFIIS le long de la structure primaire de la séquence de la protéine. Les domaines I, II et III sont respectivement représentés en blanc, vert et orange. Le linker entre les domaines II et III est représenté en jaune. Les limites des domaines sont indiquées au dessus de la figure en numérotant à partir du premier acide aminé N-terminal. (b) Structure des domaines II, III et du linker.

TFIIS est un facteur composé d'un domaine I en N-terminal, un domaine central II, et un domaine III C-terminal. La structure des trois domaines est connue chez la levure. Chez les mammifères, la structure du domaine III est connue, et peut se superposer au domaine III de la levure. Le domaine I est

composé des 130 acides aminés de l'extrémité N-terminal. Les résidus 131-240 composent le domaine II. Une région « linker » de 19 nucléotides est présente entre les domaines II et III, celui-ci contenant les résidus 260-309 chez la levure (Fish and Kane, 2002).

Le domaine I, présentant le moins de conservation entre les organismes, n'est pas indispensable à l'activité de clivage de TFIIS *in vitro*. Ce domaine est cependant impliqué dans la transcription. Le domaine I est constitué de quatre hélices, lui conférant une forme globulaire. Il peut être phosphorylé au niveau de résidus sérine et thréonine, bloquant son activité de stimulation de clivage de la RNAP II (Fish and Kane, 2002).

Les domaines II et III sont les régions les plus conservées de TFIIS. De plus, ils sont nécessaires et suffisants à l'activité de TFIIS *in vitro* (Guo and Price, 1993). Le domaine II interagit avec la RNAP II. Il est composé de trois hélices, et une dernière structure adoptant approximativement la forme d'une hélice.

Le domaine III, partageant 61% d'identité entre la levure et l'homme, est requis pour l'activité de clivage de TFIIS. La structure du domaine III consiste en trois feuillets β anti-parallèles, composant le motif à ruban de zinc (zinc ribbon). Ce domaine est stabilisé par la présence d'un tétrade de cystéines liant un ion zinc. Le motif zinc ribbon a été identifié au niveau des sous-unités Rpb9 (RNAP II), et Rpc11 (RNAP III), Rpa12 (RNAP I). Le domaine III contient le motif RSADE très conservé parmi ces protéines, bien que plus faiblement chez Rpb9. Les résidus acides aspartique et glutamique sont particulièrement critiques à l'activité de clivage de TFIIS.

La région liant les domaines II et III est relativement flexible par rapport aux autres domaines. Cependant, il ne constitue pas une simple connexion, mais est nécessaire à l'activité de TFIIS. Ce domaine n'est pas structuré quand TFIIS est libre. Au contact de la RNAP II, il forme une hélice couvrant la surface de la RNAP II.

Structure du complexe de l'ARN polymérase II et de TFIIS

Le modèle établi par Kettenberger (Kettenberger et al., 2003) d'après la structure cristallographique, montre que TFIIS s'étend le long de la surface de la RNAP II, couvrant une distance de 100 Å. Le domaine II de TFIIS s'ancre au niveau du domaine constituant la mâchoire Rpb1/9 de la RNAP II, près du point d'entrée de l'ADN dans le sillon. Le linker s'étend sur la surface de la RNAP II jusqu'à l'entonnoir. Le domaine III s'ancre au niveau du pore par son domaine zinc. Un des feuillet β s'étend le long du pore jusqu'au cœur catalytique de la RNAP II.

Deux résidus acides au sommet de cette boucle contribuent à la coordination d'un second ion métal au niveau du cœur catalytique de la RNAP II. Cette liaison entraîne le clivage nucléolytique de type

Sn2 de la liaison phosphosdiester entre les ribonucléotides du transcrit. Cette boucle B est conservée au niveau de la sous-unité Rpc11. Les deux résidus acides sont également essentiels à l'activité de clivage de la RNAP III (Landrieux et al., 2006).

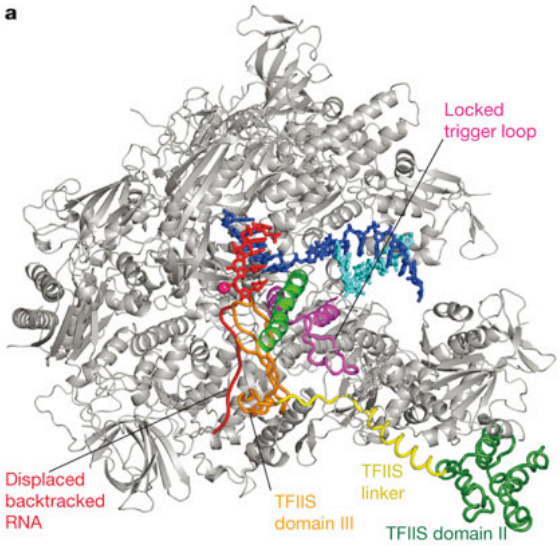


Figure 13. Structure du complexe de l'ARN polymérase II et de TFIIS (d'après Cheung et al, 2011) (Cheung and Cramer, 2011).

Structure de côté du complexe de réactivation intermédiaire entre l'ARN polymérase II et TFIIS, obtenue par rayons X. Les domaines II et III de TFIIS sont en vert et en orange respectivement, le linker est en jaune. Les brins codant, non-codant de l'ADN, et l'ARN sont représentés en bleu, cyan et rouge respectivement.

Lorsque la RNAP II rencontre un obstacle bloquant son élongation, ou en cas de mauvaise incorporation d'un nucléotide (proofreading), elle s'arrête et recule d'un nucléotide. Un recul plus important est entravée par une tyrosine (Gating tyrosine), située au niveau de l'hélice bridge, la RNAP II est capable d'hydrolyser seule, sans facteur de stimulation, un ou deux nucléotides de l'ARN, permettant de réaligner l'extrémité 3' de l'ARN. Cependant, si le recul est plus important, au-delà de cette tyrosine, l'ARN et la boucle sont piégés dans le pore, inhibant l'élongation, provoquant ainsi l'arrêt de la RNAP II. TFIIS est alors recruté pour réactiver la RNAP II, en déplaçant la boucle et l'ARN. TFIIS stimule ensuite le clivage de l'ARN, permettant à l'élongation de reprendre (Cheung and Cramer, 2011).

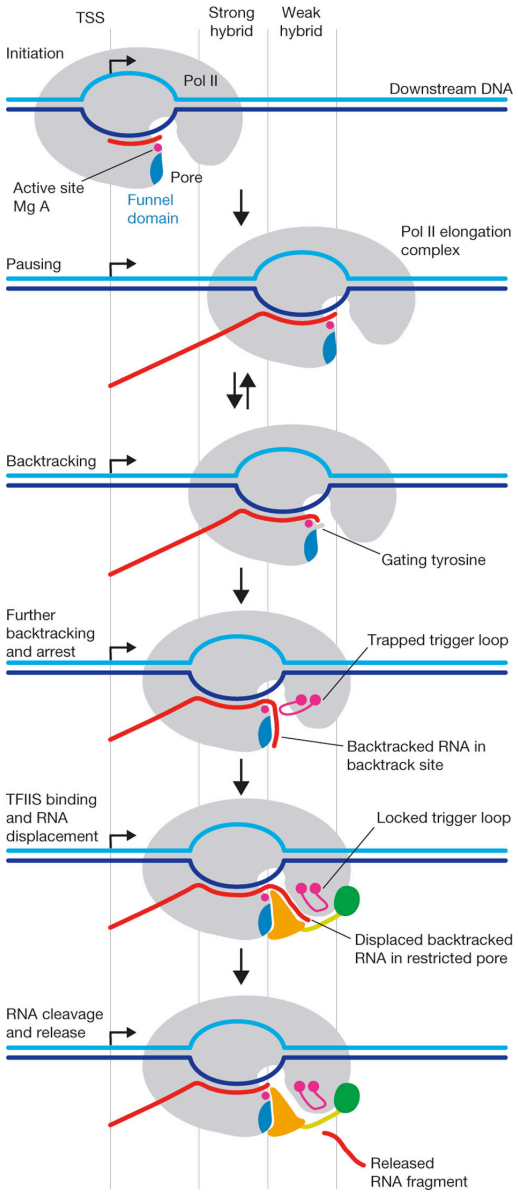


Figure 14. Schéma illustrant le mécanisme du recul (backtracking), de l'arrêt et de la réactivation de l'ARN Polymérase II (d'après Cheung and Cramer, 2011).

TFIIS nommée p37, SII ou RAP38 a été identifié initialement chez l'homme (Natori et al., 1973), et un peu plus tard chez la levure (Sawadogo et al., 1980; Sawadogo et al., 1981). TFIIS est fortement conservé dans le règne eucaryote et chez les archae (Hausner et al., 2000). Chez *S. cerevisiae*, TFIIS est codé par le gène *PPR2*, appelé *DST1*. De façon très surprenante, le gène *DST1* n'est pas essentiel à la survie de la cellule, soulignant que l'activité de clivage est intrinsèque à la RNAP II. Cependant cela ne signifie pas que l'activité de clivage du transcrit n'est pas essentielle, ou que l'élongation de la

transcription *in vivo* est un processus en réalité moins discontinue qu'observée *in vitro*. Une étude récente prouve au contraire que les obstacles et le backtracking de la RNAP II sont fréquents *in vivo*, et qu'ainsi l'activité de clivage intrinsèque de la RNAP II est essentielle à la survie de la cellule (Sigurdsson et al., 2010).

Il existe 3 isoformes de TFIIS chez la souris appelées Tcea1, 2 et 3. Ces trois protéines sont relativement mal caractérisées. Tcea1 semble être exprimée de façon ubiquitaire, Tcea2 est spécifique des spermatocytes (Ito et al., 1996), et Tcea3 est exprimée dans le foie, les reins et le cœur (Labhart and Morgan, 1998).

```
CLUSTAL 2.1 multiple sequence alignment


sp|P10711|TCEA1_MOUSE      --MEDEVVRIAKKMDKMVQKKNAAGALDLLKELKNIPMTLELLQSTRIGM 48
sp|Q9QVN7|TCEA2_MOUSE      MGKEEEIARIARRLDKMVTRKNAEGAMDLLRELKNMPITLHLLQSTRVGM 50
sp|P23881|TCEA3_MOUSE      MGLEEELLRIAKKLEKMVSRKKTEGALDLLKKLNSCQMSIQLLQTTRIGV 50
                              *:*: ***::::*** :*:: **:***::*:.  :::.***:**:*:

sp|P10711|TCEA1_MOUSE      SVNALRKQSTDEEVTSLAKSLIKSWKKLLDGPSTDKDPEEKK-------- 90
sp|Q9QVN7|TCEA2_MOUSE      SVNALRKQSSDEELIALAKSLIKSWKKLLD-VSDGKSRNQGR-------- 91
sp|P23881|TCEA3_MOUSE      AVNGVRKHCSDKEVVSLAKVLIKNWKRLLDSPRTTKGEREEREKAKKEKG 100
                           :**.:**:.:*:*: :*** ***.**:***     *. .: :

sp|P10711|TCEA1_MOUSE      --------------------KEPAISSQNSPEAREESSSSSNVSSRKDET 120
sp|Q9QVN7|TCEA2_MOUSE      --------------------GTPLPTSSSKDASRTTDLSCKKPDPPRTPS 121
sp|P23881|TCEA3_MOUSE      LGCSDWKPEAGLSPPRKKGGGEPKTRRDSVDSRSSTTSSPKRPSLERSNS 150
                                               *    ..         * .. .  :  :

sp|P10711|TCEA1_MOUSE      NARDTYVSSFPRAP----------------STSDSVRLKCREMLAAALRT 154
sp|Q9QVN7|TCEA2_MOUSE      TPR---ITTFPQVP----------------ITCDAVRNKCREMLTLALQT 152
sp|P23881|TCEA3_MOUSE      SKSKVETPTTPSSPSTPTFAPAVCLLAPCYLTGDSVRDKCVEMLSAALKA 200
                           .      .: *  *                 * *:** ** ***: **::

sp|P10711|TCEA1_MOUSE      GDDYVAIGADEEELGSQIEEAIYQEIRNTDMKYKNRVRSRISNLKDAKNP 204
sp|Q9QVN7|TCEA2_MOUSE      DHDHVAVGVNCEHLSSQIEECIFLDVGNTDMKYKNRVRSRISNLKDAKNP 202
sp|P23881|TCEA3_MOUSE      EDNFKDYGVNCDKLASEIEDHIYQELKSTDMKYRNRVRSRISNLKDPRNP 250
                            .:.   *.: :.*.*:**: *: :: .*****:************.:**

sp|P10711|TCEA1_MOUSE      NLRKNVLCGNIPPDLFARMTAEEMASDELKEMRKNLTKEAIREHQMAKTG 254
sp|Q9QVN7|TCEA2_MOUSE      GLRRNVLCGAITPQQIAVMTSEEMASDELKEIRKAMTKEAIREHQMARTG 252
sp|P23881|TCEA3_MOUSE      GLRRNVLSGAISPELIAKMTAEEMASDELRELRNAMTQEAIREHQMAKTG 300
                           .**:***.* *.*: :* **:*********:*:*: :*:**********:**

sp|P10711|TCEA1_MOUSE      GTQTDLFTCGKCKKKNCTYTQVQTRSADEPMTTFVVCNECGNRWKFC 301
sp|Q9QVN7|TCEA2_MOUSE      GTQTDLFTCNKCRKKNCTYTQVQTRSSDEPMTTYVVCNECGNRWKFC 299
sp|P23881|TCEA3_MOUSE      GTTTDLLRCSKCKKKNCTYNQVQTRSADEPMTTFVLCNECGNRWKFC 347
                           ** ***: *.**:******.******:******:*:***********

Sequences (TCEA1:TCEA2) Aligned. Score:  64
Sequences (TCEA1:TCEA3) Aligned. Score:  58
Sequences (TCEA2:TCEA3) Aligned. Score:  58
```

Figure 15. Alignement des séquences des trois isoformes de TFIIS chez la souris, avec ClustalW2 (http://www.ebi.ac.uk/Tools/services/web_clustalw2/toolform.ebi).

Les numéros d'accessions sont de la base Uniprot, TCEA1 P10711, TCEA2 Q9QVN7, et TCEA3 P23881. Les scores indiquent le pourcentage de similarité existant entre chaque isoforme.

Plusieurs études ont souligné l'importance de ce facteur au cours du développement ou dans des cellules cancéreuses. Une étude récente a mis en évidence que la diminution de l'expression du gène codant TFIIS réduit significativement la prolifération des cellules cancéreuses dans des lignées cellulaires de cancer du sein, des poumons et du pancréas (Hubbard et al., 2008). La diminution de l'expression de TFIIS dans des cellules du cancer du sein entraîne l'augmentation de l'expression des gènes c-Myc et p53. Chez la souris, TFIIS est impliqué dans l'hématopoïèse ; des souris « knock-out » pour TCEA1 ne peuvent se développer, les embryons n'arrivent pas à terme pour cause d'anémie, bien que ni la croissance, la différenciation ou le développement des embryons ne soient affectés avant le stade E13.5 (Ito et al., 2006).

Rôle de TFIIS dans la transcription de classe II

Rôle de TFIIS dans l'initiation et l'élongation

TFIIS joue un rôle au cours de l'initiation de la transcription. Ce rôle est distinct de son activité de stimulation de clivage. La délétion de *DST1* est co-léthale avec la délétion de la sous-unité *MED31* du Médiateur (Malagon et al., 2004). TFIIS est recruté au promoteur du gène *GAL1*, la perte de l'expression de TFIIS entraîne une diminution du recrutement de la machinerie de transcription au promoteur de ce gène (Prather et al., 2005). Enfin, TFIIS joue un rôle dans la formation du PIC. Ce rôle requiert le domaine II de liaison à la RNAP II (Guglielmi et al., 2007). TFIIS serait recruté par des activateurs ou des co-activateurs, comme le Médiateur, indépendamment de la RNAP II. En conjonction avec ces activateurs, TFIIS stimulerait ensuite le recrutement de la RNAP II, via une interaction directe par son domaine II. *In vitro*, le domaine I est également nécessaire à la formation du PIC (Kim et al., 2007).

Étant donné son rôle dans l'élongation, certaines études postulaient que TFIIS ne serait recruté qu'aux sites de pause de l'ARN polymérase II. Une fois la RNAP II réactivée, TFIIS se dissocierait du complexe d'élongation. L'analyse globale de l'occupation de TFIIS dans le génome de *S. cerevisiae* a été réalisée dans notre laboratoire (Ghavi-Helm et al., 2008). L'observation de l'occupation de TFIIS sur plusieurs gènes fortement transcrits par l'ARN polymérase II révèle un enrichissement constant tout au long de l'ORF dont le profil suit précisément celui de l'ARN polymérase II. Ce résultat va à l'encontre d'un précédent modèle selon lequel TFIIS ne serait recruté que lors d'une carence en nucléotides dans la cellule, bloquant l'élongation (Pokholok et al., 2002).

Rôle de TFIIS dans la transition de la pause au promoteur à l'élongation productive

Au cours de la pause au promoteur, la RNAP II recule et s'arrête (Nechaev et al., 2010). Lors de la reprise de l'élongation, l'extrémité 3' de l'ARN doit être réalignée avec le site actif de la RNAP II. L'induction de l'activité de clivage de l'ARN est effectuée par TFIIS. Bien qu'une première étude chez la drosophile ait suggérée que TFIIS permette à la RNAP II de quitter son état de pause au promoteur et de passer en élongation productive (Adelman et al., 2005), les mêmes auteurs ont récemment remis cette hypothèse en question. Le complexe d'élongation au niveau du promoteur proximal rencontre une région riche en A-T, causant de faibles interactions au sein de l'hybride ADN-ARN. La RNAP II se pause, et recule, vers une région en amont, riche en G-C, et ainsi thermodynamiquement plus stable, où elle est séquestrée. Cette région coïncide avec le DPE (Downstream Promoter Element) ou ce que les auteurs ont nommé "Pause Button Motif". Le mouvement de recul a délogé l'extrémité 3' de l'ARN du site actif. La RNAP II ne peut reprendre l'élongation. Les auteurs observent que l'ARN sorti du site actif, est clivé en condition normale. Cependant, lorsque l'expression du gène codant TFIIS est diminuée, bien que cette activité de clivage ne soit plus observée, la bulle de transcription ne se déplace pas en aval. Ces données suggèrent que TFIIS est bien nécessaire au passage de la pause, mais que ce facteur n'est pas suffisant, d'autres facteurs stimulant la transition doivent être recrutés (Nechaev et al., 2010).

Ces résultats ne vont néanmoins pas à l'encontre d'un rôle de TFIIS dans la transition de la pause à l'élongation. L'activité de TFIIS semble importante, les facteurs DSIF et NELF retenant la RNAP II au promoteur inhibent la liaison de TFIIS à la RNAP II *in vitro* (Palangat et al., 2005). A l'opposé, TFIIS serait impliqué dans l'ubiquitination de CDK9, sous-unité du facteur d'élongation P-TEFb, en recrutant UBR5, l'ubiquitine ligase E3, dans des cellules humaines. Cette ubiquitination n'entraîne pas la dégradation de CDK9, mais module son activité. Elle entraîne de plus son recrutement aux gènes de classe II, et de façon concomitante, une augmentation de la phosphorylation de la sérine 2 du CTD de la RNAP II (Cojocaru et al., 2011).

Rôle de TFIIS dans la transcription de classe III

La RNAP III possède une activité de clivage intrinsèque, dépendant du domaine C-terminal de Rpc11 (Chedin et al., 1998), similaire au domaine III de TFIIS. La sous-unité Rpc11 (RNAP III) possède deux domaines zinc ribbon. Elle contient également un motif SADE. Rpc11 est également similaire au facteur de clivage des archae TFS, qui possède également le motif SADE (Hausner et al., 2000). Ces

différentes données montrent que la RNAP III possède sa propre activité de clivage, et aucun facteur supplémentaire ne semble requis, contrairement à la RNAP II.

Ainsi, il était tout à fait inattendu que TFIIS chez *S. cerevisiae* soit retrouvé lié aux gènes de classe III activement transcrits (Ghavi-Helm et al., 2008). Des mutations altérant ou supprimant la liaison de TFIIS aux gènes de classe III, entraîne une diminution notable de la transcription de classe III. Cette découverte est paradoxale car l'activité de clivage de la RNAP III dépend de Rpc11 (Chedin et al., 1998). Le recrutement de TFIIS ne serait cependant pas lié à son activité de clivage, bien que le motif SADE soit requis. Des données *in vitro* suggèrent que TFIIS contribuerait à la sélection du site d'initiation de transcription. Ce rôle peut être mis en relation avec son rôle dans l'initiation de la transcription des gènes de classe II.

Chapitre IV : Le séquençage Haut-débit

La cartographie de la distribution des interactions protéines-ADN, et des marques épigénétiques est essentielle pour notre compréhension de la régulation transcriptionnelle. L'immunoprécipitation de chromatine est la technique de choix pour étudier la distribution des protéines sur le génome. Cette technique requiert une première étape de fixation des cellules, généralement par la formaldéhyde formant des pontages covalents protéine-ADN, suivie de la fragmentation de l'ADN, jusqu'à l'obtention de fragments de taille comprises entre 0,2 et 1 kb. Les protéines liées à l'ADN sont ensuite immunoprécipitées en utilisant un anticorps spécifique dirigé contre la protéine d'intérêt. Après réversion du cross-link ADN-protéine, l'ADN immunoprécipité est purifié. Les fragments issus du ChIP peuvent ensuite être identifiés par hybridation sur puces (microarray), permettant d'étudier à l'échelle du génome, les interactions protéine-ADN. Les sondes oligonucléotides sont aujourd'hui conçues pour couvrir l'ensemble du génome dans le cas des « tiling array » ou puces de très haute densité, ou ciblent des régions bien précises, comme les promoteurs, des chromosomes spécifiques, ou des familles de gènes.
Les développements techniques très rapides de ces dernières années ont permis l'émergence de nouveaux outils de séquençage à haut-débit, constituant la famille des NGS ou Next-Generation Sequencing. Les NGS ont été appliquées dans de nombreux domaines, du séquençage, ou re-séquençage des génomes entiers, le profil d'expression des gènes par séquençage des ARNm (RNA-seq), la caractérisation des sites hypersensibles à la DNase I…

L'immunoprécipitation de chromatine suivie du séquençage à haut-débit a été une des premières applications des NGS, les premières études ont été publiées en 2007 (Barski et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007). Dans le cas du ChIP-seq, les fragments d'ADN sont séquencés directement, au lieu d'être hybridés sur une puce. Les avantages de cette technique sont très nombreux sur le ChIP-chip ; elle présente une plus haute résolution, une meilleure couverture du génome, moins d'artefacts. La cartographie très fine des interactions ADN-protéine permet de dresser une liste de cibles plus précise, et une meilleure identification des motifs de liaison.

A. Principe et méthodes

La méthode de séquençage de Sanger est considérée comme technologie de « première génération ». Cependant, les limites de cette technologie (longueur des séquences, a priori requis pour le

design d'une amorce…) ont conduit au développement de nouvelles techniques de séquençage, les NGS ou Next-Generation Sequencing. Différentes plateformes existent à l'heure actuelle, chacune ayant mis en place sa propre méthodologie, la préparation des matrices, le séquençage, et la capture d'image, et l'analyse des données. Les NGS disponibles sur le marché sont le Roche/454, Illumina/Solexa, Life/APG, Helicos Biosciences, et l'instrument Polonator (Zhang et al., 2011). De nombreuses autres technologies sont en cours de développement

Je me suis, ici, focalisée sur le séquençage de type Illumina, puisque il a été utilisé dans notre étude, pour obtenir des informations quant aux autres technologies, le lecteur pourra se référer aux revues de M.Metzker (Metzker, 2010) ou J.Zhang (Zhang et al., 2011).

L'approche utilisée est le séquençage par synthèse (SBS, Sequencing-by-synthesis). L'ADN est tout d'abord fragmenté, puis immobilisé sur un support. La plupart des systèmes d'imagerie ne peuvent détecter un seul événement de fluorescence, les matrices ADN doivent donc être amplifiées.

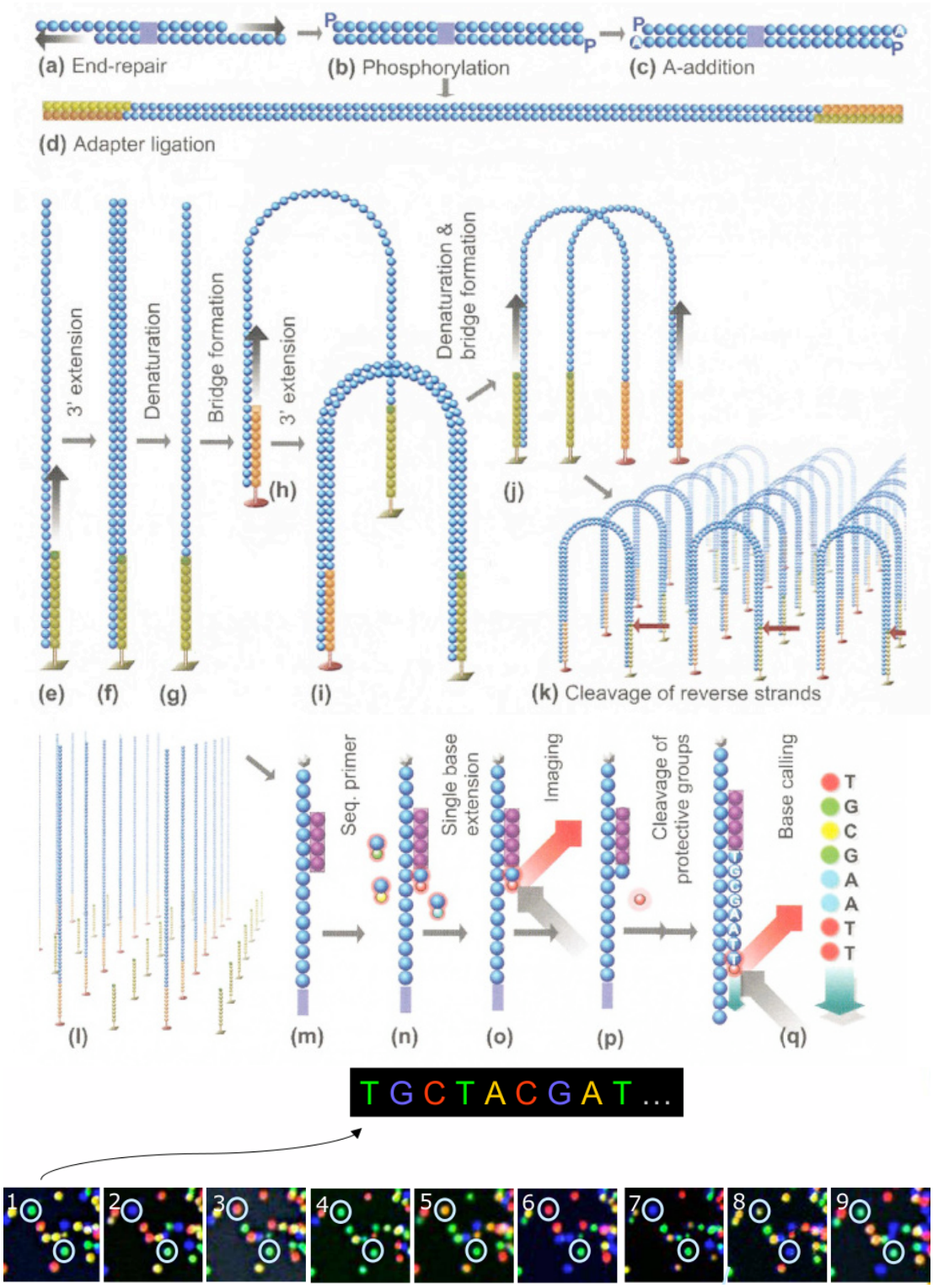


Figure 16. Description du processus de Séquence par Synthèse (d'après Lakdawalla et Van Steenhouse, 2008).

Préparation de la librairie : les extrémités des fragments d'ADN sont réparées (a), phosphorylées (b), et une adénine est ajoutée (c). Enfin, des adaptateurs directs et inverses sont ligasés (d). Formation des

colonies : les fragments d'ADN sont dénaturés, puis hybridés aux adaptateurs sur la « flow-cell » (e), étendus (f), puis dénaturés (g). L'ADN simple brin forme un pont en se liant aux amorces proches, liées à la « flow-cell » (h), de nouveau est étendu, formant un double brin (i), puis dénaturé et réhybridé pour former un nouveau pont (j). Le processus est répété plusieurs fois, formant une colonie de plusieurs milliers de fragments identiques (k). Les brins inverses sont clivés, libérant l'extrémité 3' (k). L'extrémité 3' est bloquée, et les amorces du séquençage s'hybrident aux brins (m). Séquençage : l'ADN polymérase incorpore un nucléotide terminateur réversible (n), l'image de la fluorescence est capturée (o), le fluorophore et le terminateur sont clivés (p). Le processus est répété sur plusieurs cycles.

Chez Illumina, l'amplification est réalisée sur un support solide, la flow-cell. C'est l'étape de « solid-phase amplification ». Des adaptateurs directs ou inverses sont ligasés aux extrémités des matrices. Ces mêmes adaptateurs sont fixés covalemment sur la flow-cell. Ces adaptateurs ou amorces directes ou inverses, répartis aléatoirement sur la surface de la flow-cell, permettent aux matrices de s'hybrider sur le support. Une première étape d'extension reconstitue le brin complémentaire de la matrice. Après dénaturation, ce simple brin s'hybride via son adaptateur en 3' à l'amorce adjacente fixée sur la flow-cell, et le brin complémentaire est synthétisé. Cette étape de « bridge amplification » est répétée plusieurs fois. Cette amplification clonale résulte en une population de matrices identiques, regroupés physiquement sur la flow-cell, formant des colonies (clusters). Une amorce peut ensuite s'hybrider aux extrémités 3' de ces matrices pour l'étape du séquençage. Illumina utilise la méthode « Cyclic reversible terminator » (CRT). Une ADN polymérase liée à la matrice ajoute un seul nucléotide complément de la base de la matrice. Ce nucléotide auquel est fixé un fluorophore, est modifié de façon à bloquer l'addition du nucléotide suivant. Après l'incorporation, les nucléotides non fixés sont éliminés par un lavage. L'image de la fluorescence de chacun des quatre fluorophores est capturée, permettant de déterminer l'identité du nucléotide ajouté à une colonie donnée. Le terminateur et le fluorophore sont clivés, un nouveau lavage permet d'enlever les fluorophores et terminateur, avant de passer au cycle suivant, où l'élongation peut reprendre. Actuellement, le nouveau Genome Analyser HiSeq 2000 peut séquencer des fragments de 100 pb, et générer jusqu'à 200 giga bases par cycles d'utilisation.

Le ChIP-seq offre bien des avantages par rapport au ChIP-chip. Premièrement, contrairement au ChIP-chip, le ChIP-seq permet d'obtenir une résolution à la base près. Bien que les sondes des puces « tiling » puissent couvrir le génome entier, dans le cas des mammifères, il faudrait utiliser un très grand nombre de puces, pour accéder à l'ensemble de leur génome, multipliant ainsi les coûts. Les puces sont également limitées en résolution du fait de contraintes de l'hybridation. L'hybridation des acides nucléiques est complexe et dépend de nombreux facteurs, comme le taux en GC, la longueur, la

concentration et la structure secondaire des cibles et des sondes. Un autre avantage notable de l'utilisation du ChIP-seq est qu'il est possible de couvrir une surface plus étendue du génome. La couverture du ChIP-seq n'est pas limitée par le répertoire des sondes fixé sur la puce. Ceci est d'autant plus vrai si l'on considère les génomes plus complexes des mammifères. Ceux-ci sont composés à plus de 50% de séquences répétées, or ces régions sont typiquement masquées sur les puces. Dans le cas du ChIP-seq, les variations minimes de séquences au sein des répétitions peuvent être capturées, et utilisées pour cartographier les lectures. De même, les séquences uniques flanquant les répétitions peuvent aider à la cartographie des lectures. Ainsi, par exemple, des séquences de 30 nt sont suffisantes pour cartographier 80 % du génome, et jusqu'à 90% si les séquences sont de 70 nt (Rozowsky et al., 2009).

Toute technologie présente bien évidemment un certain nombre d'artefact. Le ChIP-seq n'y échappe pas. Bien que les erreurs de séquençage aient été réduites, il existe toujours un biais spécialement à la fin de chaque lecture. Ce problème peut être contourné par les algorithmes d'alignements, qui n'aligneront qu'une partie de la séquence, dénommée graine (« seed »), et ne correspondant qu'aux x premières bases de la séquence. Cette méthodologie entraîne tout de même une perte d'information. Le biais des séquences riches en G-C influence également au moment de la préparation de la librairie (PCR), et au cours du séquençage, pour le nombre de lectures d'une séquence donnée.

B. Schéma expérimental

Qualité de l'anticorps

La qualité de n'importe quelles données de ChIP dépend fondamentalement de la qualité de l'anticorps utilisé. Un anticorps spécifique et sensible permettra d'obtenir un bon enrichissement par rapport au bruit de fond. Une validation rigoureuse par Western Blotting est nécessaire, et dans le cas d'anticorps ciblant des modifications d'histones très proches, la réactivité croisée doit être vérifiée, en utilisant par exemple la spectrométrie de masse avec les peptides modifiés.

Qualité de l'échantillon

Un avantage certain du ChIP-seq est la très faible quantité d'échantillon requise pour le séquençage. Pour la plateforme Illumina, 10 à 50 ng d'ADN sont recommandés, pouvant même descendre à 2 ng, tandis que le ChIP-chip requiert plus de 2 µg de matériel de départ. La quantité d'ADN et le nombre de cellules requis sont néanmoins dépendants de l'abondance de la chromatine associée au facteur ciblé, et de la qualité de l'anticorps.

Contrôle de l'expérience

Les étapes du protocole de ChIP engendrent de nombreux artefacts, comme l'étape de fragmentation de la chromatine. La sonication ou la digestion Mnase ne résultent pas en une fragmentation uniforme du génome. Les régions « ouvertes » du génome sont plus facilement fragmentées, contrairement aux régions dites fermées, créant une distribution inégale des lectures. Les régions répétées semblent également être enrichies à cause du manque de précision du nombre de copies des répétitions dans les assemblages des génomes. Il est important également de souligner que le ChIP consiste en un enrichissement, et non pas en une purification des sites liés par le facteur. Ceci est spécialement vrai dans le cas d'une seule étape d'immunoprécipitation, avec un anticorps spécifique. La majorité des fragments d'ADN immunoprécipités et donc des lectures sera du bruit de fond, tandis qu'une minorité constituera les fragments spécifiques. La distribution des lectures du bruit de fond dépendra de la taille et de la composition du génome séquencé. C'est pourquoi un pic d'une expérience de ChIP-seq doit être comparé à une même région d'une expérience contrôle, afin de déterminer sa validité. Trois types de contrôles sont communément utilisés dans les expériences de ChIP : l'input ou ADN total, avant immunoprécipitation, la « mock » IP, ou une IP réalisée sans anticorps, et enfin, une IP réalisée avec un anticorps non-spécifique (comme l'immunoglobuline G). Il n'existe aucun consensus quant au choix du contrôle le plus approprié, chacun produisant son lot d'artefact. L'input d'ADN a été utilisé dans la plupart des analyses ChIP-seq. Il permet de déterminer les zones enrichies de façon aspécifique, du fait de la fragmentation inégale, ou des variations d'amplification lors de la préparation de la librairie. Cependant, il est nécessaire de séquencer en profondeur, car les lectures se répartissant sur le génome en entier, les biais seront plus difficiles à localiser si le séquençage n'est pas suffisant. Le contrôle de type « mock IP » ne permet d'immunoprécipiter que très peu d'ADN, conduisant à des variations entre les contrôles eux-mêmes. La distribution du bruit de fond est souvent déterminée empiriquement, cependant il peut être modélisé, par exemple suivant une loi de Poisson, à partir de l'échantillon lui-même (Mikkelsen et al., 2007). Enfin, quelque soit l'approche utilisée, il faut souligner que la distribution des lectures du bruit de fond n'est pas uniforme, ni identique selon le tissu ou le type cellulaire, et dépend même de l'expérience en elle-même, et du protocole.

Profondeur de séquençage

Le succès d'une expérience de ChIP-seq dépend tout d'abord de la qualité de l'immunoprécipitation, de façon à obtenir un enrichissement suffisant par rapport au bruit de fond non-spécifique, et de la complexité de la librairie générée à partir de l'ADN immunoprécipité. Pour une expérience de ChIP-seq, l'unité de base du séquençage est une ligne de « flow-cell » ; au tout début, 4 à 6 millions de lectures étaient générés par le Genome Analyser avant alignement. Aujourd'hui, il est possible

d'obtenir plus de 30 millions de lectures. De plus, le nombre de sites occupés, la taille des régions enrichies, et la gamme d'enrichissement du ChIP affectent le nombre de lectures nécessaires. Si une protéine se liant à l'ADN présente un grand nombre de sites, ou si une modification particulière d'histone couvre une large fraction du génome, le nombre de lectures correspondant devra être grand pour couvrir chaque site de liaison, avec la même densité de lectures. Afin de déterminer si la profondeur de séquençage a été atteinte, un critère raisonnable est que l'augmentation du nombre de lectures séquencées ne change pas les résultats. En termes de nombre de sites de liaison, ce critère traduit l'existence de « point de saturation », au-delà duquel aucun site de liaison supplémentaire ne sera identifié, malgré une profondeur de séquençage supérieure. Ce point de saturation existe si un seuil d'enrichissement entre les régions et le contrôle est fixé, et si seuls les pics présentant un nombre minimal de lectures sont considérés. Si tous les pics sont considérés, même ceux avec un nombre très faible de lectures, lorsque la profondeur de séquençage est augmentée, deviendront statistiquement significatifs.

Multiplexage

Pour les petits génomes, comme celui de la levure *S. cerevisiae*, le nombre de lectures obtenu pour une ligne d'une « flow-cell » Illumina peut être bien supérieur au nombre de lectures nécessaires pour atteindre la couverture du génome. Le nombre de lectures continuant à augmenter au fur et à mesure que la technique s'améliore, il est maintenant possible de séquencer sur une même ligne plusieurs échantillons. Cette technique permet d'augmenter le nombre d'échantillons séquencés, tout en diminuant le coût.

Pour préparer les échantillons pour le multiplexage, une étiquette avec un identifiant unique est ajoutée à chaque librairie. Lors du séquençage, cet identifiant unique permet de distinguer la provenance de chaque séquence, et de la ré-attribuer à l'échantillon d'origine.

Paired-end

Les fragments de ChIP-seq sont généralement séquencés seulement à une extrémité (séquençage « single read»). Ils peuvent cependant être séquencés aux deux extrémités (séquençage « paired end »). Ce type de séquençage n'est habituellement pas utilisé dans le cas du ChIP-seq, il est plutôt réalisé pour détecter les variations structurelles (insertions, délétions, larges réarrangements chromosomiques) du génome. Dans le cas du ChIP-seq, il peut être utile pour cartographier les lectures dans les séquences répétées, ou si l'on recherche des interactions à distance (Fullwood et al., 2010).

C. Analyse des données

Alignement

L'acquisition des images, et l'étape d'assignation des bases (« base calling ») sont propres aux plateformes, qui utilisent des logiciels spécifiques. L'alignement des lectures sur le génome est au contraire une étape clé au cours du traitement des données, puisque l'ensemble des résultats découle de ces alignements. De nombreux algorithmes ont été développés, et chaque outil est un compromis entre rapidité, usage de la mémoire, et flexibilité (Trapnell and Salzberg, 2009). Les aligneurs doivent autoriser un certain nombre de mésappariements, tenant compte des erreurs de séquençage, des polymorphismes d'un seul nucléotide (SNPs, « Single Nucleotide Polymorphism »), et des différences existant entre le génome de référence et le génome d'intérêt. Les aligneurs les plus populaires sont Eland ou Efficient and fast aligner for short lecture, développé par Illumina, intégré à leur système d'analyse, Mapping and Assembly with Qualities (MAQ) (Li et al., 2008), un outil très utilisé pour les analyses NGS, excellent pour détecter les SNPs et Bowtie (Langmead et al., 2009). Ces méthodes peuvent utilisés les scores de qualité associés aux séquences, qui indiquent la fiabilité de chaque base. La plupart des aligneurs ne conservent pas les lectures non-uniques, s'alignant aux régions répétées du génome, cependant il existe des méthodes pour attribuer, avec une certaine confiance une lecture non-unique à une région, en tenant souvent compte de son environnement, c'est à dire des lectures flanquantes.

Identification des régions liées

L'étape suivante consiste à détecter les régions enrichies en lectures de façon significative par rapport au contrôle. Plusieurs programmes appelés « peak callers » parcourent le génome à la recherche de ces régions enrichies (Fejes et al., 2008; Jothi et al., 2008; Kharchenko et al., 2008; Robertson et al., 2007; Rozowsky et al., 2009; Valouev et al., 2008; Zhang et al., 2008). Les premiers algorithmes se basaient sur un simple décompte des lectures dans une fenêtre de taille déterminée, et selon un certain enrichissement par rapport au contrôle. Les algorithmes suivants ont pris en compte la directionalité des lectures. L'une des extrémités 5' ou 3' des fragments est séquencée, ainsi la localisation des lectures forme deux distributions, une sur le brin sens, l'autre sur le brin antisens, avec une distance constante entre les pics de chaque distribution. La méthode de ces algorithmes consiste à construire un profil de la distribution des lectures sur chaque brin, puis à combiner ces deux profils en un seul, déterminant la région de liaison du facteur (Boyle et al., 2008; Valouev et al., 2008). Ce profil est construit en décalant chaque distribution vers le centre ou en étendant les brins de façon orientée, et en joignant ces fragments ensemble. Cette dernière approche est normalement la plus précise, mais elle implique une estimation de la taille d'origine

des fragments séquencés, tout en faisant l'hypothèse que cette taille soit uniforme. Les algorithmes actuels ignorent les faux-positifs comme les agrégations ou empilements de lectures.

Une des difficultés majeures dans l'identification des régions réside dans la forme des pics : plutôt fins, couvrant quelques centaines ou moins de nucléotides, larges localisés sur quelques kilobases, ou très larges couvrant plusieurs kilobases, et enfin un mélange de tous les types. Les pics fins sont plutôt associés aux facteurs de transcription ou aux modifications d'histones présentes au niveau d'éléments régulateurs. Les profils « mélangés » de type ponctués ou larges mais localisés sont associés aux protéines comme la RNAP II. La liaison de la RNAP II au niveau du promoteur est plutôt est caractérisée par un pic fin, tandis que les pics au corps du gène sont plus étendus, et d'intensité beaucoup plus faible. Les pics larges couvrant de très grandes régions sont retrouvés pour les modifications d'histones marquant les domaines comme les régions réprimées ou transcrites. Les peaks callers sont généralement optimisés pour détecter préférentiellement un type de pic.

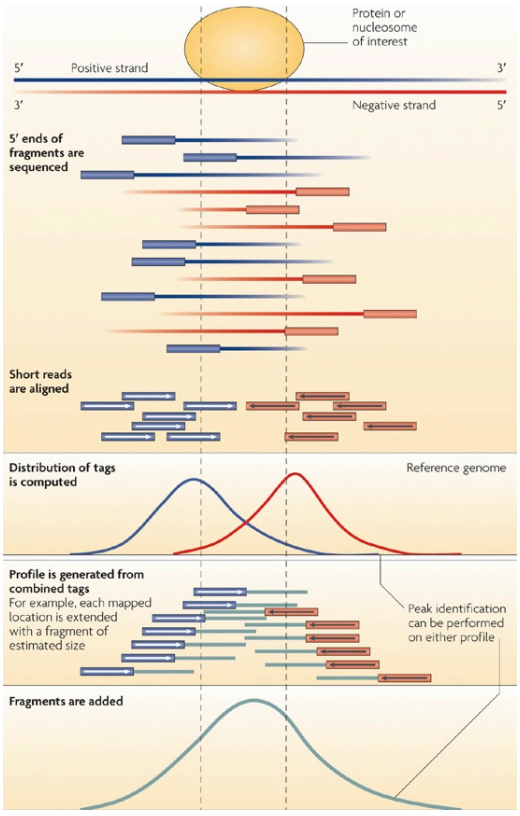


Figure 17. Distribution des lectures sur le génome (d'après Park,PJ, 2009 (Park, 2009)).
Les fragments d'ADN issus de l'immunoprécipitation d'ADN sont séquencés en 5'. L'alignement des lectures résulte alors en deux pics, un sur chaque brin, flanquant le site de liaison du facteur étudié. Cette distribution brin-spécifique peut être utilisée pour détecter les régions enrichies. Pour créer un profil de

distribution approximatif de chaque fragment, chaque lecture peut être étendu à partir de sa coordonnée 5', d'une taille équivalent à celle du fragment de départ par le programme. Le nombre de fragments est alors décompté pour chaque position.

Analyse en aval

Le ChIP-seq nous révèle souvent bien plus que ce que nous avons pu nous demander au départ. Il y a plusieurs façons d'aborder les implications biologiques des résultats de ChIP-seq. Pour les protéines se liant à l'ADN, l'analyse la plus classique consiste à identifier les motifs de liaison de la protéine. Différents outils sont disponibles sur internet, ou peuvent être installé en local, comme MEME (Bailey et al., 2009), Tmod (Sun et al., 2010), RSAT (Thomas-Chollier et al., 2008). Parfois, plusieurs motifs peuvent être identifiés. L'analyse de ces motifs peut contribuer à révéler des interactions du facteur étudié avec d'autres protéines se liant à l'ADN. Le ChIP-chip a permis d'identifier de nombreux motifs de liaison, cependant le ChIP-seq fournit une résolution plus fine, la région de liaison du facteur étant identifiée plus précisément.

Une autre étude assez basique suivant une expérience de ChIP-seq passe par l'annotation des régions liées, comme la localisation dans des introns-exons, régions intergéniques, etc. De même, il est intéressant de rapporter les sites liés à une échelle relative. Par exemple, une même échelle peut être adoptée pour les gènes liés par le facteur, afin de déterminer le profil de distribution global du facteur (analyse « gene profil » de seqMINER (Ye et al., 2011)). Enfin, pour trouver les relations existant entre les profils de différents facteurs, une analyse de corrélation, ou une approche par « clustering » peut être réalisée (Hon et al., 2008; Ye et al., 2011). Les données de ChIP-seq peuvent être corrélées à des données d'expression, obtenues par puces transcriptomiques ou RNA-seq. Si le taux d'expression d'un gène corrèle avec la liaison d'un facteur, ceci peut suggérer une régulation directe de ce gène par ce facteur. Lorsque une modification d'histone est associée à un ensemble de gènes dont l'expression est activée ou réprimée, alors il peut être inféré que cette marque corrèle avec un état transcriptionnel particulier. Une analyse par Gene Ontology analysis peut permettre de savoir si parmi les gènes liés par un facteur, une fonction moléculaire, ou un processus biologique particuliers sont représentés.

Pour se documenter plus en détail, se référer aux revues suivantes (Farnham, 2009; Hawkins et al., 2010; Metzker, 2010; Park, 2009; Pepke et al., 2009; Zhang et al., 2011) .

RESULTATS

Contexte de l'étude

Notre connaissance de la transcription de classe III provient d'études biochimiques et génétiques, réalisées au cours de ces 40 dernières années. Le recrutement des facteurs de transcription, l'organisation des promoteurs, les types de gènes sont bien caractérisés. Le transcriptome de classe III a été extensivement défini chez la levure *S. cerevisiae* (Harismendy et al., 2003; Moqtaderi and Struhl, 2004; Roberts et al., 2003). Ces études ont permis de dresser un panorama complet des gènes de classe III, et d'établir clairement les facteurs recrutés. Elles ont de plus, révélé l'existence de sites uniquement liés par le facteur de transcription, dénommés Extra-TFIIIC loci (ETC) par Moqtadery. Jusque très récemment, le même type d'études était très difficilement réalisable chez les mammifères. L'utilisation de puces spécifiques générait un biais, puisque les puces ne couvraient pas l'étendue de leur génome. Une étude plus large impliquait l'utilisation de « tilling array », dont le prix est plutôt prohibitif. Ainsi, en dépit du rôle essentiel de la RNAP III, la définition des gènes de classe III chez les mammifères se limitait essentiellement à l'étude de quelques gènes ou famille de gènes. La localisation des gènes de classe III résulte de prédictions bioinformatiques, puissantes, basées sur la conservation très forte des séquences et des structures secondaires entre les espèces et au sein même des familles, mais n'établissant pas la réalité des sites liés par la machinerie de classe III. Le transcriptome chez les mammifères est donc très mal caractérisé, il est difficile distinguer les gènes réellement transcrit par la RNAP III, des pseudogènes. En outre, en dehors de l'ensemble des transcrits bien caractérisés, comme les gènes d'ARNt, ou les gènes d'ARN U6, il a été suggéré que la RNAP III est responsable de la transcription des miARNs, ou d'ARNs de fonction inconnue, dont le promoteur est de type III (Pagano et al., 2007). En parallèle, les mammifères présentent des spécificités. Deux formes de TFIIIB ont été décrites, chacune associée à un type de promoteur, et par conséquence un ensemble particulier de gènes. Cependant, très peu de données *in vivo* concernant le recrutement de la RNAP III et de ses facteurs de transcription, chez les mammifères, sont disponibles.

Les génomes des mammifères contiennent de très nombreux éléments répétés. Les SINEs composent une de ces classes. Les SINEs contiennent un promoteur de classe III, la dépendance de la RNAP III pour leur transcription avait été démontrée *in vitro*. Un très grand nombre de ces éléments est décrit dans le génome, près d'un million pour les SINEs. Cependant, il n'est pas défini si l'ensemble était transcrit, ou si seul un sous-groupe était exprimé.

Enfin, chez *S. cerevisiae*, ont été découvert les sites ETC. Leur rôle d'insulateur a été caractérisé chez *S. pombe*. Cette fonction est conservée entre les différentes espèces de levure. La conservation des facteurs de transcription, des sites de liaison de ses facteurs, et des mécanismes en général, entre la levure

et les mammifères, nous ont conduit à rechercher si de tels sites de recrutement de TFIIIC existaient chez ces derniers organismes.

Finalement, une étude menée dans notre laboratoire illustre un rôle tout à fait inattendu du facteur d'élongation TFIIS chez *S. cerevisiae*. Ghavi-Helm a montré que ce facteur est recruté au niveau de tous les gènes de classe III. De même que pour les sites ETC, nous avons cherché si ce recrutement était conservé chez les mammifères.

Plusieurs questions ont ainsi été à l'origine de notre projet, (i) Peut-on identifier les gènes d'ARNt transcrits ? (ii) Peut-on identifier de nouveaux transcrits de classe III ? (iii) Peut-on identifier les éléments SINEs transcrits ? (iv) La transcription des éléments SINEs est-elle plus ou moins active que la transcription des autres gènes de classe III ? (v) Quels sont les gènes dont la transcription dépend de Brf2 ? (vi) Existe t'il chez la souris, des régions uniquement liées par TFIIIC, similaires aux ETC ? (vii) TFIIS est-il un facteur de transcription de classe III comme chez *S. cerevisiae* ?

Dans cette étude, nous avons déterminé la distribution de la RNAP III, de Brf1, Brf2, du complexe TFIIIC, et du facteur de transcription TFIIS, sur l'ensemble du génome murin des cellules souches embryonnaires. Afin d'identifier précisément les régions liées par ces différentes protéines, nous avons utilisé la méthode du ChIP-seq, qui combine l'immunoprécipitation de chromatine au séquençage haut-débit.

Chaque gène codant une protéine d'intérêt a été fusionnée en 3', juste avant le codon stop à une séquence codant pour une étiquette, par recombinaison homologue en cellules ES. Cette étiquette code pour trois épitopes : 6 histines, flag, et HA. Nous avons ainsi pu développer un protocole d'immunoprécipitation en tandem hautement spécifique, applicable à toutes les protéines possédant ces épitopes. Leur utilisation nous a également permis de nous affranchir des anticorps contre les protéines d'intérêt, parfois peu spécifiques. Nous avons choisi de travailler en cellules ES, afin d'exploiter leur capacité à se différencier, ainsi que pour la transmission germinale, afin de pouvoir étendre, par la suite, notre étude à d'autres types cellulaires, ou organes. Ces cellules, de plus sont particulièrement bien adaptées au recombineering car elles présentent un fort taux de recombinaison.

Nous avons construit des lignées de cellules ES, dans lesquelles nous avons étiqueté des protéines de la RNAP III, RPC1 et RPC4. Deux immunoprécipitations de chromatine indépendantes ont été menées dans ces lignées. Par la suite, lorsque nous avons défini une région comme étant liée par la RNAP III, nous n'avons retenu que les régions liées simultanément par ces deux protéines, afin d'être le plus spécifique possible. Nous avons procédé à l'immunoprécipitation de la chromatine associée à Brf1 et Brf2, pour définir les gènes dépendants de l'une ou l'autre forme de TFIIIB.

Afin d'étudier plus précisément le rôle de TFIIIC chez les mammifères, nous avons étiqueté trois sous-unité du complexe. Nous avons immunoprécipité et séquencé l'ADN pour chacune de ces trois sous-unités. Nous avons défini les régions enrichies ne retenant que celles liées simultanément par ces trois protéines. Le profil de liaison de TFIIIC est clairement distinct de celui de la Pol III. Tandis que ce facteur est présent au niveau des gènes d'ARNt et des nouveaux gènes, environ 2200 sites sont dépourvus du reste de la machinerie de classe III. De même que dans les cellules humaines, ces sites uniquement liés par TFIIIC, ou ETC, sont colocalisés avec la protéine CTCF. Cette protéine a un rôle bien connu d'insulateur. Cette proximité entre ces deux protéines pourrait souligner une interdépendance ou un rôle commun dans l'établissement d'une région frontière. Le complexe de la cohésine occupe les sites liés par CTCF, et contribue à la fonction insulatrice de CTCF. De façon remarquable, nous observons que les sous-unités Smc1 et Smc3 du complexe de la cohésine colocalisent avec CTCF et les sites ETC. Ces observations permettent de postuler que le rôle de TFIIIC dans l'organisation de l'architecture nucléaire, est conservé de la levure aux mammifères.

Enfin, notre étude s'est poursuivie avec l'étude du facteur d'élongation TFIIS. L'isoforme TCEA1, exprimé de façon ubiquitaire chez la souris, a été étiquetée en N-terminal. L'analyse de la distribution de TCEA1 a révélé la conservation du recrutement de ce facteur aux gènes de classe III, chez les mammifères.

Cette étude s'est conclue par la parution d'un article dans le journal Nucleic Acid Research, sous le titre :

Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells.

Carrière L, Graziani S, Alibert O, Ghavi-Helm Y, Boussouar F, Humbertclaude H, Jounier S, Aude JC, Keime C, Murvai J, Foglio M, Gut M, Gut I, Lathrop M, Soutourina J, Gérard M, Werner M.

Nucleic Acids Res. 2011 Sep 12.

Article.

Nucleic Acids Research Advance Access published September 12, 2011

Nucleic Acids Research, 2011, 1–14
doi:10.1093/nar/gkr737

Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells

Lucie Carrière[1], Sébastien Graziani[1], Olivier Alibert[2], Yad Ghavi-Helm[1], Fayçal Boussouar[1], Hélène Humbertclaude[1], Sylvie Jounier[1], Jean-Christophe Aude[1], Céline Keime[3], Janos Murvai[1], Mario Foglio[4], Marta Gut[4], Ivo Gut[4], Mark Lathrop[4], Julie Soutourina[1], Matthieu Gérard[1,*] and Michel Werner[1,*]

[1]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), iBiTec-S, F-91191 Gif-sur-Yvette cedex, [2]CEA, iRCM, F-91057 Evry cedex, [3]Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS, INSERM, Université de Strasbourg, F-67404, Illkirch cedex and [4]CEA, iG, F-91057 Evry cedex, France

Received June 23, 2011; Revised August 22, 2011; Accepted August 24, 2011

ABSTRACT

RNA polymerase (Pol) III synthesizes the tRNAs, the 5S ribosomal RNA and a small number of untranslated RNAs. *In vitro*, it also transcribes short interspersed nuclear elements (SINEs). We investigated the distribution of Pol III and its associated transcription factors on the genome of mouse embryonic stem cells using a highly specific tandem ChIP-Seq method. Only a subset of the annotated class III genes was bound and thus transcribed. A few hundred SINEs were associated with the Pol III transcription machinery. We observed that Pol III and its transcription factors were present at 30 unannotated sites on the mouse genome, only one of which was conserved in human. An RNA was associated with >80% of these regions. More than 2200 regions bound by TFIIIC transcription factor were devoid of Pol III. These sites were associated with cohesins and often located close to CTCF-binding sites, suggesting that TFIIIC might cooperate with these factors to organize the chromatin. We also investigated the genome-wide distribution of the ubiquitous TFIIS variant, TCEA1. We found that, as in *Saccharomyces cerevisiae*, TFIIS is associated with class III genes and also with SINEs suggesting that TFIIS is a Pol III transcription factor in mammals.

INTRODUCTION

In eukaryotes, three nuclear RNA polymerases (Pol) are responsible for the transcription of the genome. Pol I transcribes a single RNA species, the precursor of the large ribosomal RNAs (rRNA) and of the 5.8S rRNA. Pol II transcribes all messenger RNAs and many non-coding RNAs implicated in various processes ranging from splicing, RNAs modification (snoRNAs) or gene regulation (miRNAs). Pol III transcribes the 5S rRNA, the tRNAs and a small number of stable non-coding transcripts (1,2). The U6 snRNA (mRNA splicing), RNAse P RNA (tRNA maturation) and 7SL RNA (signal recognition particle) are produced by Pol III in all eukaryotes examined so far, whereas, the RNase MRP RNA (mitochondrial rRNA maturation) is transcribed by Pol II in *Saccharomyces cerevisiae* and by Pol III in animals. Other short Pol III products of rather poorly defined functions, such as the vault particle, the Y, 4.5S and BC1 RNAs are largely specific to mammals (2). Finally, the short interspersed repeated elements (SINEs) are retrotransposons originating from class III (i.e. Pol III-transcribed) genes

*To whom correspondence should be addressed. Tel: +33 16908 9342; Fax: +33 16908 4712; Email: michel.werner@cea.fr
Correspondence may also be addressed to Matthieu Gérard. Tel: +33 16908 9429; Fax: +33 16908 4712; Email: matthieu.gerard@cea.fr
Present addresses:
Sébastien Graziani, DGA Maîtrise NRBC, Département Evaluation des effets des agents chimiques, ANC/TOGA, 3-5 rue Lavoisier, F-91710 Vert-le-Petit, France.
Yad Ghavi-Helm, Genome Biology Unit, European Molecular Biology Laboratory, D-69117 Heidelberg, Germany.
Fayçal Boussouar, Institut Albert Bonniot, Domaine de la Merci, F-38706 La Tronche cedex, France.
Marta Gut and Ivo Gut, Centre Nacional d'Anàlisis Genòmica, Parc Científic de Barcelona, Torre I, Baldiri Reixac 4, E-08028 Barcelona, Spain.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

present in hundreds of thousands of copies in mammalian genomes (3). SINEs can be transcribed by the Pol III transcription machinery *in vitro* (4).

The promoters of class III genes are divided into three categories depending on their organization and transcription factor dependence (reviewed in refs 5,6; unless stated otherwise, the nomenclature used for the class III machinery subunits is that of the mammalian transcription system according to ref. 5). Type II promoters of tRNA genes harbor intragenic A and B boxes that are recognized by TFIIIC, a six subunits factor (7 and references therein). Once associated with DNA, TFIIIC positions a second factor, called TFIIIB in yeast or TFIIIB-β in mammals (8), upstream of the transcription start site (TSS). Yeast TFIIIB and mammalian TFIIIB-β consist of the TATA box-binding protein (TBP), BDP1 and BRF1, which are related to TFIIB Pol II general transcription factor. Yeast TFIIIB and mammalian TFIIIB-β recruit Pol III through direct protein–protein interactions (9,10). The type I promoter, unique to the 5S rRNA, is located within the transcribed region and requires TFIIIA that acts as an adapter for the binding of TFIIIC. The mammalian U6 promoter, a classical type III promoter, is located upstream of the transcribed region and consists of a proximal sequence element (PSE) recognized by a four-subunit factor variously called the PSE-binding protein (PBP), the PSE transcription factor (PTF), or the snRNA activating protein complex (SNAPc) (11–13), and a TATA box which is bound by TFIIIB-α, in which the BRF1 subunit of TFIIIB-β is replaced by BRF2 (8,14,15).

Since Pol III transcribes several rRNAs and the tRNAs, it plays a central role in determining the translational capacity of the cell (16). The regulation of Pol III transcription plays a critical role in cell proliferation and cancer (17). Indeed, artificially increasing tRNA and 5S rRNA transcription causes increased cell proliferation and oncogenic transformation (18).

We discovered that TFIIS, a Pol II transcription elongation (19) and initiation factor in *S. cerevisiae* (20,21), also functions as a Pol III general transcription factor (22). Indeed, TFIIS binding could be detected on all class III genes and mutations that affected specifically Pol II or Pol III transcription were identified. Biochemical study of TFIIS role in Pol III transcription in yeast indicated that it stimulates faithful transcription initiation *in vitro*. In mouse, three isoforms of TFIIS, encoded by TCEA1, −2 and −3 exist (23–25). TCEA1 is expressed ubiquitously, contrary to TCEA2 and −3, which are expressed in spermatocytes or in the liver and kidney, respectively. Whether or not they are implicated in Pol III transcription is presently not known.

RNAs synthesized by Pol III often originate from repeated genes, be it the repeated 5S gene or multiple copies of some tRNA genes, raising the question whether all copies of a gene are transcribed at a given time. In addition, class III transcripts are difficult to predict bioinformatically. The nature of the class III transcriptome was first investigated in the yeast *S. cerevisiae* by analyzing the genome-wide distribution of the Pol III transcription machinery (26–28). The Pol III transcription machinery was associated with nearly all tRNA genes irrespective of their genomic localization, suggesting that they are transcribed. Only one new class III transcript, snR52 snoRNA, was identified (26–28). The ChIP-Seq method was very recently applied to the analysis of the Pol III transcription machinery in various human cell lines allowing a complete description of the class III transcriptome (29–33). Unlike the situation in *S. cerevisiae* where all tRNA genes are transcribed, only a subset of the tRNA genes is bound by the Pol III machinery in human. The identity of the bound tRNA genes varied from one cell type to another. In addition, these studies allowed the identification of a few dozen new loci bound by Pol III.

In *S. cerevisiae*, TFIIIC was present, independently of Pol III, on a small number of genomic locations, called *ETC* loci for extra TFIIIC (26). These *ETC* loci are conserved among yeast species suggesting that they might have a functional role in chromatin organization. Repressing artificially the expression of histones *in vivo* leads to decreased nucleosome abundance. In such a situation, the expression of the *ETC* loci is induced (34). In *Schizosaccharomyces pombe*, the presence of TFIIIC independently of Pol III and TFIIIB marks several boundaries between euchromatin and heterochromatin domains (35). In *S. pombe*, TFIIIC plays an active role in delimiting the boundaries since *cis*-acting mutations that abolish its binding lead to the spreading of heterochromatic marks in regions that are usually euchromatic and as a consequence, lead to transcriptional silencing. Interestingly, a large number of ETCs was also found in human cell lines (1865 in K562 cells and 307 in HeLa cells; (30,31). Moreover, the ETCs that were highly enriched for TFIIIC were also often associated with CCCTC-binding factor (CTCF) (31), a protein implicated in insulation and chromosome looping and conformation (36), suggesting a role for TFIIIC in defining repressive domains in human.

In this study, we investigated the distribution of Pol III, BRF1, BRF2, TFIIIC and one of the TFIIS homologs, TCEA1, on the genome of mouse embryonic stem (ES) cells using a highly specific ChIP-seq method that entailed tagged ES cell lines. Our work provided a detailed analysis of the active class III genes in mouse ES cells and led to the discovery of new genes transcribed by this enzyme. We found that only a few hundreds SINEs are transcribed by Pol III. Interestingly, the presence of TCEA1 was detected on the majority of the active class III genes suggesting that it acts as a Pol III transcription factor in mammals. We also took advantage of the genome-wide analysis of TCEA1 to show that it is present at similar levels on paused Pol II peaks of active and inactive genes suggesting that it is not TFIIS recruitment that triggers the passage of Pol II into elongation.

MATERIALS AND METHODS

Construction of the mouse ES cell lines

We used the recombineering technology (37) to generate the targeting vectors that introduce the triple affinity tag in the subunits of Pol III, TFIIIB-α or -β, TFIIIC and TFIIS. The 46C ES cell line (38) was transfected by

electroporation with each targeting vector. Cells were plated in D15 medium as described (39) and selected with G418. Individual ES cell colonies were collected 7 days after electroporation, amplified and genotyped by Southern blotting, in order to identify the clones that underwent a homologous recombination event. In these clones, a sequence encoding a 6 Histidine-Flag-HA tag, followed by a neomycin resistance marker flanked by loxP sites, was inserted just after the last codon of the gene encoding the protein to be tagged. For TCEA1, the tag was inserted just after the start codon. The sequence of the insertion cassette is given in the Supplementary Table S2. The integration of the cassette at the right loci was verified by Southern blotting using three or four different restriction enzymes and DNA-polymerase chain reaction (PCR). Transient transfection with a Cre recombinase expression vector was used to remove the selection cassette. The karyotypes of all cell lines were verified. The genes that were modified in the cell lines encoded RPC1 (MGI:2681836), RPC4 (MGI:1914315), BRF1 (MGI:1919558), BRF2 (MGI:1913903), TFIIIC220 (MGI:107887), TFIIIC110 (MGI:1919002), TFIIIC90 (MGI:2138937) and TCEA1 (MGI:1196624), respectively. The expression of the tagged versions of the proteins was verified by western blotting using HA7 anti-HA antibodies (Sigma; Supplementary Figure S2).

ES cell lines culture

ES cells were cultured on embryonic fibroblast feeder cells blocked with mitomycin C in D15 medium [Dulbecco's Modified Eagle's Medium High Glucose supplemented with 15% fetal bovine serum, 2 mM L-glutamine, 50 U/ml penicillin, 50 µg/ml streptomycin, 0.1 mM non-essential amino acids (all from GIBCO), 0.1 mM 2-mercaptoethanol (Sigma) and 1000 U/ml LIF]. ES cells were maintained at 37°C, 5% carbon dioxide, fed with fresh media daily, and transferred to new plates after trypsinization.

Chromatin immunoprecipitation

Typically $(150–200) \times 10^6$ cells were collected for each ChIP experiment. The cellular proteins and DNA were cross-linked by the addition of formaldehyde (0.4% final concentration). The plates were incubated for 10 min at room temperature and then the reaction was stopped by the addition of glycine (0.125 M final concentration). The cells were washed twice with 10 ml chilled phosphate buffered saline (PBS). Each plate was scraped with 2 ml PBS with protease inhibitors (Roche complete, 10 mM PMSF dissolved in ethanol), pelleted by centrifugation at 2500 rpm and resuspended in 1 ml per dish of FA/SDS buffer (50 mM Tris–HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS) with protease inhibitors and incubated on ice for 15 min. The subsequent operations were performed at 4°C. Chromatin was collected by centrifugation for 20 min at 12 000 rpm, resuspended in the same amount of FA/SDS and put on a rotating wheel for 1 h, centrifuged and resuspended in 0.4 volume of FA/SDS with protease inhibitors. The 400 µl batches of chromatin were fragmented with a sonicator (Diagenode) to generate DNA fragments of 300–400 bp mean size. The chromatin was collected by centrifugation at 14 000 rpm for 10 min and was kept at −80°C until needed.

The immunoglobulin-coupled magnetic beads (Dynal) used in the chromatin immunoprecipitation experiments were prepared essentially as described previously (40). Anti-HA antibody was HA7 H-3663 (Sigma) and anti-Flag anti-body was F-1804 (Sigma). For each ChIP experiment, 240 µg DNA (RPC1, RPC4, 46C) or 750 µg DNA (BRF1, BRF2) of sonicated chromatin were incubated with 150 µl beads for 2 h at room temperature. The beads were magnetically pelleted and the supernatant was removed. The beads were washed with 1 ml of low salt buffer (pH 8.0, 20 mM Tris–HCl, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100, 0.1% SDS), then with the same volume of high salt buffer (pH 8.0, 20 mM Tris–HCl, 2 mM EDTA, 500 mM NaCl, 1% Triton X-100, 0.1% SDS), then with low LiCl buffer (pH 8.0, 10 mM Tris–HCl, 1 mM EDTA, 250 mM LiCl, 1% NP-40, 1% deoxycholate) and finally twice with TE buffer (pH 8.0, 10 mM Tris–HCl, 1 mM EDTA). The first ChIP was performed using the anti-HA antibody. The bound chromatin was eluted using FA/SDS buffer containing the HA peptide (0.5 mg/ml; Ansynth Service) for 4 h at 16°C then overnight at 4°C. For the second ChIP, the supernatant was incubated with 50 µl beads coupled to anti-Flag antibodies which were treated as above except for the elution that was performed by incubation with a buffer containing 1% SDS and 0.1 M $NaHCO_3$. Reversal of the cross-links and DNA purification were performed as described previously (41). TFIIIC subunits and TCEA1 ChIPs were fragmented using MNase I. The protocol will be described in detail elsewhere (M. Gérard, manuscript in preparation).

The immunoprecipitated DNA was analyzed by quantitative real time PCR on an ABI Prism 7000 or 7300 machine (Applied Biosystem; 40). Relative quantification using a standard curve method was performed and the occupancy level for a specific fragment was defined as the ratio of immunoprecipitated DNA over total DNA.

Sequencing of the immunoprecipitated DNA and analysis of the regions bound by the RNA pol III

For each chromatin immunoprecipitation, the DNA was sequenced on a single Solexa genome analyzer GS or GA IIx channel using the procedures recommended by the manufacturer (Illumina). The characteristics of each sequencing experiment are indicated in Supplementary Table S3. The bound regions were identified using Quest version 2.3 (42) and an in-house program that used Quest method for peak calling. The data have been deposited to the ArrayExpress database under accession number E-MTAB-767.

RNA preparation

Total RNAs were isolated from 46C wild-type ES cells, using 1 ml of Trizol (Invitrogen) per 10 cm-dish, as indicated by the manufacturer. Total RNAs were prepared following the manufacturer's protocol, except that the

RNA pellet were washed with 80% ethanol, without shaking. The RNAs were suspended in diethylpyrocarbonate treated water, at around 2 mg/ml and stored at −80°C until needed. Integrity of RNA was tested on 1.5% agarose gel. Total RNA concentration and purity were verified using a NanoDrop spectrophotometer ND-1000 (Thermo Scientific), measuring absorbance at OD 260/280. Total RNA samples were treated with 1 U of RQ1 DNase (Promega) per µg of RNA, for 1 h at 37°C. DNase was inactivated by the addition of 20 mM ethylene glycol tetraacetic acid, pH 8.0 (Stop solution) and heated at 65°C for 10 min. The RNA was then precipitated with isopropanol.

cDNA synthesis

cDNA synthesis was performed using SuperScript II reverse transcriptase with random hexamer primers (Invitrogen) according to the manufacturer's instructions. In brief, 5 µg of total RNA, 2 µl of random hexamer primers (10 µM), 1 µl of dNTP mix (5 mM each) to 11 µl in total, were incubated at 65°C for 5 min. After chilling on ice for 2–3 min and brief centrifugation, 5 µl of first-strand synthesis buffer (5×, containing 250 mM Tris–HCl [pH 8.3], 375 mM KCl, 15 mM MgCl2), 2 µl of 0.1 M DTT, were added and the tubes were incubated at 25°C for 1 min. Then, 1 µl of (200 U/µl) of SuperScript II reverse transcriptase was added and the reaction was first incubated at 25°C for 10 min, followed by incubation at 42°C for 1 h. Reverse transcriptase activity was terminated by incubation at 70°C for 15 min. 1 µl RNase H (Invitrogen, 2U/µl) was added and further incubated at 37°C for 20 min. Samples were stored at −20°C until needed. The cDNA solution was diluted 10-fold before use in reverse transcriptase-PCR (RT–PCR). PCR were performed using specific primers covering the region of interest. The reactions contained 25 ng of cDNA, and primers at a final concentration of 150 nM. RT-PCR reactions were analyzed by gel electrophoresis. Further, total RNA samples were analyzed for the possible presence of DNA contamination by PCR using RNA not reverse-transcribed.

RESULTS

Setting up a high-specificity tandem affinity chromatin immunoprecipitation method in mouse ES cells

To identify the regions bound by the Pol III transcription machinery, we developed a tandem affinity chromatin immunoprecipitation method in mouse ES cells. Two different specific subunits of Pol III, RPC1 and RPC4, the BRF1 and BRF2 subunits of TFIIIB-β or TFIIIB-α, respectively and the TFIIIC220, TFIIIC110 and TFIIIC90 subunits of TFIIIC were tagged (Supplementary Table S1). To investigate whether or not TFIIS is a Pol III transcription factor, TCEA1 was also tagged. We introduced a cassette encoding consecutively six histidines, one Flag and one HA epitope just after the last sense codon of the Pol III, BRF and TFIIIC genes. A neomycin marker flanked by loxP sites follows the tag cassette. The construction was introduced in 46C mouse ES cell genome using the recombineering method (Supplementary Figure S1 and Supplementary Table S2) (37). The neomycin marker was removed by expressing the cre recombinase. The correct integration of the cassette at the endogenous locus and the excision of the neomycin marker were verified by Southern blotting and PCR. The expression and correct size of the C-terminally tagged proteins were verified by western blotting with a monoclonal anti-HA antibody (Supplementary Figure S2).

Chromatin from the RPC1-, RPC4-, BRF1- or BRF2-tagged ES cell lines was prepared. ChIP experiments in RPC1 or RPC4 ES lines with anti-HA antibodies enriched strongly a tRNA-val gene and the H1 gene (from 40- to 110-fold) but not the ARBP gene, which is transcribed by Pol II (Figure 1A). We then tested whether genes that depend on BRF1 or BRF2 could be distinguished. BRF1 ChIP enriched specifically two tRNA genes. Conversely, BRF2, but not BRF1, was associated with H1 gene, encoding the RNA subunit of the RNase P, and U6 (Figure 1B) (5,6).

To improve the specificity of the ChIP experiments, tandem immunoprecipitations were performed. The chromatin was first ChIPed with the anti-HA antibody as above, and then eluted by competition with an HA peptide. The enriched chromatin was submitted to a second round of ChIP with a monoclonal anti-Flag antibody. The second ChIP further improved the enrichment ratio above background up to 8-fold depending on the protein and gene considered (Figure 1C and Supplementary Figure S3). The protocol used here is thus, highly specific for ChIP experiments of tagged proteins expressed from their native locus in mouse ES cells. It is also generic since it does not depend on the generation of high specificity antibodies.

Analysis of the regions bound by RNA polymerase III

DNA from the chromatin associated with Pol III was sequenced using a Solexa Genome Analyzer. The comparison of ChIP-Seq experiments using tagged RPC1 and RPC4 ES cell lines allowed us to cross-validate the Pol III-binding sites. We considered as bound regions, only those that showed co-occupancy by RPC1 and RPC4. These regions were validated if the number of tags mapping within the defined interval in the ChIP-Seq experiments with RPC1 and RPC4 were both 5-fold higher than in the untagged ES cell line, which was used as a negative control. Regions that satisfied only one of the two conditions were visually inspected using the UCSC genome browser and rejected if the 46C track showed a high background around the bound region. The protein-bound regions were annotated using the mm9 UCSC mouse database. The regions that were associated with more than one annotation were visually inspected and eventually split to associate one region with each annotation. If one annotation was associated with two or more regions, these were inspected and eventually fused. Figure 2 and Supplementary Figure S4 show examples of tag density profiles on various regions representative of several class III genes. Data concerning the bound regions can be found in Supplementary Table S4.

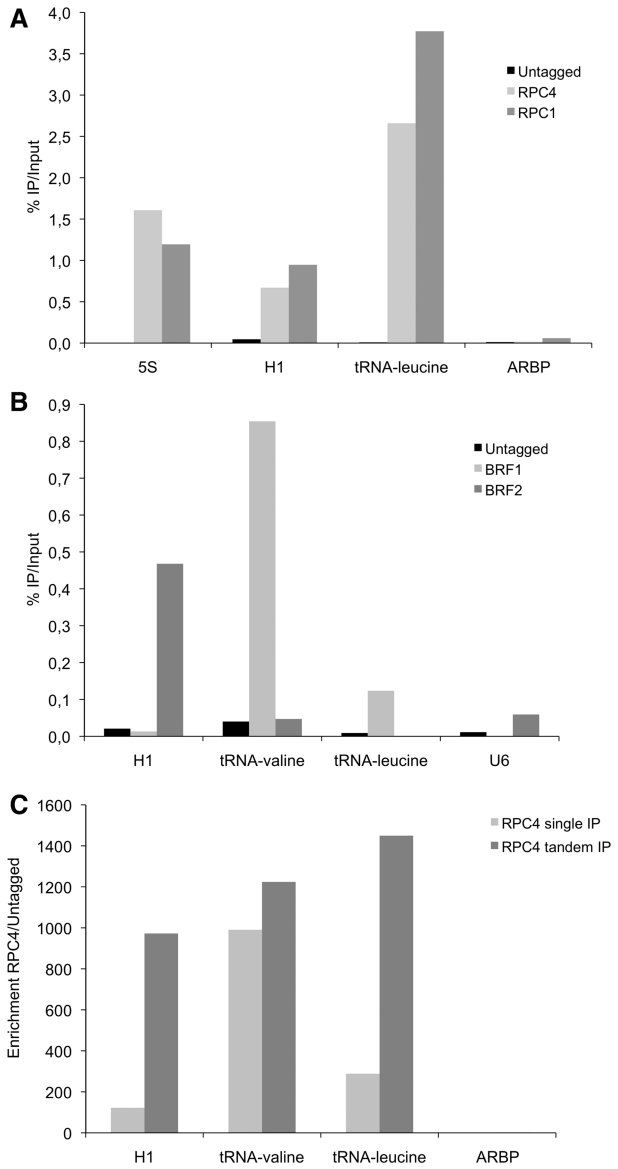


Figure 1. Chromatin immunoprecipitation of the Pol III transcription machinery in mouse ES cell lines. ChIP experiments were performed with the HA7 H-3663 anti-HA antibody (**A** and **B**) as described in 'Materials and Methods' section. (A) Chromatin extracts from RPC1 or RPC4 cell lines were used to show Pol III enrichment on different types of class III genes but not on the ARBP class II gene. (B) Extracts from BRF1 or BRF2 cell lines were used to demonstrate specific enrichment of TFIIIB-β or TFIIIB-α on genes with type 2 (tRNAs) or type 3 promoters (H1, U6), respectively. An untagged cell line was used as a negative control. (**C**) Tandem ChIPs improved signal to noise ratios. Single round ChIP (RPC4 single IP) were performed as above with HA7 H-3663 anti-HA antibody and were compared with experiments in which a second round (RPC4 tandem IP) of immunoprecipitation was done with F-1804 anti-Flag antibody after elution of proteins with an HA peptide. Enrichment-folds of sequences from an RPC4 ES cell line chromatin extract relative to an untagged cell line chromatin are indicated.

tRNAs. A total of 284 tRNA genes were bound by Pol III, including one selenocysteine tRNA gene and one possible suppressor. Of these, 271 were predicted by Coughlin *et al.* (43) and 281 by the Genomic tRNA Database (GtRNADb). Coughlin *et al.* predict that 461 tRNA genes exist based on the sequence of the expressed tRNAs. GtRNADb, which relies upon the tRNAscan-SE program (44), predicts the existence of 433 tRNA genes (including two selenocysteine tRNAs and one possible suppressor) in the mouse genome. Altogether, 526 tRNAs are annotated on the mouse genome by these two databases. Totally, 59 and 65% of the genes predicted by Coughlin *et al.* and GtRNADb, respectively, were bound by Pol III in mouse ES cells. This observation raised the possibility that some tRNAs were not detected because they cannot be identified by unique tags. The percentage of nucleotides that could not be mapped (NM score) within 150 nt on either side or within the tRNA genes that were actually bound by Pol III was computed. For each tRNA, we considered the lowest NM score from the three regions. We found that the worst NM score for a bound tRNA was 75%. We figured out that among the 242 tRNA genes predicted by Coughlin *et al.* or GtRNAdb that were not bound by Pol III, only 14 had an NM score >75%. Hence, we did not underestimate the number of bound tRNA genes by >13.2%. We performed independent ChIP experiments for three bound tRNA genes and two mappable unbound tRNA genes, which agreed with the ChIP-Seq experiments (Supplementary Figure S5). The genomic distribution of bound tRNA genes and other class III genes is shown in Supplementary Figure S6.

snRNAs. In addition to tRNA genes, we also found binding of Pol III transcription machinery on the U6, 5S, 7SK, 7SL, 4.5S, BC1, HY1 and HY3 genes. However, the number of bound genes was very small compared to that predicted in Repbase (Table 1). For example, Repbase predicts 1269 U6 genes and Ensembl, 617 when we actually found only 5 that were bound. As for the tRNA genes, we performed an independent ChIP experiment and verified for two genes of each 7SL, 7SK, U6 and HY3, one bound in the ChIP-Seq experiment and one mappable but unbound, that Pol III was present or not as expected (Supplementary Figure S5).

New Pol III transcripts. About 30 regions were bound by both RPC1 and RPC4, but were not annotated as Pol III-transcribed genes or SINEs and could thus correspond to new Pol III transcripts. Of note, these regions were neither annotated as transcribed by Pol II, nor corresponded to miRNAs. We verified by an independent ChIP experiment for nine of those regions that they were indeed bound by RPC1 and RPC4 (Supplementary Figure S7A). Furthermore, we looked for the presence of an RNA associated with the Pol III-bound region by RT–PCR (Supplementary Figure S7B). An RNA was transcribed from 16 regions out of the 18 that were examined. Most of these regions were conserved in the rat genome but only one was highly conserved in human (Supplementary Figure S7C and Supplementary Table S5). The latter region, which was also associated with Pol III in human, had a typical organization for class III genes with conserved A-box and B-box and terminator. Altogether, these results strongly suggest that we have identified new class III transcripts.

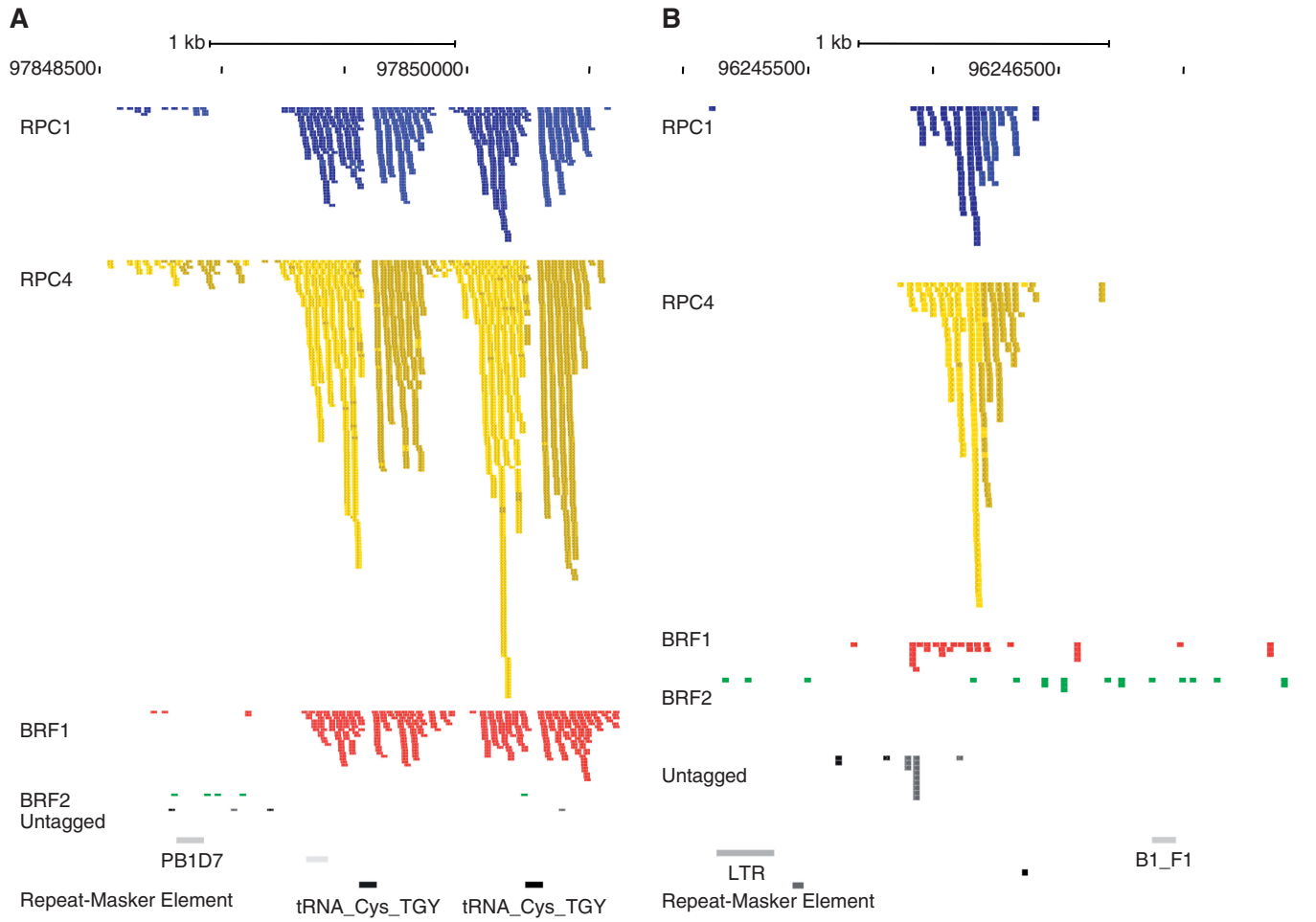


Figure 2. Representative examples of sequence reads that map on bound regions. (**A**) Binding of Pol III and BRF1 to two tRNA genes and a SINE. The sequences that map to chromosome 11 are displayed using the UCSC genome browser. A 1 kb scale is shown at the top of the Figure. The sequence tags are represented by colored rectangles. Tags identified in the ChIP-Seq experiments performed with the RPC1-tagged, RPC4-tagged, BRF1-tagged or BRF2-tagged ES cell lines and 46C negative control cell line are colored in blue, yellow, red, green and gray, respectively. Light colors indicate the sequences that match the top strand. Dark colors indicate the tags that match the bottom strand. The location of repeated sequences, according to RepBase, is indicated at the bottom of the Figure by gray or black boxes. The two black boxes refer to two tRNA-Cys-TGY genes. The bound SINE is a PB1D7 repeat of the Alu family. (**B**) Binding of Pol III transcription machinery to an unannotated region on chromosome 3.

Distribution of BRF1 and BRF2 on class III genes

We also tagged the BRF1 or BRF2 subunits of TFIIIB-β or TFIIIB-α, respectively, to investigate which genes depend on either form of the Pol III factor. Type I and II promoters require the presence of BRF1 in TFIIIB-β transcription factor for their transcription while type III uses BRF2 as a subunit of TFIIIB-α. As expected, BRF1 was exclusively associated with types I and II promoters (Figures 2A and 3A–C), whereas, BRF2 was bound to type III promoters only (Figure 3D and Supplementary Figure S4), indicating that no class III gene could use both TFIIIB variant. BRF2 was associated with a small number of snRNAs that included U6, 7SK, the H1 and MRP RNA genes, U6atac, the $tRNA^{sec}$ and two of the Y RNA genes, HY1 and HY3. Apart from these genes that were already predicted to depend on BRF2, no new region was found. The new genes were associated with low, or sometimes background levels of BRF1 (Figure 3E). This situation probably stems from the fact that the number of tags associated with the new genes is often low even for the Pol III ChIP. The number of tags associated with class III genes in experiments with the BRF1 cell line is around 10-fold smaller than with the RPC4 cell line (Supplementary Table S4, compare RPC4.SUM_DENSITY to BRF1.SUM_DENSITY). A thorough assessment of BRF1 and/or BRF2 presence on the new genes would thus require a 10-fold deeper sequencing.

We wondered where the two forms of TFIIIB were positioned, relative to the mature RNA 5′-end (the TSS of mouse class III genes have not been systematically determined). On tRNA genes, BRF1 peak density was located 10 nt upstream of RNA 5′-ends (Figure 3A), which is within a few nucleotides of the position determined by *in vitro* footprinting experiments. A similar situation was found for BRF2 on genes with type III promoters with a peak of tag density located 14 nt

upstream RNA 5′-ends (Figure 3D). Using data from Kagey *et al.* (45), we also looked at the distribution of TBP relative to the class III genes. TBP was positioned 17 or 12 nt upstream of RNA 5′-ends for class II or class III promoters, respectively.

Table 1. Pol III-associated regions

RNA	Number bound	Number predicted[a]	Percent bound
tRNA	284	526	53.8
4.5S	13	1475	0.81
5S	3	984	0.30
7SK	1	665	0.15
7SL	2	276	0.72
BC1	2	6351	0.03
HY1	1	37	2.70
HY3	1	15	6.67
HY4	0	2	0.00
HY5	0	1	0.00
U6	5	1269	0.39

[a]The number of predicted class III transcripts are extracted from GtRNAdb and Coughlin *et al.* (43) for the tRNAs and from Repbase for the other ones.

Transcription of SINEs

Pol III *in vitro* transcription systems can drive the transcription of SINEs (4). We wondered whether some of the SINEs could be bound by Pol III, and thus most probably be transcribed *in vivo*. Indeed, we found 241 locations on the mouse genome that were associated with both RPC1 and RPC4 and had a SINE annotation according to Repeatmasker. We first looked at the distribution of the sequence divergence of the SINEs (Supplementary Figure S8A). The distribution is bi-modal with maxima at 7 and 27% divergence and a minimum at 13%. The SINEs were separated in two classes according to their divergence using 13% as threshold. Notably, most (80%) of the bound SINEs belonged to the highly conserved category (Supplementary Figure S8B).

We wondered if the small number of observed bound SINEs in our ChIP-Seq experiments could stem from the low mappability of the regions encompassing the SINEs. To answer this question, we used an approach that was similar to the one applied to the tRNA genes. The bound SINEs were divided in two populations depending on whether their divergence was ≥13%. We determined the lowest NM score for bound SINEs of each category. The

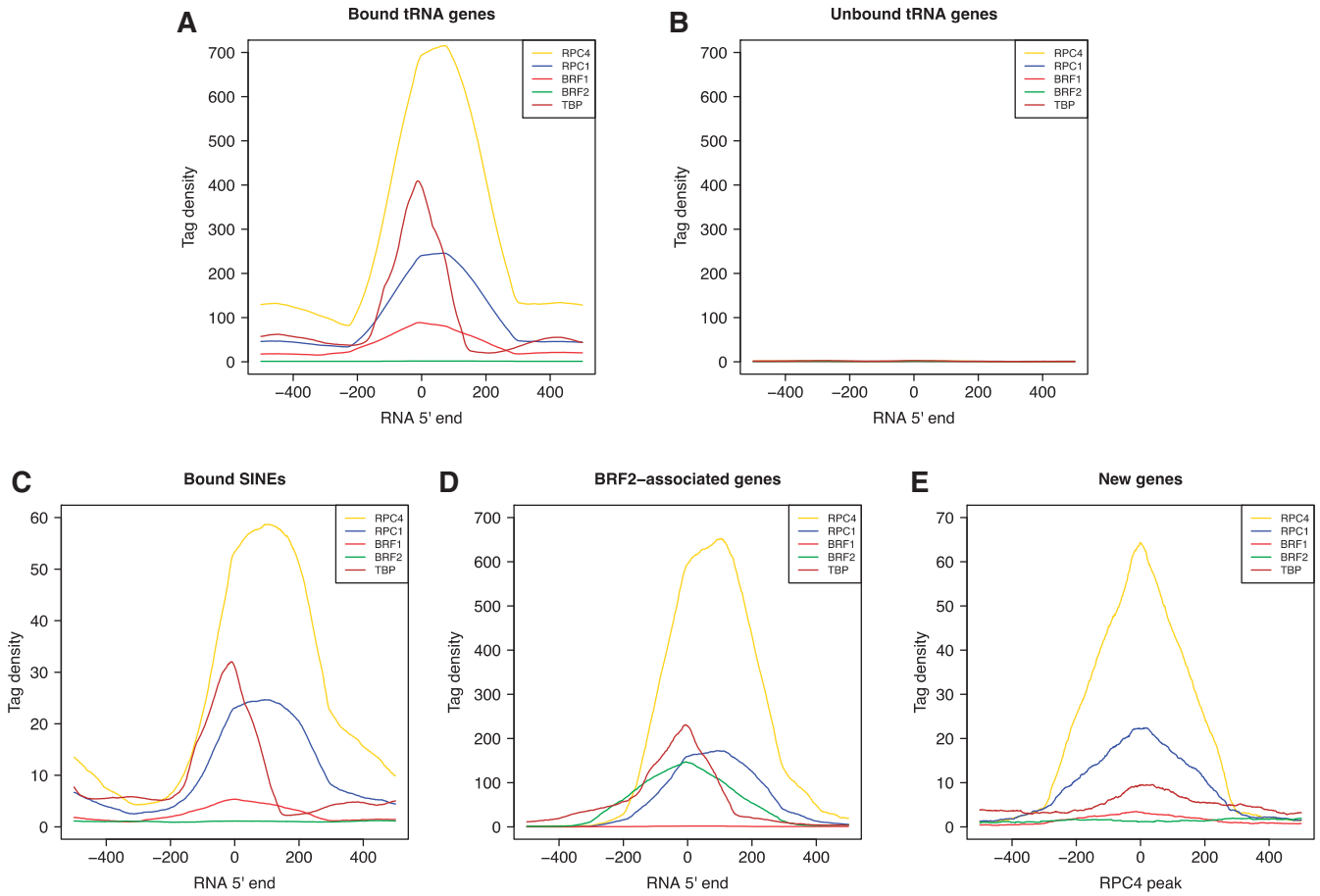


Figure 3. Distribution of Pol III, TBP, BRF1 and BRF2 relative to class III genes. Tag densities are shown for RPC1, RPC4, BRF1, BRF2 and TBP (45) relative to the location of the RNA 5′-end for (**A**) tRNA genes, (**B**) unbound tRNA genes, (**C**) SINEs, (**D**) BRF2-associated genes and (**E**) new class III genes. For these, neither the transcription orientation, nor the TSS or end is known. Hence, the distribution is shown relative to the RPC4 density peak.

worst NM score within the SINE or the 150 nt flanking-regions was 65% for the conserved category and 50% for the diverged category. We then computed the NM score for all the SINEs predicted by Repeatmasker for each of the two divergence classes and determined the number of SINEs present in the mouse genome that had their two flanking regions with NM scores above those observed in bound SINEs. We postulated that the proportion of bound SINEs in the non-mappable population was similar to that in the mappable population. We could thus estimate that only around 10 non-mappable SINEs might be bound.

We searched in the bound SINEs for the presence of the TRGYTYARTGG and GTTCRAWTC sequences, which correspond to the A- and B-boxes in human (30). Of them, 64% contained an A-box and 87% a B-box (at *P*-value 10^{-3}), indicating that most of them had highly conserved type II promoters. Pol III-associated SINEs were found throughout the mouse genome and were not particularly associated with other categories of class III genes (Supplementary Figure S6B). Altogether, these observations indicate that a limited set of the SINEs direct Pol III transcription *in vivo*.

Chromatin environment of class III genes

We looked for the presence of various chromatin marks around the bound and unbound class III genes, comparing the distribution of H3K4me1, -me2, -me3, which are considered as active chromatin marks, and H3K27me3, that is an inactive chromatin mark (46–48). We also analyzed the distribution of H3K9me3, which are deposited at many euchromatic loci in mouse ES cells (49). Interestingly, both Pol III bound tRNA genes and BRF2-dependent genes were surrounded by peaks of H3K4me3 located around 400 nt on either side of the RNA 5′-ends (Supplementary Figure S9). H3K4me3 mark was absent from unbound tRNA genes (data not shown). The euchromatic mark H3K9me3 might be slightly enriched on either side of active class III genes. On the contrary, H3K27me3, which is typically heterochromatic was completely absent. This pattern is extremely similar to that of Pol II-transcribed genes.

Distribution of TFIIIC

TFIIIC is required for the transcription of class III genes that are under the control of types I and II promoters, i.e. 5S and tRNA genes. However, in *S. pombe*, *S. cerevisiae* and human, TFIIIC binds some regions independently of Pol III (26,30,31,35). We explored the genome-wide distribution of TFIIIC in mouse ES cell lines where one subunit of TFIIIC, TFIIIC220, −110 or −90, was tagged. Since the standard ChIP protocol that used sonication for DNA shearing did not give satisfactory results, it was modified by the inclusion of a MNase I DNA digestion step (M.G., manuscript in preparation). Tandem ChIP experiments allowed the identification of 2652 regions that were considered as bound since they were consistently enriched in ChIP experiments with the three TFIIIC-tagged cell lines but not in an untagged control cell line (Supplementary Table S6).

Of the 283 tRNA genes bound by Pol III (excluding the selenocysteine tRNA which has a type III promoter), 261 were also associated with significant levels of all three TFIIIC subunits. The three subunits displayed similar distributions downstream of BRF1 on the body of the tRNA genes (Figure 4A). The distribution of the tags on tRNA genes was very similar for the three subunits tested, showing an extended association with tRNA genes. TFIIIC was absent from the tRNA genes that were not bound by Pol III or BRF1 (data not shown). In line with their promoter organization, Pol III-bound SINEs were also associated with TFIIIC and BRF1, but not BRF2 (Figure 4B). As expected from previous *in vitro* experiments (5), the BRF2-associated genes were completely devoid of TFIIIC (Figure 4C). Interestingly, the three TFIIIC subunits were generally associated with the new class III genes while no BRF2 could be found suggesting that they depend on TFIIIC for their transcription (Figure 4D).

Several recent studies pointed at the presence in human cell lines of numerous ETC loci where TFIIIC is present but devoid of Pol III or TFIIIB-α or -β (26,30,31,35). Similarly, 2233 ETC loci bound by the three TFIIIC subunits were detected by our ChIP-Seq experiments. TFIIIC has been shown to play a role in the organization of chromatin in *S. cerevisiae* and *S. pombe* and to act as a barrier in the extension of heterochromatin into euchromatic territories (35,50,51). It has been shown that CTCF protein plays a similar insulator role in mammals (36). Moreover, CTCF interacts with cohesins to position it at many sites (52). A correlation between CTCF and ETC sites in human cells has been found (30,31). To investigate the potential correlation between the ETCs- and CTCF-binding sites in mouse ES cells, the ETCs were ordered according to their distance relative to the CTCF-binding sites (data taken from ref. 53). In parallel, an identical number of randomly selected regions were ordered using the CTCF-binding site distance criterion. Heat maps of the tag density for TFIIIC220 and CTCF were generated (Figure 5A) and compared with the heat maps of the randomly selected regions (Figure 5B). In agreement with the observations made in human cells, we found that 85% of ETCs also had CTCF-binding sites within 20 kb (Figure 5A). The distribution of CTCF was clearly different when regions were selected randomly (Figure 5B; compare the sigmoid curves in the two panels). The CTCF-binding sites were previously shown to be associated with the Smc1A and Smc3 subunits of cohesin (45). Heat maps were drawn for Smc1 and Smc3 and confirmed the clear co-occupancy of these proteins with CTCF. Intriguingly, in addition to being closely associated with CTCF, we observed that Smc1A and, to a lesser extent, Smc3 were enriched at the ETCs themselves, as indicated by the increased tag density centered on the ETCs (Figure 5A) which is absent from the maps of the randomly selected regions (Figure 5B).

Enrichment of TCEA1 isoform of TFIIS at class II and class III genes

TFIIS is a transcription elongation factor that stimulates Pol II transcription elongation and initiation. Moreover,

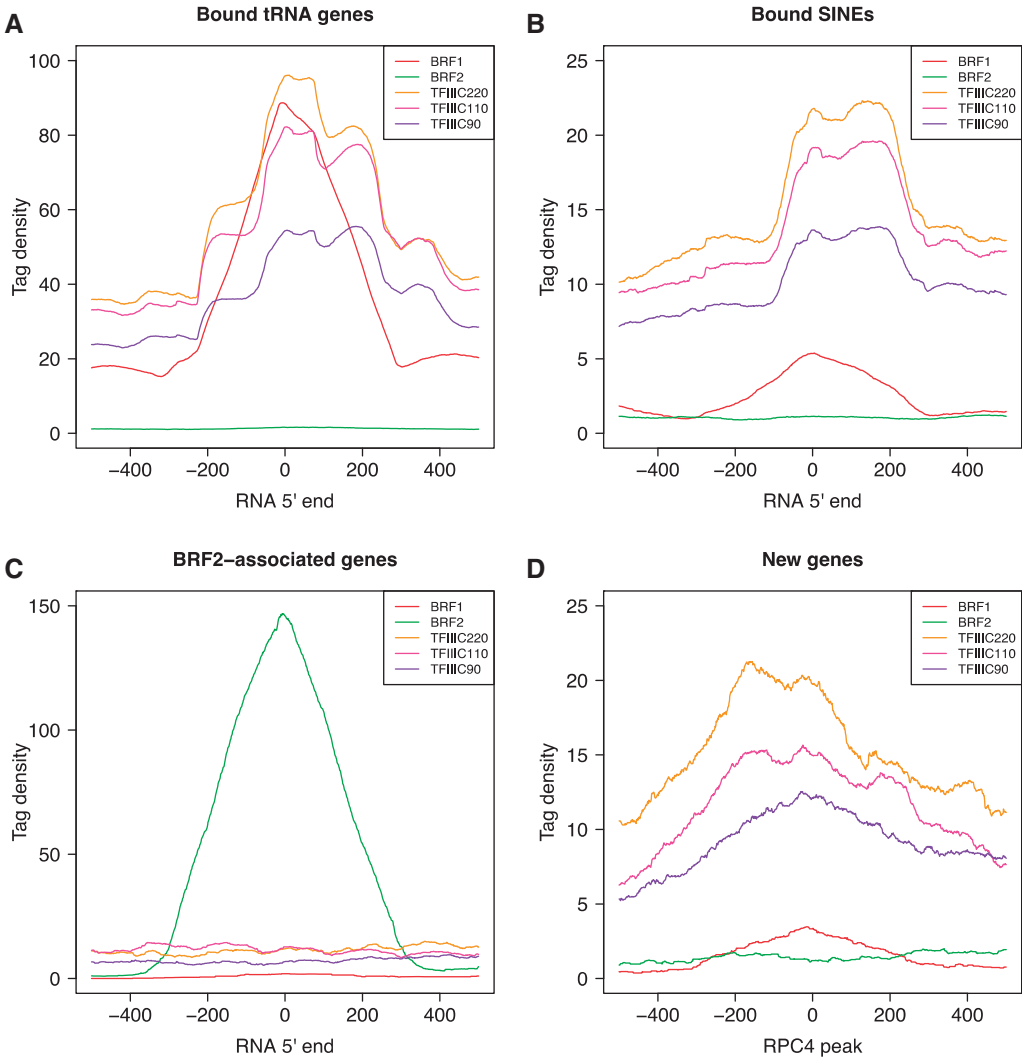


Figure 4. Distribution of TFIIIC relative to class III genes. Tag densities are shown for TFIIIC220, −110, −90, BRF1 and BRF2 relative to the location of the RNA 5′-end for (**A**) tRNA, (**B**) SINEs, (**C**) BRF2-associated. For the new class III genes (**D**) the distribution is shown relative to the RPC4 density peak.

we have shown that it acts positively on Pol III transcription in *S. cerevisiae* (22). The distribution of TFIIS on the genome of mouse ES cells was thus investigated. Three variants of TFIIS are encoded by the mouse genome. The ubiquitously expressed variant TCEA1 was tagged at its N-terminus because C-terminal tagging abolishes TFIIS function in yeast. Chromatin immunoprecipitation was performed using the same protocol that was used for TFIIIC.

We found that 57.6% of the tRNA genes and 50% of the BRF2-dependent genes that were associated with Pol III were also bound by TCEA1 (Figure 6 and Supplementary Figure S10) using a 3-fold signal to background threshold. About 41.1% of the bound SINEs and 16.7% of the new class-III genes were also bound by TCEA1 using the same criterion. We wondered if the presence of TCEA1 on class III genes does require Pol II (data from ref. 54). On tRNA genes and active SINEs, only very low levels of hypophosphorylated Pol II is found upstream of the genes (Figure 6A and B). Low levels of hypophosphorylated Pol II was also present downstream of the SINEs. Pol II phosphorylated on serine 2, 5 or 7 of the CTD was not significantly enriched on the regions surrounding active tRNA genes or SINEs. The distribution of TCEA1 on tRNAs and SINEs resembled that of Pol III but was shifted upstream toward the 5′-end in line with its possible role in transcription initiation by Pol III (22). Intriguingly, the hypophosphorylated, S7P and S5P forms of Pol II were all significantly present upstream of BRF2-associated genes (Figure 6C). The level of unphosphorylated Pol II was around 5-fold more abundant on BRF2-associated genes, than on active tRNA genes. However, two peaks of TCEA1 were found on BRF2-dependent genes, one associated with Pol II, the other with Pol III. These observations suggest that TFIIS plays a role in Pol III transcription.

The distribution of TCEA1 on class II genes was also analyzed. When ordered according to RNA steady-state levels transcribed from a given gene (measured by RNA-seq; C. Keime and M. Gérard, manuscript in

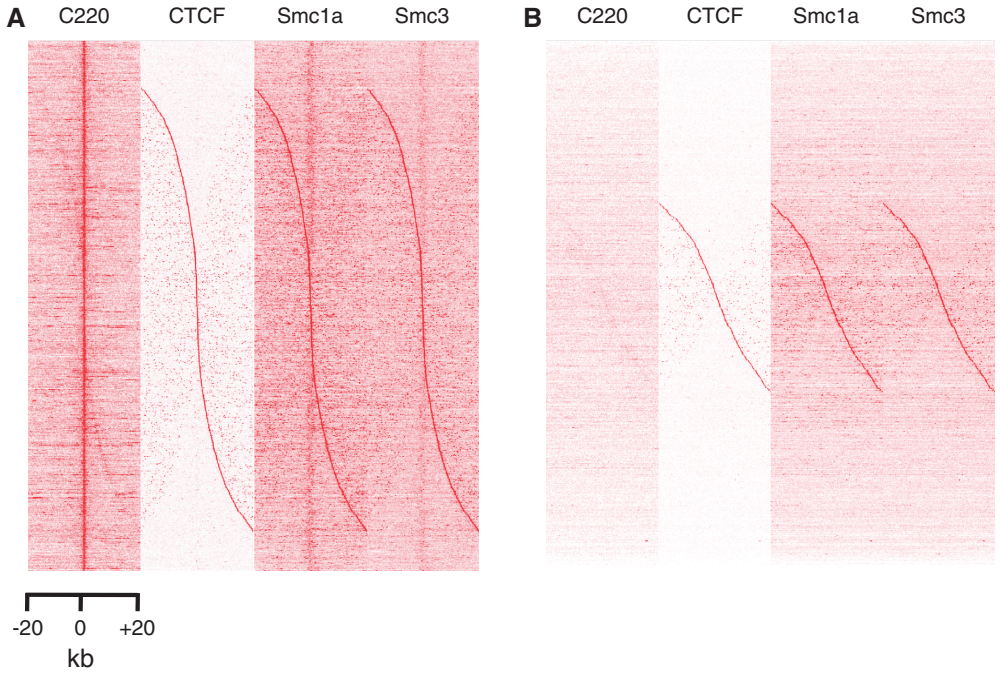


Figure 5. Distribution of CTCF, Smc1a, Smc3 relative to ETC-bound TFIIIC220. (**A**) Distribution of CTFC, Smc1a and Smc3, relative to TFIIIC220. The ETC regions bound by TFIIIC220 (C220) were ordered relative to their distance to CTFC-binding sites using seqMINER (63). The windows span 20 kb upstream and downstream the center of TFIIIC bound regions. The distribution of Smc1a and Smc3 on the same regions is shown on the two right panels according to the same order. (**B**) Distribution of CTCF, Smc1a and Smc3 relative to randomly selected regions. Randomly selected regions were sorted according to the distance of CTCF-binding site to the center of the region. Smc1a- and Smc3-binding sites are shown according to the same in the two right panels.

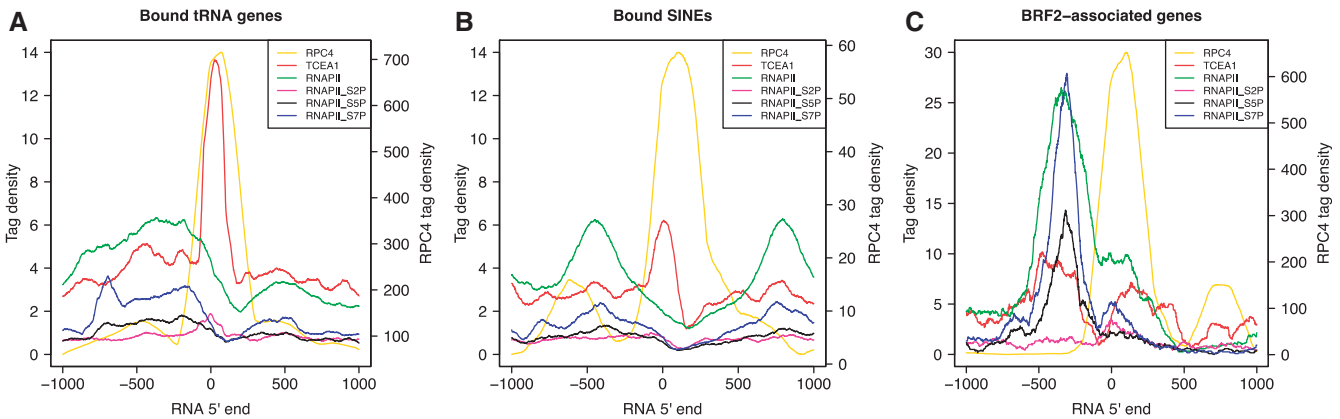


Figure 6. Distribution of TCEA1 and Pol II relative to class III genes. Tag densities for RPC4, TCEA1 (this work) and the hypophosphorylated (RNAPII) (64), serine 2 (RNAPII_S2P) (54), serine 5 (RNAPII_S5P) and serine 7 (RNAPII_S7P) phosphorylated forms of Pol II, are shown relative to the location of the RNA 5′-end for bound (**A**) tRNA, (**B**) SINEs and (**C**) BRF2-associated genes. The tag density scale for RPC4 is indicated on the right side of the plots.

preparation), Pol II and TCEA1 distributions were similar (Figure 7). Indeed for the 10 000 most highly expressed genes, the Spearman correlation coefficient between Pol II and TCEA1 occupancies on class II genes was 0.77. A large fraction of the genes, active or inactive, have paused Pol II just after the TSS. We wondered if TCEA1 occupancy might correlate with the transition from pausing to elongation. This is not the case since the levels of Pol II and TCEA1 on the pausing regions (defined here as the 300 bp after the TSS) of actively transcribed or non-productive genes with paused Pol II were similar, with Spearman correlation coefficient of 0.59 and 0.54, respectively (see Figure 7B and C for examples of Pol II and TCEA1 distributions). The Spearman correlation coefficient between TCEA1 and Pol II occupancies downstream of the pausing region were 0.75 and 0.76, respectively, for the two classes of genes.

DISCUSSION

The genome-wide distribution of Pol III, TFIIIB-β, its variant form TFIIIB-α and TFIIIC has been established

in mouse ES cells using a highly specific ChIP-Seq procedure. Only 284 tRNA genes out of 526 predicted genes were indeed associated with Pol III and, hence, were likely to be transcribed. The fact that Pol III was found only on a very small number of SINEs suggests that mechanisms exist to prevent their transcription. Additionally, 30 sites on the mouse genome were associated with Pol III, but lacked any annotation, most of them being transcribed. The regions encoding the new transcripts were usually conserved in the rat genome but except in one case, this conservation did not extend to man. The chromatin environment of class III genes resembles strongly that of class II genes with high levels of H3K4me3 upstream and downstream of the gene and low levels around the transcription start site. Studies in human have indicated that Pol II is associated with class III transcription (29,30,32). However, we found different situations depending on whether we looked at BRF1- or BRF2-associated genes. Upstream of tRNA genes, only very low levels of hypophosphorylated Pol II were observed. On the contrary, Pol II was present upstream of BRF2-associated genes. Finally, we observed that TCEA1 isoform of TFIIS was associated with classes II and III genes. The distribution pattern of TCEA1 suggests that it

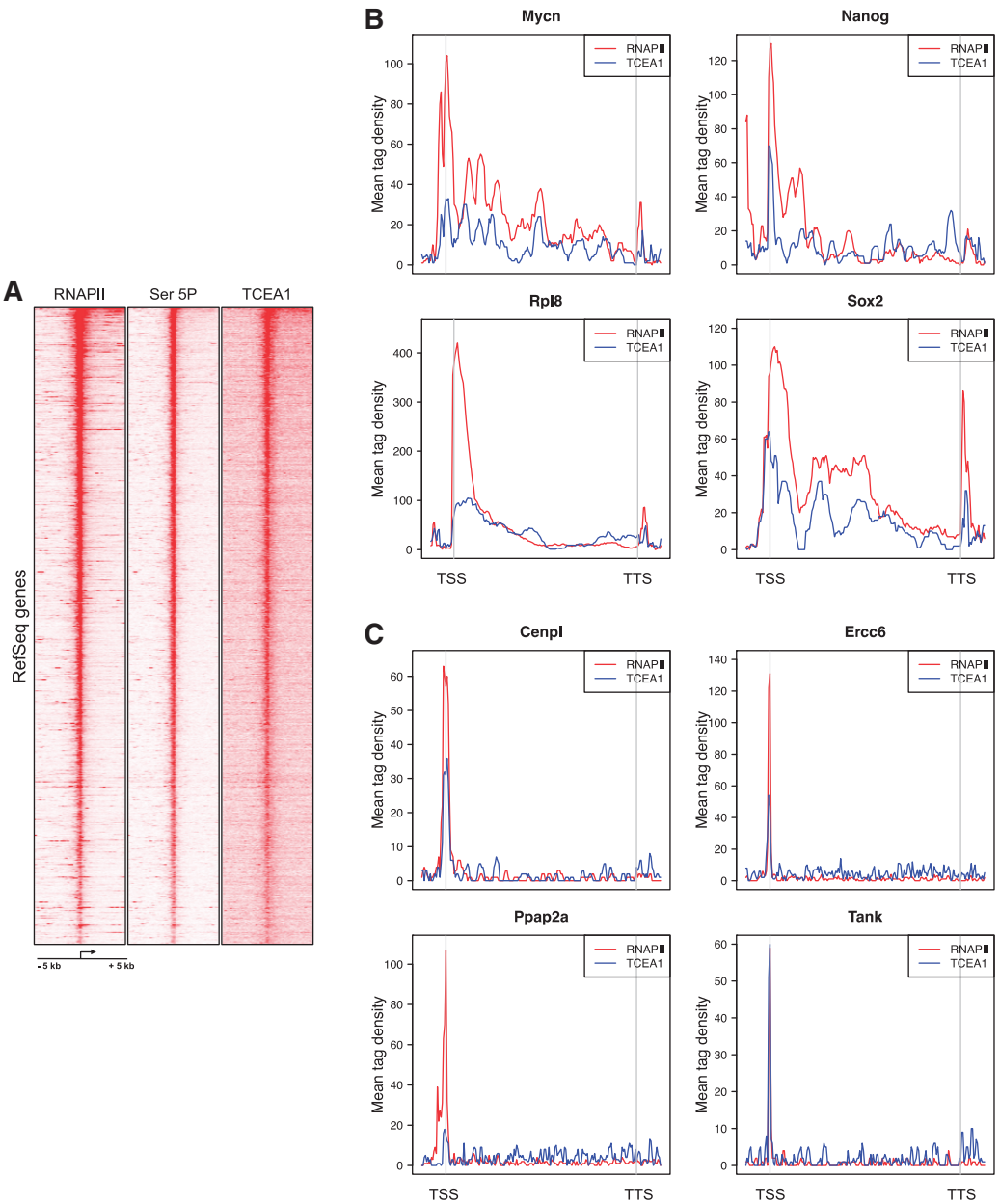


Figure 7. Distribution of TCEA1 on class II genes. (**A**) Hypophosphorylated Pol II (RNAPII) and TCEA1 have similar distribution on genes. The distribution of Pol II, S5P Pol II and TCEA1 5 kb upstream and downstream of the TSS is plotted for the 10 000 most expressed genes in ES cells based on RNA-Seq (M.G., manuscript in preparation). The genes were ordered according to their expression levels using seqMINER. (**B**) Examples of Pol II and TCEA1 profiles on highly transcribed genes. TTS: Transcription termination site. (**C**) Examples of Pol II and TCEA1 profiles on paused unproductive genes.

is not sufficient to stimulate the transition of paused Pol II into elongation. Moreover, it indicates that TFIIS could, as in yeast, be a Pol III general transcription factor.

Thanks to our highly specific tandem ChIP-Seq protocol and cross-validation by experiments that investigated independently the distribution of several subunits belonging to Pol III or TFIIIC, the mouse Pol III transcriptome is now precisely defined. As in human cell lines (29–33), the Pol III transcription machinery is associated with only a subset of the tRNA genes. This situation also holds true for the BRF2-associated genes. We estimated that <5% of the tRNA genes might have escaped our analysis. A total of 30 unannotated regions were associated with significant levels of Pol III. Several arguments indicate that they are indeed transcribed by Pol III. First, we could verify the presence of Pol III in independent experiments. Second, in addition to Pol III, most of these regions were associated with TBP and TFIIIC. In contrast, BRF2 was present only on regions that were already known to require TFIIIB-α for their transcription *in vitro*. Third, an RNA was found in >80% of the new regions that we tested. Finally, none of these regions was annotated. Experiments have suggested that a small number of miRNAs is transcribed by Pol III in human (55,56). This view was later challenged by other experiments (57). In mouse ES cells, none of the newly identified regions overlapped with miRNA annotations suggesting that the transcription of this class of RNA by Pol III is very limited. The function, if any, of the RNA transcribed from the new Pol III-associated regions will await further studies.

In vitro transcription experiments have demonstrated that Pol III is able to transcribe the SINEs (4). It is also known that SINEs, which number in hundreds of thousands in mammalian genomes, result from the insertion of retrotranscribed Pol III transcripts (3). We thus wondered if we could estimate how many of them are indeed associated with the Pol III transcription machinery. Unexpectedly, only 241 SINEs were bound, usually at rather low levels compared with tRNAs. The transcribed SINEs were distributed throughout the genome independently of tRNA or other class III genes. The small number of transcribed SINEs suggests that a general mechanism might repress their expression. This situation might, however, be specific to ES cells.

In human cells, class II genes are closely associated with class III genes. Pol II transcribed genes are often present upstream of class III genes (29–32). Even though Pol II transcription inhibition has only a modest effect on the transcription of adjacent class III genes, it has been argued that Pol II might regulate Pol III transcription (32). Our observations in mouse indicated that, at actively transcribed tRNA genes, the level of Pol II is extremely low. Moreover, while the hypophosphorylated non-elongating form of Pol II was present upstream of the bound tRNA genes at very low levels, the other forms were virtually absent. In contrast, hypophosphorylated, S7P and S5P Pol II were all present upstream of BRF2-associated genes. This observation suggests that, in mouse, the chromatin organization around transcribed tRNA genes, allows low levels of hypophosphorylated Pol II to bind but prevents the transition to elongation. On the other hand, the promoters of BRF2-associated genes, which are gene external and resemble those of Pol II transcribed genes, would allow the association of class II transcription machinery and divergent transcription relative to Pol III. The observation of different behaviors for tRNA genes and BRF2-associated genes strongly suggests that, on tRNA genes, Pol II presence has probably limited functional significance in mouse. Type III promoters are gene external and their transcription by Pol III is determined by the presence of a TATA box which, if absent, leads to transcription by Pol II (58,59). It has also been shown that U6 transcription is decreased ~2-fold in human cell lines when Pol II transcription is inhibited (60). It is thus possible that, as in human, the presence of Pol II upstream of BRF2-associated genes in mouse might stimulate their transcription by Pol III.

The binding pattern of TFIIIC was clearly distinct from that of Pol III. Whereas the factor was present on tRNA genes and on the new Pol III-associated regions, more than 2200 TFIIIC-binding sites were located far away from class III genes. This number is similar to that of the ETC loci in K562 human cells where 1865 such sites were found (31). The ETC sites have first been detected in *S. cerevisiae* and *S. pombe* and have been shown to play a role in the organization of chromatin (26,35,51). Even though the ChIP-Seq experiments were performed with different protocols, antibodies and in different cell lines, the ETC site number in human is remarkably similar to that found in mouse, both organisms having comparably sized genomes (3.1 Gb and 2.9 Gb for man and mouse, respectively). CTCF has been shown to have an enhancer-blocking activity (36). In mouse ES cell lines, CTCF is loosely associated with ETC loci, a relationship that might be indicative of an interdependence and/or common role in silencing transcription of adjacent sequences. Cohesins are required for the localization of CTCF at boundary elements (52). In line with previous observations, in mouse and human, Smc1a and Smc3 distributions closely follow that of CTCF. Remarkably, Smc1a and Smc3 were also present on the TFIIIC peak. In *S. cerevisiae* mutations in Smc1 or Smc3 affects the boundary function of tRNA genes (35,50,51). Moreover, a recent report has shown that cohesins connect enhancers and promoters through DNA loops via interactions with the Mediator, stimulating Pol II transcription (45). These observations raise the intriguing possibility that the presence of cohesins at ETCs and CTFC-binding sites might stimulate the formation of DNA loops, promoting an enhancer-blocking activity of TFIIIC and shape the organization of chromatin.

In *S. cerevisiae*, we showed that TFIIS occupies nearly all class III genes. In addition, some TFIIS mutations specifically affect Pol II or Pol III transcription *in vitro* and *in vivo* demonstrating that TFIIS plays a key role in Pol III transcription in yeast (22). The observation that, in mouse ES cells, TCEA1 is associated with the majority of class III genes independently of Pol II, strongly suggests that TFIIS is also a Pol III transcription factor in mammals. Intriguingly, its distribution on class III genes is skewed toward the TSS when compared to that of Pol III, in line

with a putative role in transcription initiation as we proposed in yeast.

Previous studies in *Drosophila* have shown that TFIIS plays a role in stimulating paused Pol II to enter active elongation (61,62), a situation that could also hold in mouse ES cells. We wondered if the presence or absence of TFIIS could be related to pausing on class II genes. TFIIS occupancy on class II genes followed that of Pol II both on elongating genes and on non-productive genes that have high levels of paused Pol II, indicating that the pause is not the consequence of the absence of TFIIS. Our observations thus raise the intriguing possibility that TFIIS activity, but not its recruitment, might be somehow stimulated upon the entry into elongation.

In summary, this study provides a high-resolution map of class III genes active in mouse ES cells and it shows that the TFIIIC transcription factor is present at around 2000 sites on the genome independently of Pol III and its other factors TFIIIB-α or -β, possibly playing a role in chromatin organization. Finally, we provide evidence that support a role for TFIIS Pol II elongation factor in class III transcription.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank I. Davidson for inspiring us to start this project. We thank P. Thuriaux and H. Neil-Bernet for critical reading of the manuscript. We thank T. Ye and A. Krebs for providing seqMINER before publication, M. de Dieuleveult, R. Fenouil, Y Duffour and N. Naouar for help with bioinformatics, I. Hmitou for help with some experiments, U. Rogner and A. Smith for the gift of the 46C ES cell line.

FUNDING

The Agence Nationale de la Recherche (ANR-05-BLAN-0396); the Association pour la Recherche sur le Cancer (3164); the Association Française contre les Myopathies (MNM2 2008-13630); the Ile-de-France Region (grant 2745 to L.C.); the Fondation pour la Recherche Médicale (grant 382-2010 to L.C.). Funding for open access charge: Commissariat à l'Energie Atomique et aux Energies Alternatives.

Conflict of interest statement. None declared.

REFERENCES

1. Paule,M.R. and White,R.J. (2000) Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res.*, **28**, 1283–1298.
2. Dieci,G., Fiorino,G., Castelnuovo,M., Teichmann,M. and Pagano,A. (2007) The expanding RNA polymerase III transcriptome. *Trends Genet.*, **23**, 614–622.
3. Kramerov,D.A. and Vassetzky,N.S. (2005) Short retroposons in eukaryotic genomes. *Int. Rev. Cytol.*, **247**, 165–221.
4. White,R.J. (1998) *RNA Polymerase III Transcription*, 2nd edn. Springer, Berlin.
5. Schramm,L. and Hernandez,N. (2002) Recruitment of RNA polymerase III to its target promoters. *Genes Dev.*, **16**, 2593–2620.
6. Geiduschek,E.P. and Kassavetis,G.A. (2001) The RNA polymerase III transcription apparatus. *J. Mol. Biol.*, **310**, 1–26.
7. Dumay-Odelot,H., Marck,C., Durrieu-Gaillard,S., Lefebvre,O., Jourdain,S., Prochazkova,M., Pflieger,A. and Teichmann,M. (2007) Identification, molecular cloning, and characterization of the sixth subunit of human transcription factor TFIIIC. *J. Biol. Chem.*, **282**, 17179–17189.
8. Teichmann,M. and Seifart,K.H. (1995) Physical separation of two different forms of human TFIIIB active in the transcription of the U6 or the VAI gene in vitro. *EMBO J.*, **14**, 5974–5983.
9. Brun,I., Sentenac,A. and Werner,M. (1997) Dual role of the C34 subunit of RNA polymerase III in transcription initiation. *EMBO J.*, **16**, 5730–5741.
10. Wang,Z. and Roeder,R.G. (1997) Three human RNA polymerase III-specific subunits form a subcomplex with a selective function in specific transcription initiation. *Genes Dev.*, **10**, 1315–1326.
11. Sadowski,C.L., Henry,R.W., Lobo,S.M. and Hernandez,N. (1993) Targeting TBP to a non-TATA box cis-regulatory element: a TBP-containing complex activates transcription from snRNA promoters through the PSE. *Genes Dev.*, **7**, 1535–1548.
12. Murphy,S., Yoon,J.B., Gerster,T. and Roeder,R.G. (1992) Oct-1 and Oct-2 potentiate functional interactions of a transcription factor with the proximal sequence element of small nuclear RNA genes. *Mol. Cell Biol.*, **12**, 3247–3261.
13. Waldschmidt,R., Wanandi,I. and Seifart,K.H. (1991) Identification of transcription factors required for the expression of mammalian U6 genes in vitro. *EMBO J.*, **10**, 2595–2603.
14. Teichmann,M., Wang,Z. and Roeder,R.G. (2000) A stable complex of a novel transcription factor IIB- related factor, human TFIIIB50, and associated proteins mediate selective transcription by RNA polymerase III of genes with upstream promoter elements. *Proc. Natl Acad. Sci. USA*, **97**, 14200–14205.
15. Schramm,L., Pendergrast,P.S., Sun,Y. and Hernandez,N. (2000) Different human TFIIIB activities direct RNA polymerase III transcription from TATA-containing and TATA-less promoters. *Genes Dev.*, **14**, 2650–2663.
16. Warner,J.R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.*, **24**, 437–440.
17. White,R.J. (2008) RNA polymerases I and III, non-coding RNAs and cancer. *Trends Genet.*, **24**, 622–629.
18. Marshall,L., Kenneth,N.S. and White,R.J. (2008) Elevated tRNA(iMet) synthesis can drive cell proliferation and oncogenic transformation. *Cell*, **133**, 78–89.
19. Wind,M. and Reines,D. (2000) Transcription elongation factor SII. *Bioessays*, **22**, 327–336.
20. Kim,B., Nesvizhskii,A.I., Rani,P.G., Hahn,S., Aebersold,R. and Ranish,J.A. (2007) The transcription elongation factor TFIIS is a component of RNA polymerase II preinitiation complexes. *Proc. Natl Acad. Sci. USA*, **104**, 16068–16073.
21. Guglielmi,B., Soutourina,J., Esnault,C. and Werner,M. (2007) TFIIS elongation factor and Mediator act in conjunction during transcription initiation in vivo. *Proc. Natl Acad. Sci. USA*, **104**, 16062–16067.
22. Ghavi-Helm,Y., Michaut,M., Acker,J., Aude,J.C., Thuriaux,P., Werner,M. and Soutourina,J. (2008) Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription. *Genes Dev.*, **22**, 1934–1947.
23. Ito,T., Seldin,M.F., Taketo,M.M., Kubo,T. and Natori,S. (2000) Gene structure and chromosome mapping of mouse transcription elongation factor S-II (Tcea1). *Gene*, **244**, 55–63.
24. Ito,T., Xu,Q., Takeuchi,H., Kubo,T. and Natori,S. (1996) Spermatocyte-specific expression of the gene for mouse testis-specific transcription elongation factor S-II. *FEBS Lett.*, **385**, 21–24.
25. Taira,Y., Kubo,T. and Natori,S. (1998) Molecular cloning of cDNA and tissue-specific expression of the gene for SII-K1, a novel transcription elongation factor SII. *Genes Cells*, **3**, 289–296.
26. Moqtaderi,Z. and Struhl,K. (2004) Genome-wide occupancy profile of the RNA polymerase III machinery in Saccharomyces

cerevisiae reveals loci with incomplete transcription complexes. *Mol. Cell Biol.*, **24**, 4118–4127.
27. Roberts,D.N., Stewart,A.J., Huff,J.T. and Cairns,B.R. (2003) The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. *Proc. Natl Acad. Sci. USA*, **100**, 14695–14700.
28. Harismendy,O., Gendrel,C.G., Soularue,P., Gidrol,X., Sentenac,A., Werner,M. and Lefebvre,O. (2003) Genome-wide location of yeast RNA polymerase III transcription machinery. *EMBO J.*, **22**, 4738–4747.
29. Barski,A., Chepelev,I., Liko,D., Cuddapah,S., Fleming,A.B., Birch,J., Cui,K., White,R.J. and Zhao,K. (2010) Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat. Struct. Mol. Biol.*, **17**, 629–634.
30. Oler,A.J., Alla,R.K., Roberts,D.N., Wong,A., Hollenhorst,P.C., Chandler,K.J., Cassiday,P.A., Nelson,C.A., Hagedorn,C.H., Graves,B.J. *et al.* (2010) Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat. Struct. Mol. Biol.*, **17**, 620–628.
31. Moqtaderi,Z., Wang,J., Raha,D., White,R.J., Snyder,M., Weng,Z. and Struhl,K. (2010) Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nat. Struct. Mol. Biol.*, **17**, 635–640.
32. Raha,D., Wang,Z., Moqtaderi,Z., Wu,L., Zhong,G., Gerstein,M., Struhl,K. and Snyder,M. (2010) Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc. Natl Acad. Sci. USA*, **107**, 3639–3644.
33. Canella,D., Praz,V., Reina,J.H., Cousin,P. and Hernandez,N. (2010) Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. *Genome Res.*, **20**, 710–721.
34. Guffanti,E., Percudani,R., Harismendy,O., Soutourina,J., Werner,M., Iacovella,M.G., Negri,R. and Dieci,G. (2006) Nucleosome depletion activates poised RNA polymerase III at unconventional transcription sites in Saccharomyces cerevisiae. *J. Biol. Chem.*, **281**, 29155–29164.
35. Noma,K., Cam,H.P., Maraia,R.J. and Grewal,S.I. (2006) A role for TFIIIC transcription factor complex in genome organization. *Cell*, **125**, 859–872.
36. Wallace,J.A. and Felsenfeld,G. (2007) We gather together: insulators and genome organization. *Curr. Opin. Genet. Dev.*, **17**, 400–407.
37. Liu,P., Jenkins,N.A. and Copeland,N.G. (2003) A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome Res.*, **13**, 476–484.
38. Ying,Q.L., Stavridis,M., Griffiths,D., Li,M. and Smith,A. (2003) Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat. Biotech.*, **21**, 183–186.
39. Tessarollo,L. (2001) Manipulating mouse embryonic stem cells. *Methods Mol. Biol.*, **158**, 47–63.
40. Esnault,C., Ghavi-Helm,Y., Brun,S., Soutourina,J., Van Berkum,N., Boschiero,C., Holstege,F. and Werner,M. (2008) Mediator-dependent recruitment of TFIIH modules in preinitiation complex. *Mol. Cell*, **31**, 337–346.
41. Kuras,L., Borggrefe,T. and Kornberg,R.D. (2003) Association of the Mediator complex with enhancers of active genes. *Proc. Natl Acad. Sci. USA*, **100**, 13887–13891.
42. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
43. Coughlin,D.J., Babak,T., Nihranz,C., Hughes,T.R. and Engelke,D.R. (2009) Prediction and verification of mouse tRNA gene families. *RNA Biol.*, **6**, 195–202.
44. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
45. Kagey,M.H., Newman,J.J., Bilodeau,S., Zhan,Y., Orlando,D.A., van Berkum,N.L., Ebmeier,C.C., Goossens,J., Rahl,P.B., Levine,S.S. *et al.* (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, **467**, 430–435.
46. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M., Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
47. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
48. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
49. Bilodeau,S., Kagey,M.H., Frampton,G.M., Rahl,P.B. and Young,R.A. (2009) SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev.*, **23**, 2484–2489.
50. Donze,D. and Kamakaka,R.T. (2001) RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in Saccharomyces cerevisiae. *EMBO J.*, **20**, 520–531.
51. Valenzuela,L., Dhillon,N. and Kamakaka,R.T. (2009) Transcription independent insulation at TFIIIC-dependent insulators. *Genetics*, **183**, 131–148.
52. Parelho,V., Hadjur,S., Spivakov,M., Leleu,M., Sauer,S., Gregson,H.C., Jarmuz,A., Canzonetta,C., Webster,Z., Nesterova,T. *et al.* (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, **132**, 422–433.
53. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
54. Rahl,P.B., Lin,C.Y., Seila,A.C., Flynn,R.A., McCuine,S., Burge,C.B., Sharp,P.A. and Young,R.A. (2010) c-Myc regulates transcriptional pause release. *Cell*, **141**, 432–445.
55. Borchert,G.M., Lanier,W. and Davidson,B.L. (2006) RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.*, **13**, 1097–1101.
56. Ozsolak,F., Poling,L.L., Wang,Z., Liu,H., Liu,X.S., Roeder,R.G., Zhang,X., Song,J.S. and Fisher,D.E. (2008) Chromatin structure analyses identify miRNA promoters. *Genes Dev.*, **22**, 3172–3183.
57. Bortolin-Cavaille,M.L., Dance,M., Weber,M. and Cavaille,J. (2009) C19MC microRNAs are processed from introns of large Pol-II, non-protein-coding transcripts. *Nucleic Acids Res.*, **37**, 3464–3473.
58. Lobo,S.M. and Hernandez,N. (1989) A 7 bp mutation converts a human RNA polymerase II snRNA promoter into an RNA polymerase III promoter. *Cell*, **58**, 55–67.
59. Mattaj,I.W., Dathan,N.A., Parry,H.D., Carbon,P. and Krol,A. (1988) Changing the RNA polymerase specificity of U snRNA gene promoters. *Cell*, **55**, 435–442.
60. Listerman,I., Bledau,A.S., Grishina,I. and Neugebauer,K.M. (2007) Extragenic accumulation of RNA polymerase II enhances transcription by RNA polymerase III. *PLoS Genet.*, **3**, e212.
61. Nechaev,S., Fargo,D.C., dos Santos,G., Liu,L., Gao,Y. and Adelman,K. (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, **327**, 335–338.
62. Adelman,K., Marr,M.T., Werner,J., Saunders,A., Ni,Z., Andrulis,E.D. and Lis,J.T. (2005) Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Mol. Cell*, **17**, 103–112.
63. Ye,T., Krebs,A.R., Choukrallah,M.A., Keime,C., Plewniak,F., Davidson,I. and Tora,L. (2010) seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.*, **39**, e35.
64. Seila,A.C., Calabrese,J.M., Levine,S.S., Yeo,G.W., Rahl,P.B., Flynn,R.A., Young,R.A. and Sharp,P.A. (2008) Divergent transcription from active promoters. *Science*, **322**, 1849–1851.

Supplementary Methods

New Pol III transcripts

For new Pol III transcripts, we calculated the density of RPC4 on QuEST regions. Centered on the peak of density of RPC4, a region of 200 nt was defined.

TFIIIC cartography

The location of the TFIIIC subunits was determined using QuEST software. We used the predefined relaxed peak calling parameters (threshold parameter set at 3) and a 600 nt region size for each subunit. We next retained only regions bound by the three subunits.

ChIP-Seq density

Reads were extended to an assumed fragment length of 300 nt for Pol III and TFIIIB subunits, and 150 nt for TFIIIC subunits and TCEA1. ChIP-seq enrichment for the indicated factor, is defined as count of read, and was determined in 1 bp bin. Average density is the summed of this density per nt calculated in a determined window.

Protein profile density

Genes were aligned according to the location of the 5' RNA end. Read density was determined for each protein in a 500 bp or a 1000 bp window centered on the 5' RNA end, in 1bp bin (or per nt). Mean read density for the selected genes was then calculated and plotted over the window.

Determination of promoter proximal regions and gene bodies for class II transcripts

When several class II transcripts originated from alternative promoters, we calculated the density of Pol II around the TSS (-30 to +300) and on the gene body (+300 to the TTS) for the transcript with the highest density on the proximal promoter. RefSeq gene annotations (UCSC mm9 table) were used.

Active and non-productive genes

The active genes set was defined as the first 2000 genes according the RNA abundance as measured by RNA-Seq. These genes displayed Pol II, and H3K4me3 and H3K79me2 chromatin modifications. The non-productive set of genes consisted of those that were ranked between the 6001 and 8000 position according to the RNA-Seq experiment. We verified that Pol II was present in the promoter proximal region and that the H3K79me2 elongation mark was absent within the 5 kb downstream of the TSS.

Supplementary Figure legends

Figure S1. Construction of the ES tagged cell lines by homologous recombination.

The cartoon depicts how the sequence encoding the 6His-Flag-HA tag was inserted in the 46C cell line at the 3' end of the ORFs encoding the proteins of interest as described in the Methods section.

Figure S2. Western blotting of tagged proteins in ES cell lines.

Tagged proteins from untagged 46C cell line, which served as a negative control (-) or tagged cell lines, as indicated, were analyzed by western blotting with anti-HA antibodies to verify the expression and size of the fusion proteins.

Figure S3. Tandem chromatin immunoprecipitation of the Pol III transcription machinery in mouse ES cell lines.

Tandem ChIP experiments of chromatin extracts of (A) RPC1, (B) BRF1 and (C) BRF2 cell lines were performed with the HA7 H-3663 anti-HA antibody (IP #1) and with F-1804 anti-Flag antibody after elution of proteins with an HA peptide (IP #2) as described in the legend of Figure 1.

Figure S4. Distribution of tags on an active U6 gene.

The region shown is located on chromosome 10. The color code used is as in Figure 2.

Figure S5. Presence or absence of Pol III on class III genes.

The presence or absence of Pol III on genes predicted to be associated or not by ChIP-Seq experiments was verified in independent ChIP experiment using chromatin extracts from RPC1 or RPC4 cell lines. 7SL unbound, 7SK unbound, U6 unbound, HY3 unbound, tRNA-Proline and tRNA-Lysine (AAG) correspond to genes that were not found to be associated with Pol III in ChIP-Seq experiments. The 46C untagged cell line was used as a negative control. The mean and standard deviation (error bars) of three independent biological replicates of the ChIP experiments are plotteds.

Figure S6. Genomic distribution of Pol III-associated genes.

(A) Number of Pol III-associated class-III genes on the mouse chromosomes. (B) Distribution of predicted tRNA genes, Pol III-bound tRNA genes and SINEs, new genes, other BRF1-associated class III genes and BRF2 on the mouse physical map.

Figure S7. Identification of RNA associated with Pol III-bound un-annotated regions.

(A) Verification of the association of Pol III with un-annotated regions. ChIP experiments were performed as described in Figure 1. The mean and standard deviation (error bars) of three independent biological replicates of the ChIP experiments are plotted. (B) Detection of RNAs associated with the un-annotated Pol III-associated regions. RNAs originating from unannotated bound regions (2039 and 1283; see Table S5) were detected by RT-PCR experiments performed as described in Materials and Methods. (C) Alignement of the *Mus musculus* (M.m.) sequence (chromosome 3 from 96,246,175 to 96,246,382) surrounding conserved region 2819 and the homologous region of *Homo sapiens* (chromosome 8 from 96415864 to 96416175). The A-box, B-box and terminators sequences (Ter) are highlighted by blue, green and red rectangles respectively.

Figure S8. Conservation of Pol III-associated SINEs.

(A) Number of SINEs as a function of the percentage of divergence. The divergence percentage of the SINEs relative to the canonical SINE sequence was obtained from the UCSC mm9 database. The number of SINEs was plotted for one percent bins. (B) Conservation of the Pol III-associated SINEs.

Figure S9. Distribution of chromatin marks relative to class III genes.

Tag densities are shown for H3K4me1, -me2 (Meissner et al. 2008), -me3, H3K27me3 (Rugg-Gunn et al. 2010) and H3K9me3 (Bilodeau et al. 2009) relative to the RNA 5'end for (A) tRNA genes and (B) BRF2-associated genes.

Figure S10. Distribution of TCEA1 with respect to active class III genes.

The distribution of TCEA1 and RPC4 5 kb upstream and downstream of class-III genes by Pol III is shown. The genes are ordered according to the strength of TCEA1 binding.

Table S1: Pol III, TFIIIC, TFIIIB and TFIIS data.

The mouse Pol III subunits are labeled using the nomenclature proposed by Schramm and Hernandez (2002) for the human enzyme. The homologous *S. cerevisiae* subunits are labeled as in Werner et al. (2009). The SwissProt accession numbers are given. The subunits investigated in this study are set in bold type.

Pol III subunits	MW (kD)	Swissprot accession	*S. cerevisiae* names
RPC1	155.7	B2RXC6	Rpc160
RPC2	127.6	P59470	Rpc128
RPC3	60.7	Q9D483	Rpc82
RPC4	44.3	Q91WD1	Rpc53
RPC5	79.9	Q9CZT4	Rpc37
RPC6	35.7	Q921X6	Rpc34
RPC7	25.9	Q6NXY9	Rpc31
RPC8	22.9	Q9D2C6	Rpc25
RPC9	16.7	O35427	Rpc17
RPC10	12.3	Q9CQZ7	Rpc11
RPAC1	39.1	P52432	Rpc40
RPAC2	15.1	P97304	Rpc19
RPABC1	24.6	Q80UW8	Rpb5
RPABC2	14.5	P61219	Rpb6
RPABC3	17.1	Q923G2	Rpb8
RPABC4	7.0	Q63871	Rpb12
RPABC5	7.6	P62876	Rpb10
TFIIIC subunits			
TFIIIC220	237.5	Q8K284	Tfc3
TFIIIC110	100.3	Q8BL74	Tfc6
TFIIIC102	Unknown	Unknown	Tfc4
TFIIIC90	91.7	Q8BMQ2	Tfc8
TFIIIC63	60.5	Q8R2T8	Tfc1
TFIIIC35	25.5	Q9D8P7	Tfc7
TFIIIB subunits			
BRF1	73.8	Q8CFK2	Brf1
BRF2	47.1	Q3UAW9	Brf2
BDP1	270.8	Q571C7	Bdp1
TBP	34.7	P29037	TBP
TFIIS isoforms			
TCEA1	33.9	P10711	Dst1
TCEA2	33.7	Q9QVN7	Dst1
TCEA3	38.9	P23881	Dst1

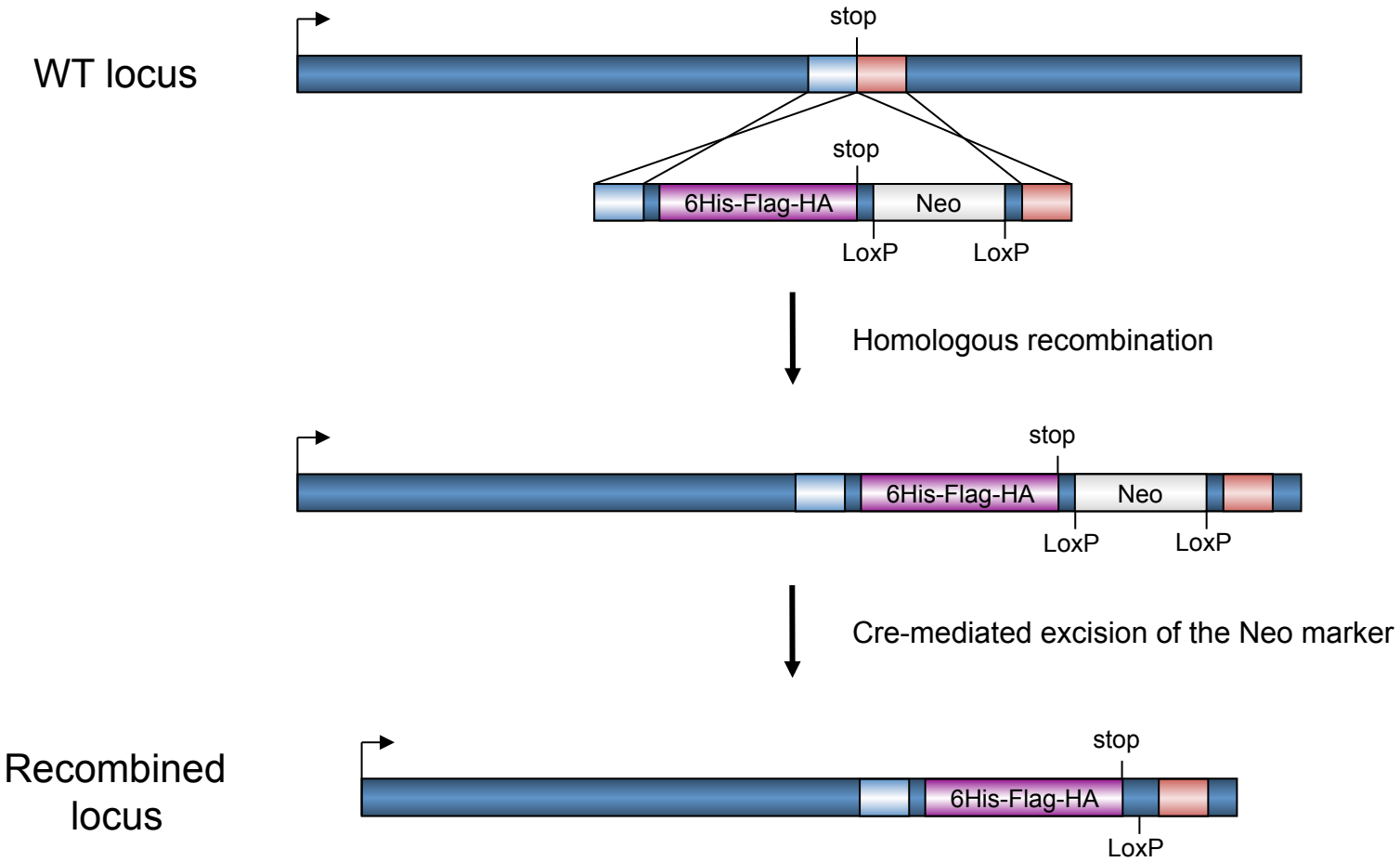
Table S2: Sequence of the C-terminal tag and of the insertion cassette.
A: Sequence of the 36 amino acid 6 histidines-Flag-HA tag.
GAPHHHHHHGAAGGDYKDDDDKSAAGGYPYDVPDYA

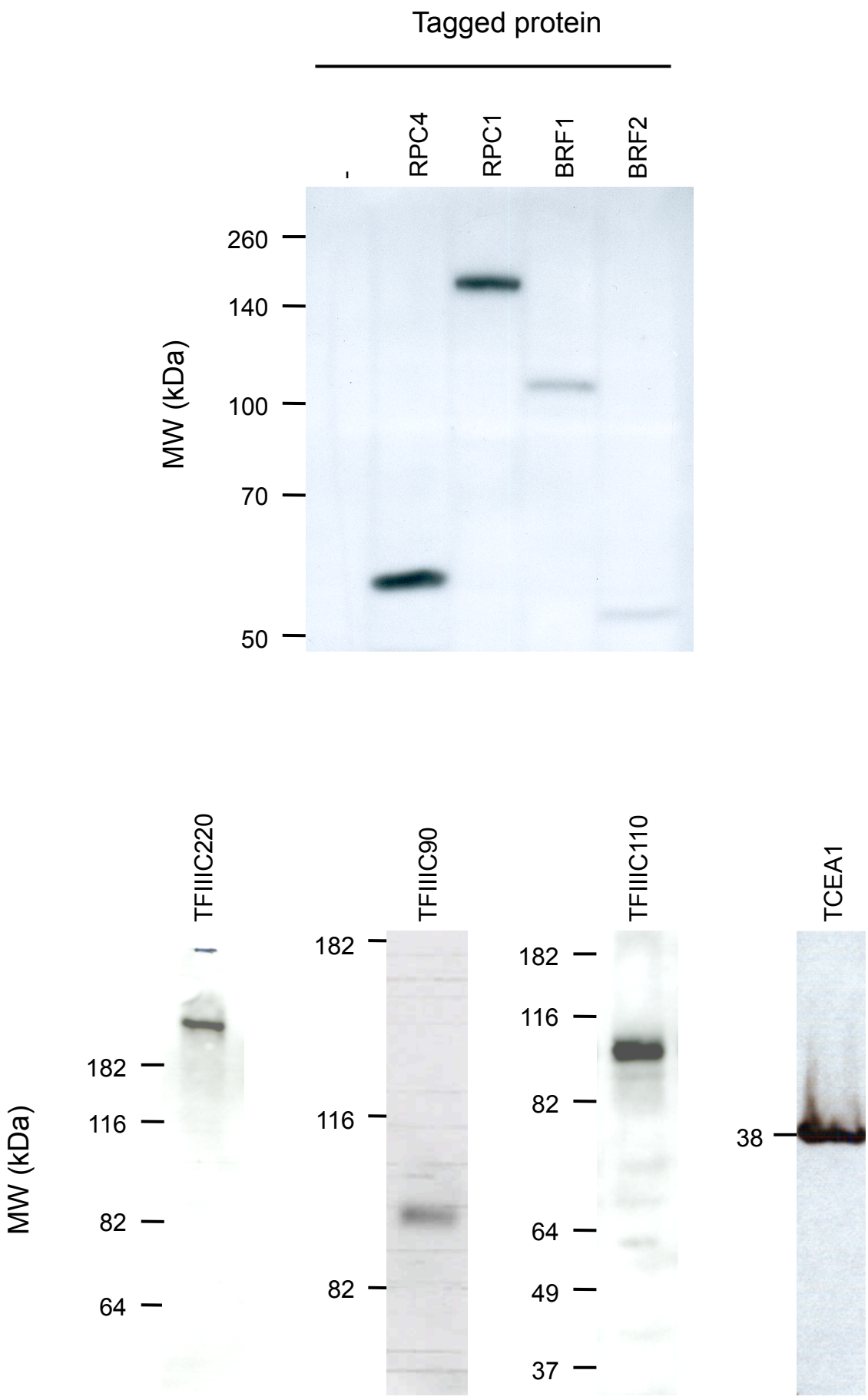
B: DNA sequence of the cassette encoding the tag and the neomycin marker.
GTCGACCTCGAGGGCGCGCCCCATCACCATCACCACCATGGGGCCGCTGG
AGGAGACTACAAGGACGACGATGACAAGTCGGCCGCTGGAGGATACCCCT
ACGACGTGCCCGACTACGCCTAGGGTACCTCTAGAGAATTCCTGCAGCCC
AATTCCGATCATATTCAATAACCCTTAATATAACTTCGTATAATGTATGC
TATACGAAGTTATTAGGTCTGAAGAGGAGTTTACGTCCAGCCAAGCTAGC
TTGGCTGCAGGTCGTCGAAATTCTACCGGGTAGGGGAGGCGCTTTTCCCA
AGGCAGTCTGGAGCATGCGCTTTAGCAGCCCCGCTGGGCACTTGGCGCTA
CACAAGTGGCCTCTGGCCTCGCACACATTCCACATCCACCGGTAGGCGCC
AACCGGCTCCGTTCTTTGGTGGCCCCTTCGCGCCACCTTCTACTCCTCCC
CTAGTCAGGAAGTTCCCCCCCGCCCCGCAGCTCGCGTCGTGCAGGACGTG
ACAAATGGAAGTAGCACGTCTCACTAGTCTCGTGCAGATGGACAGCACCG
CTGAGCAATGGAAGCGGGTAGGCCTTTGGGGCAGCGGCCAATAGCAGCTT
TGCTCCTTCGCTTTCTGGGCTCAGAGGCTGGGAAGGGGTGGGTCCGGGGG
CGGGCTCAGGGGCGGGCTCAGGGGCGGGGCGGGCGCCCGAAGGTCCTCCG
GAGGCCCGGCATTCTGCACGCTTCAAAAGCGCACGTCTGCCGCGCTGTTC
TCCTCTTCCTCATCTCCGGGCCTTTCGACCTGCAGCCTGTTGACAATTAA
TCATCGGCATAGTATATCGGCATAGTATAATACGACAAGGTGAGGAACTA
AACCATGGGATCGGCCATTGAACAAGATGGATTGCACGCAGGTTCTCCGG
CCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGGGCACAACAGACAATC
GGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGT
TCTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACG
AGGCAGCGCGGCTATCGTGGCTGGCCACGACGGGCGTTCCTTGCGCAGCT
GTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTGGCTGCTATTGGGCGA
AGTGCCGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAG
TATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCT
ACCTGCCCATTCGACCACCAAGCGAAACATCGCATCGAGCGAGCACGTAC
TCGGATGGAAGCCGGTCTTGTCGATCAGGATGATCTGGACGAAGAGCATC
AGGGGCTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCGCATGCCC
GACGGCGATGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATAT
CATGGTGGAAAATGGCCGCTTTTCTGGATTCATCGACTGTGGCCGGCTGG
GTGTGGCGGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATATTGCT
GAAGAGCTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTAT
CGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGT
TCTTCTGAGGGGATCAATTCTCTAGAGCTCGCTGATCAGCCTCGACTGTG
CCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTT
GACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAA
TTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTG
GGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCATGCTGG
GGATGCGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAACCAGCTGGGGCT
CGACTAGAGCTTGCGGAACCCTTAATATAACTTCGTATAATGTATGCTAT
ACGAAGTTATTAGGTCCCTCGAGGGGATCCACTAGTTCTAGAGCGGCCGC
C

Table S3: Sequencing experiments characteristics.

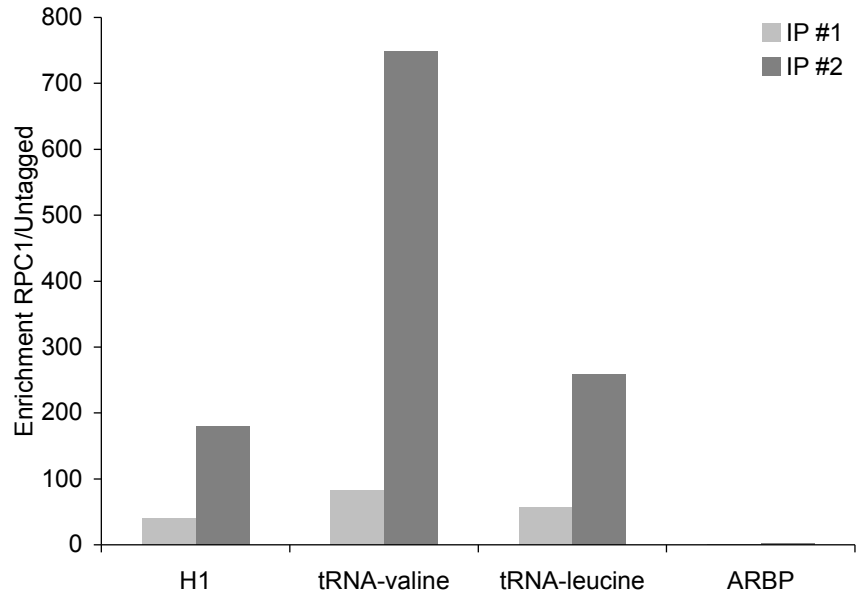
Cell Line	DNA fragmentation method	Number of mappable tags	Tag length
46C	sonication	3 796 321	26
RPC1	sonication	4 293 724	26
RPC4	sonication	6 209 726	26
BRF1	sonication	3 698 139	26
BRF2	sonication	5 932 485	26
46C	MNase	10 058 319	36
TFIIIC220	MNase	13 532 738	36
TFIIIC110	MNase	11 927 602	36
TFIIIC90	MNase	12 955 511	36
TCEA1	MNase	15 030 761	101*

*Paired-end sequencing was used. The sequences were aligned with Bowtie instead of Eland.

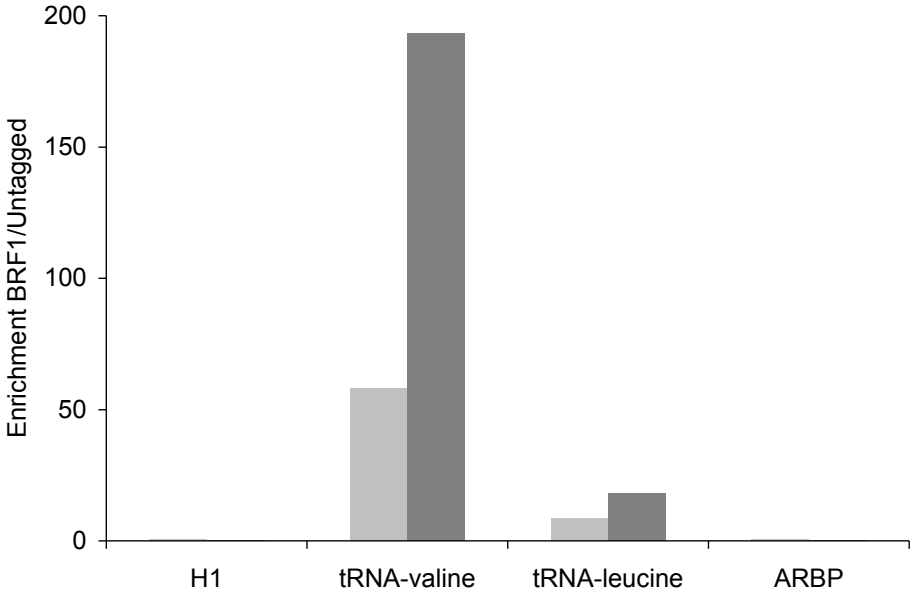

WT locus
stop
stop
6His-Flag-HA
Neo
LoxP
LoxP
Homologous recombination
stop
6His-Flag-HA
Neo
LoxP
LoxP
Cre-mediated excision of the Neo marker
Recombined locus
stop
6His-Flag-HA
LoxP


Tagged protein
-
RPC4
RPC1
BRF1
BRF2
MW (kDa)
260
140
100
70
50
TFIIIC220
TFIIIC90
TFIIIC110
TCEA1
MW (kDa)
182
116
82
64
182
116
82
182
116
82
64
49
37
38

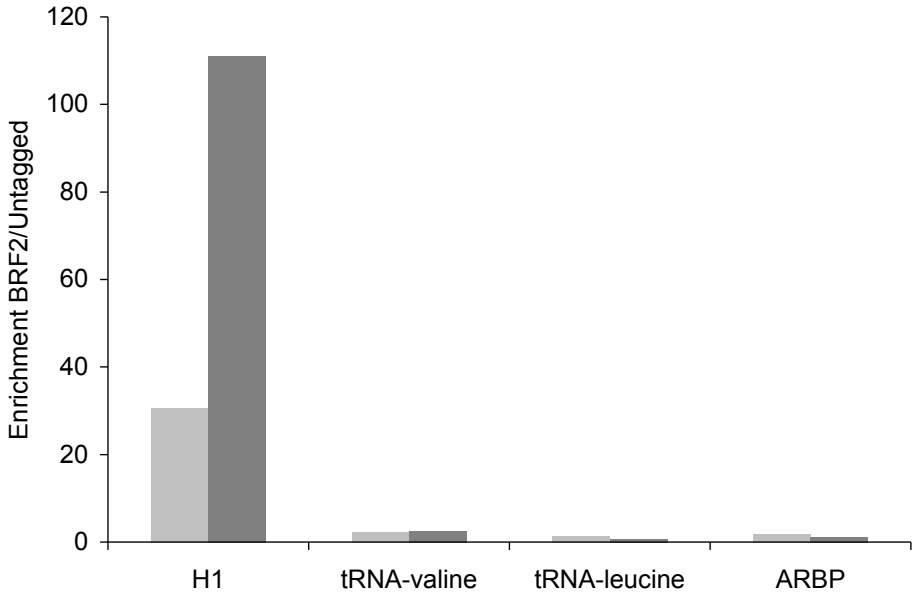
A


IP #1
IP #2
Enrichment RPC1/Untagged
800
700
600
500
400
300
200
100
0
H1
tRNA-valine
tRNA-leucine
ARBP

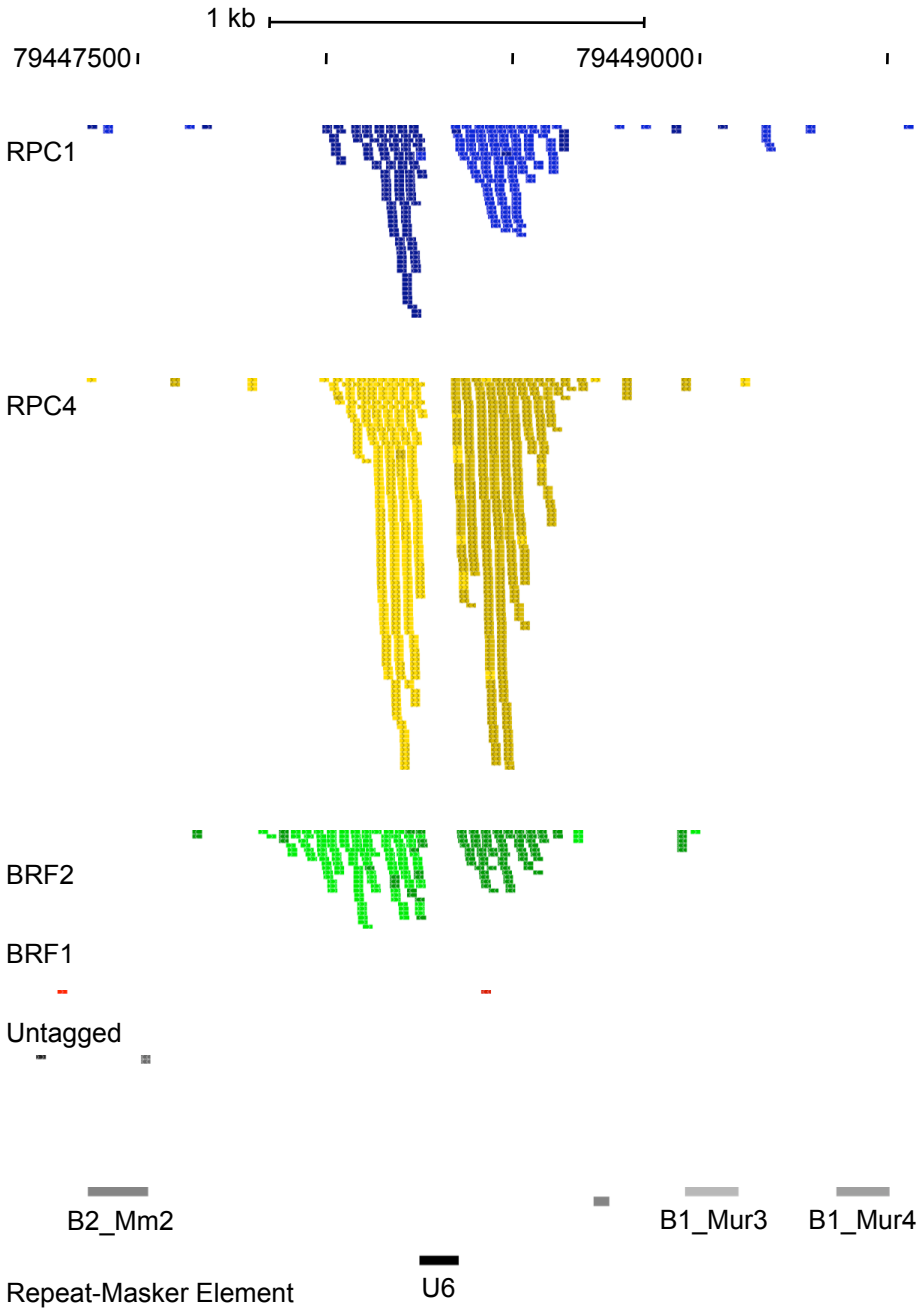
B


Enrichment BRF1/Untagged
200
150
100
50
0
H1
tRNA-valine
tRNA-leucine
ARBP

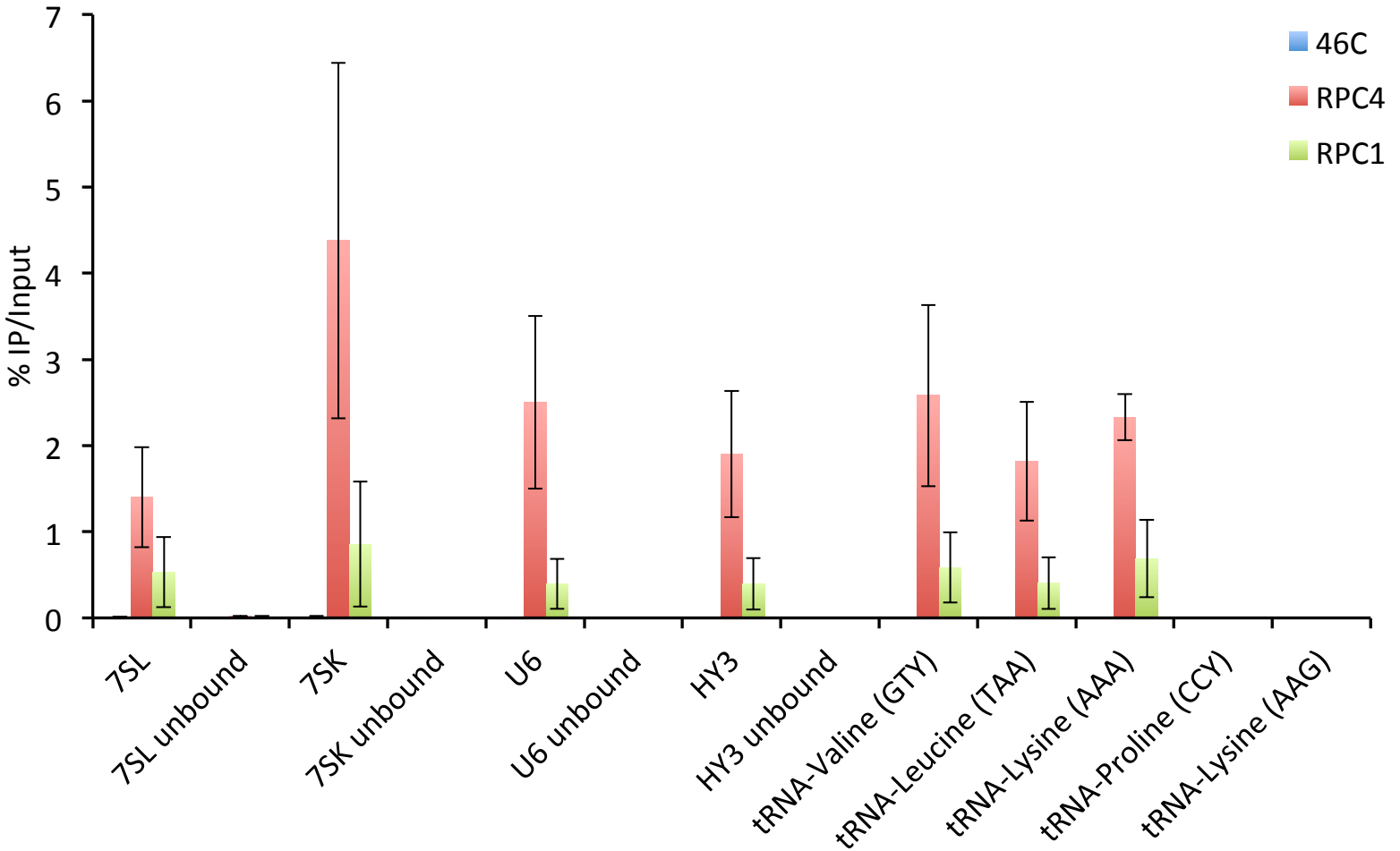
C


Enrichment BRF2/Untagged
120
100
80
60
40
20
0
H1
tRNA-valine
tRNA-leucine
ARBP

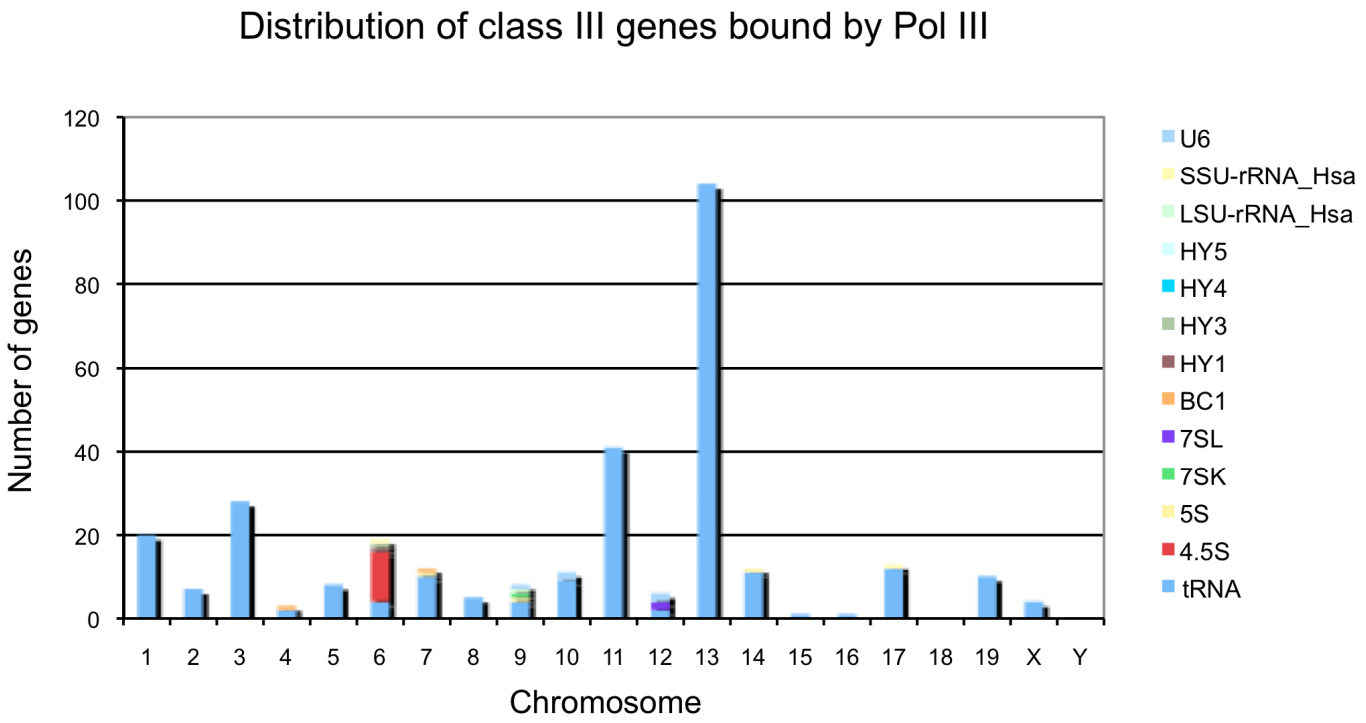
Carrière et al. Fig. S4



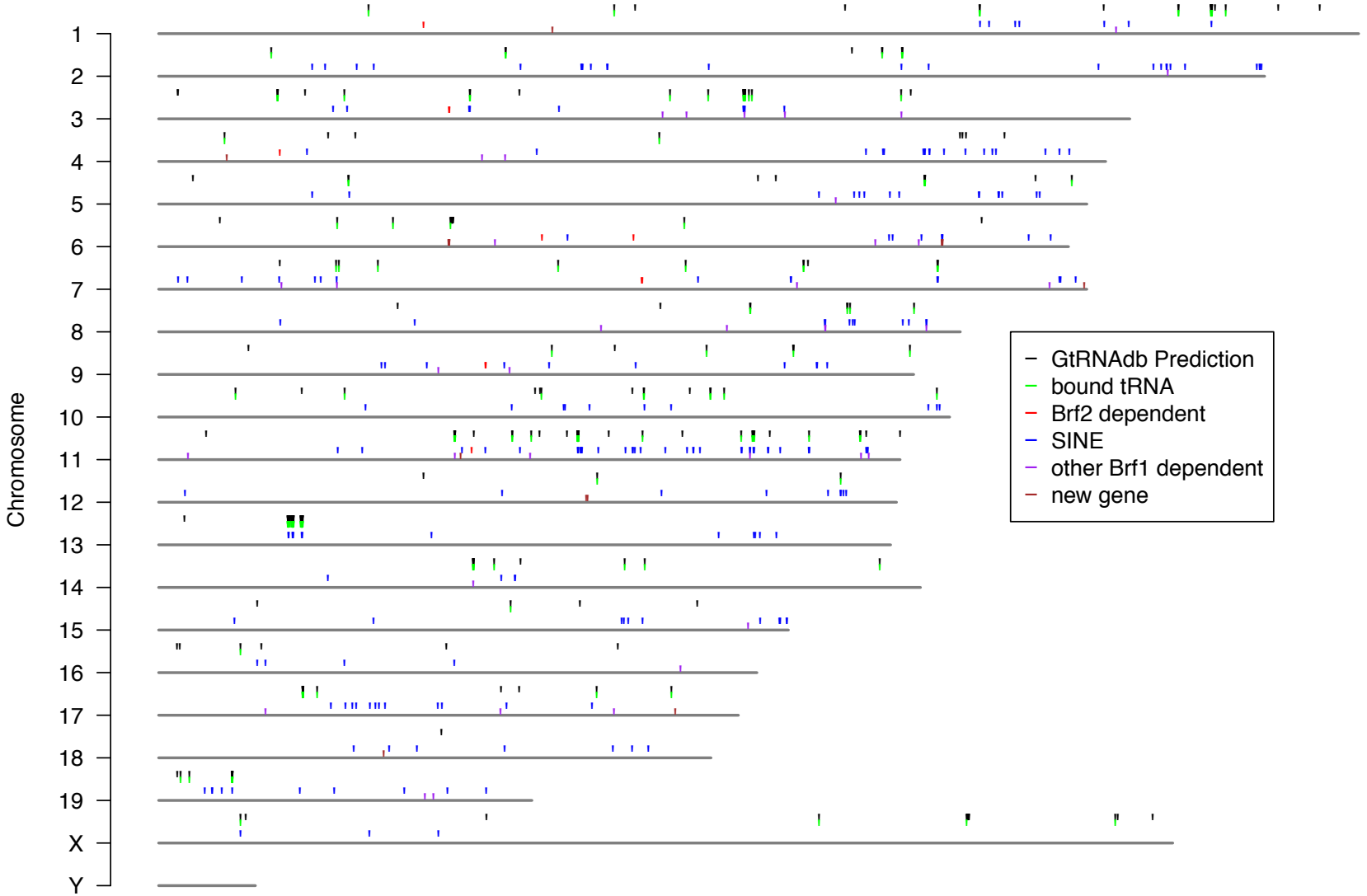
Carrière et al. Fig. S5



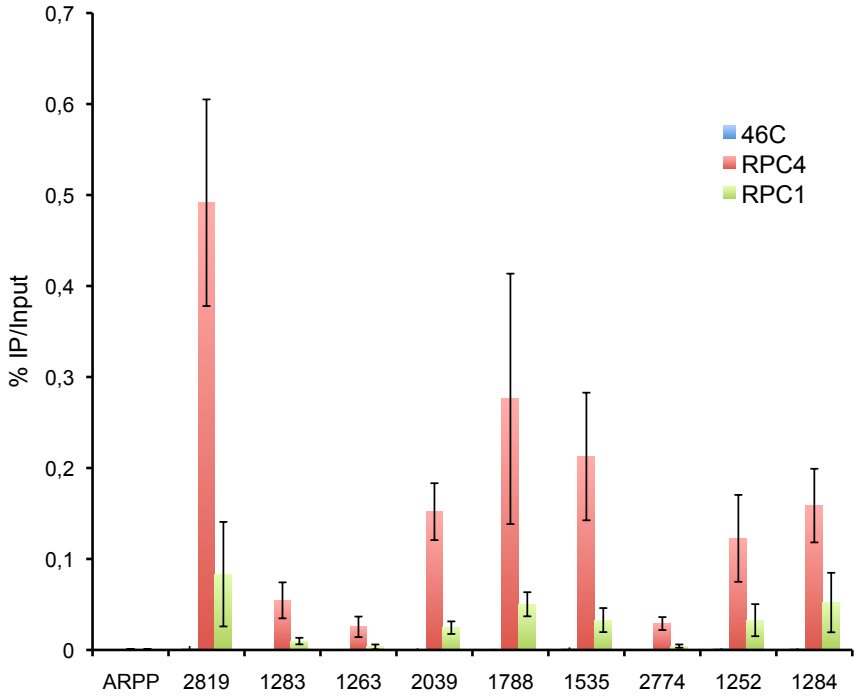
Carrière et al. Figure S6A



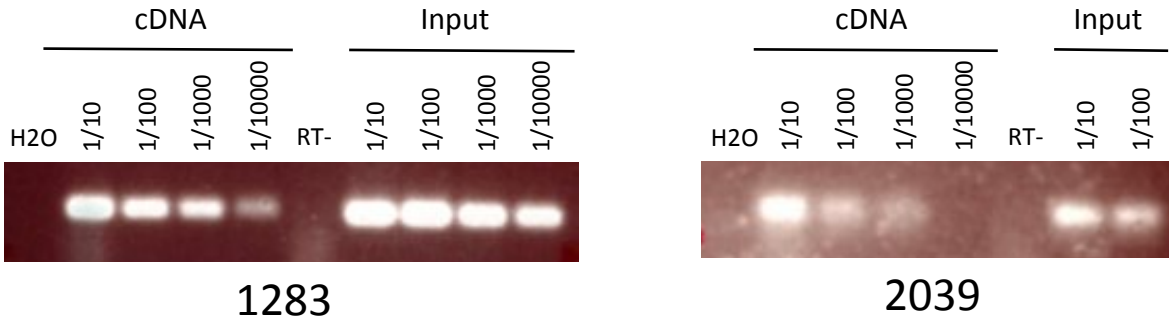
Carrière et al. Figure S6B



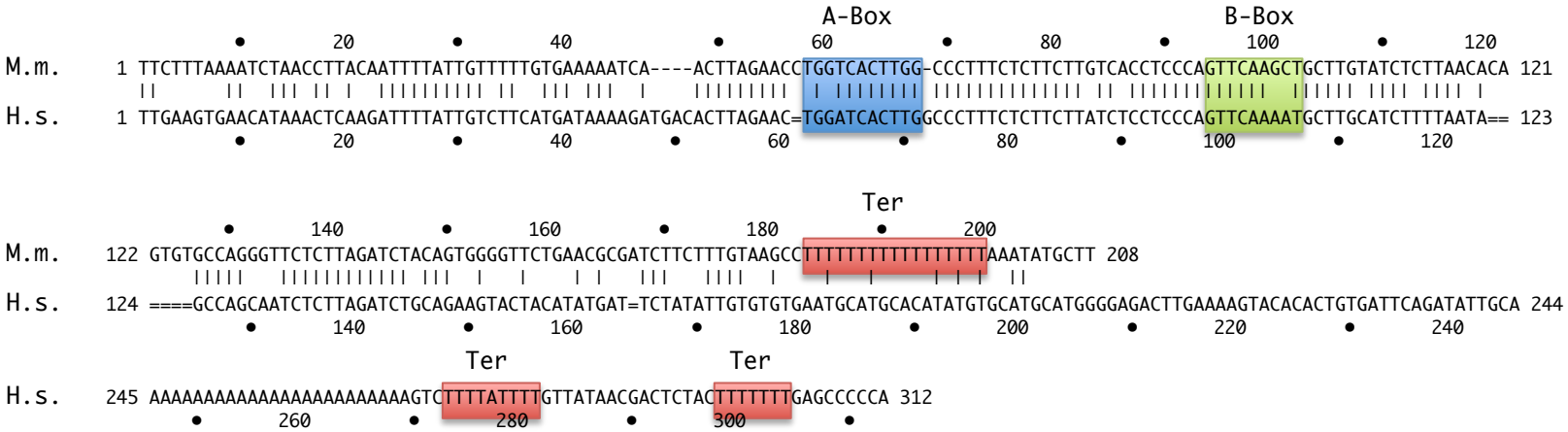
A


0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
% IP/Input
46C
RPC4
RPC1
ARPP
2819
1283
1263
2039
1788
1535
2774
1252
1284

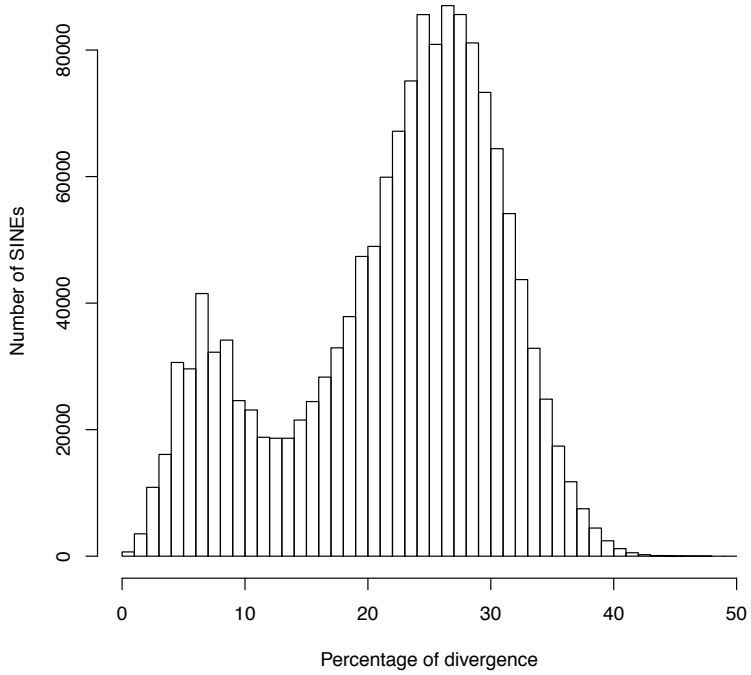
B


cDNA
Input
H2O
1/10
1/100
1/1000
1/10000
RT-
1/10
1/100
1/1000
1/10000
1283
cDNA
Input
H2O
1/10
1/100
1/1000
1/10000
RT-
1/10
1/100
2039

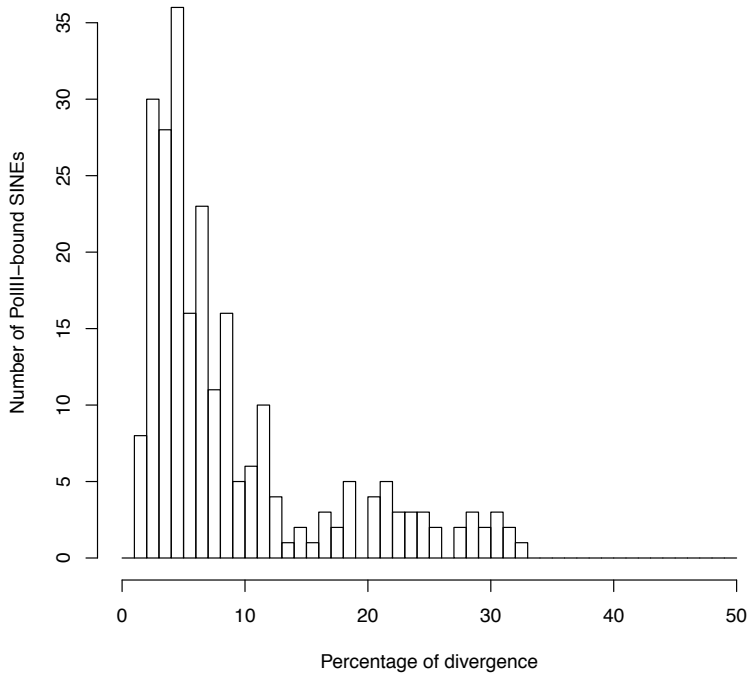
C

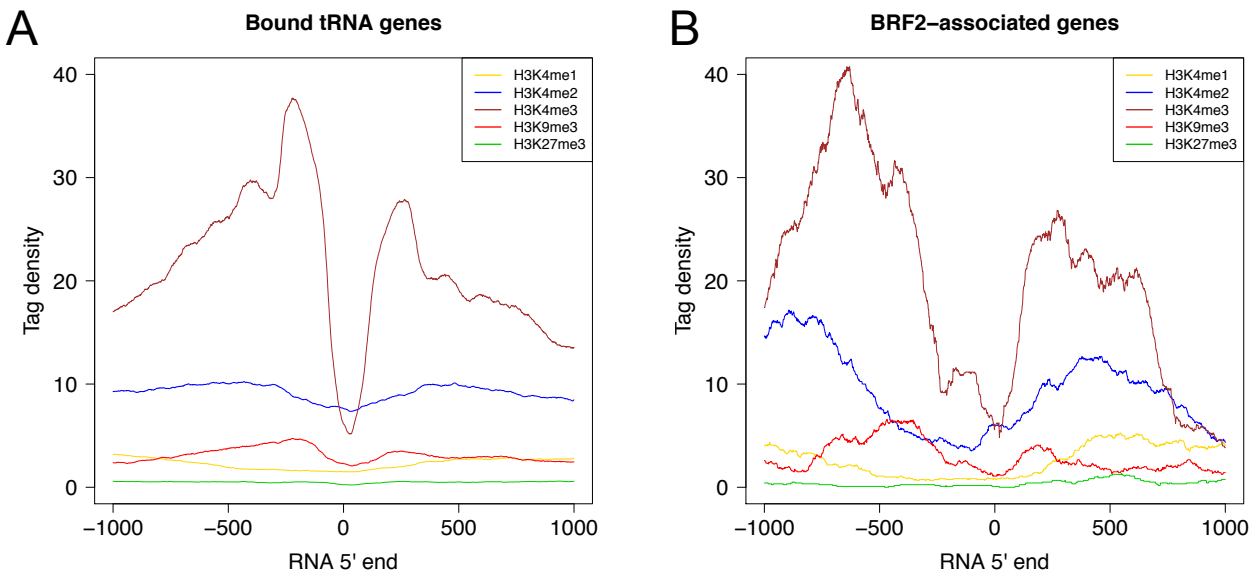


A

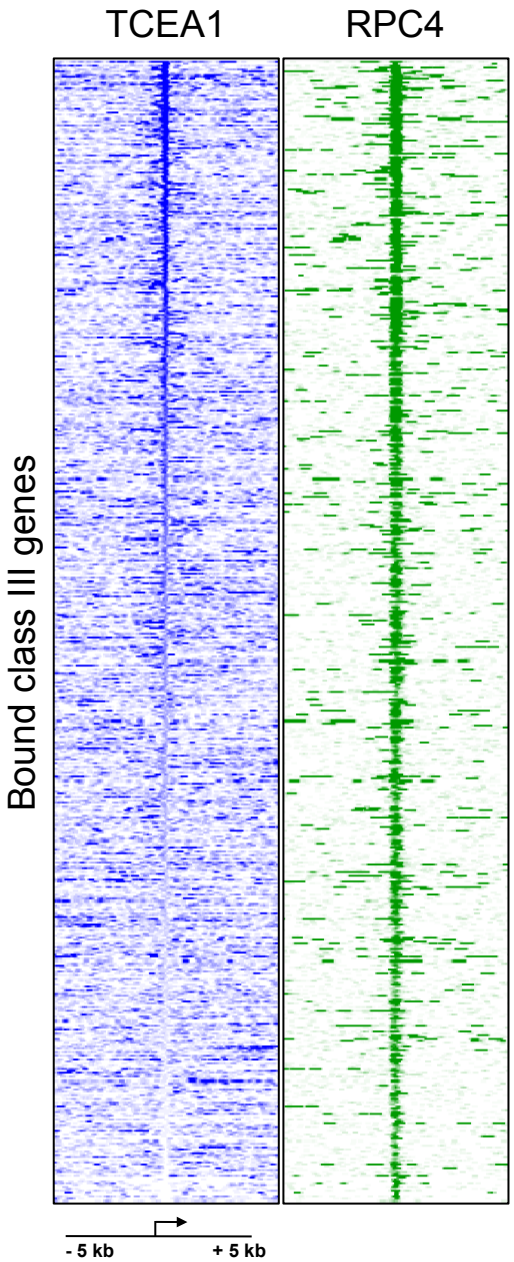

Number of SINEs
0
20000
40000
60000
80000
0
10
20
30
40
50
Percentage of divergence

B


Number of PolIII−bound SINEs
0
5
10
15
20
25
30
35
0
10
20
30
40
50
Percentage of divergence


A
Bound tRNA genes
B
BRF2–associated genes
H3K4me1
H3K4me2
H3K4me3
H3K9me3
H3K27me3
Tag density
RNA 5' end
0
10
20
30
40
−1000
−500
0
500
1000

Carrière et al. Fig. S10



DISCUSSION ET PERSPECTIVES

Génération d'outils

Au cours de cette étude, nous avons généré un nombre important d'outils. Plusieurs sous-unités de la RNAP III, et de ses facteurs de transcription ont été étiquetées. L'ajout d'étiquette nous a permis de développer des protocoles génériques d'immunoprécipitation de chromatine très spécifiques.

Le choix de travailler en cellules ES nous ouvre de larges horizons. Nous disposons d'un type cellulaire pouvant se différencier en un très grand nombre d'autres types. Ainsi, nous pourrons étudier les variations de distribution de nos protéines au cours de la différentiation, l'augmentation ou la diminution du nombre de cibles de la RNAP III. L'analyse précise des séquences, de l'environnement chromatinien, des facteurs de remodelage impliqués, pourrait nous indiquer quels mécanismes dirigent l'expression différentielle des gènes de classe III observée chez l'humain.

L'augmentation de la transcription de classe III est observée dans un grand nombre de cancer. De plus, BRF1 peut avoir une activité oncogénique. Différentes études soulignent l'importance de TFIIS au cours du développement ou dans les cellules cancéreuses (Hubbard et al., 2008; Ito et al., 2006). Nous pourrions étudier le rôle de ces protéines chez des souris, portant des mutations conduisant à l'apparition de tumeurs. Le laboratoire du Dr M. Gérard possède des lignées de souris transgéniques ApcMin (Min, multiple intestinal neoplasia). Cette mutation ponctuelle du gène APC provoque chez les souris l'apparition d'adénome intestinaux (Su et al., 1992). L'introduction de la mutation APCmin ou d'autres mutations affectant des oncogènes ou des suppresseurs de tumeurs permettra de mieux comprendre le rôle joué par la transcription de classe III dans le processus cancéreux.

Analyse des données

Le choix de ne retenir que les régions liées par les deux sous-unités de la RNAP III, comme l'inspection visuelle et répétée d'un grand nombre de ces régions, nous permet d'affirmer que les gènes identifiés sont réellement liés par la machinerie de transcription de classe III. Bien évidemment, l'efficacité d'immunoprécipitation de chromatine diffère selon ces deux sous-unités. Les sites liés par RPC4 sont plus nombreux que ceux liés par RPC1. Le fait de les écarter provient de la volonté d'établir une liste claire et juste, plutôt que d'entrer dans une simple description de régions potentiellement liées. Il est toujours possible d'explorer plus en détail ces régions, de déterminer si le seuil de détection des pics imposé pour RPC1 ne peut être revu à la baisse, et ainsi identifier de nouvelles régions liées. Je ne pense pas que nous identifions de nouveaux gènes d'ARN de transfert, car ceux-ci ont la caractéristique d'être très fortement liés. Il est, par contre possible d'identifier de nouveaux ARNs de classe III.

30 nouvelles régions ont été identifiées. La liaison de la machinerie de classe III a été confirmée au niveau de neuf d'entre elles, bien que l'intensité de liaison soit variable d'une région à l'autre. 16 transcrits ont pu être détectés. Cependant certains de ces nouveaux gènes sont localisés dans des introns de gènes de classe II. Il manque donc une confirmation claire de la dépendance de la RNAP III pour leur transcription. Nous pourrions tester cette dépendance par une expérience de transcription *in vitro*. *In vivo*, l'inhibition de la transcription de classe II par l'α-amanitine, ou de classe III par la tagetitoxine permettrait d'infirmer ou de confirmer la classe de ces transcrits.

La faible conservation de séquences de ces gènes entre la souris et l'homme peut être en leur défaveur, puisque la plupart des transcrits de classe III sont conservés entre les espèces. Il existe cependant des exceptions de gènes spécifiques comme les gènes BC1, ou BC200. De même, certains SINEs sont spécifiques d'une espèce, ou d'un genre. Un seul nouveau gène fait exception, il montre une forte similarité de séquences, ainsi que la conservation des éléments A et B, nécessaires à la transcription des gènes de classe III, celui-ci mériterait que l'on s'y attarde plus longuement.

Conservation des SINEs

80% des SINEs liés ont une séquence conservée, puisque leur pourcentage de divergence estimé par Repeat Masker est assez faible. Repeat masker utilise les séquences consensus des éléments répétés regroupés dans Repbase. Après avoir fixé un seuil de conservation, des limites de divergence de séquence, ce programme recherche tout ce qui peut ressembler de près ou de loin à un SINE. Cela explique peut-être pourquoi autant de SINEs sont prédits dans le génome, pourquoi certains ont gardé peu de similarité de séquence avec une séquence consensus SINE. Ces séquences sont peut-être des signatures d'anciens SINEs, ayant fortement divergés.

Les SINEs sont peut-être les gènes les plus difficilement identifiables, car ils présentent un niveau de liaison assez faible. Diminuer le seuil de détection aurait peut-être révélé d'autres SINEs. Cependant, de façon assez surprenante, il ne semble pas que la difficulté de cartographier ces régions ait été le critère le plus contraignant. Les régions flanquantes sont assez souvent suffisantes pour cartographier de manière unique les lectures. La transcription des SINEs varie en réponse à différents stimuli, comme l'infection virale, le choc thermique. Cette expression varie également au cours de l'embryogenèse. Nous pourrions alors déterminer au cours de la différentiation, ou dans au sein de différents tissus, l'expression des SINEs.

Expression différentielle des transcrits de classe III

L'établissement d'une carte précise de la localisation de la RNAP III et de ses facteurs de transcription a tout d'abord permis de souligner et confirmer les connaissances existantes de la transcription de classe III. Elle a en plus révélé des propriétés spécifiques des mammifères. A peine la moitié des gènes d'ARNt prédits sont liés dans les cellules ES de souris. Ceci contraste fortement avec la levure, où tous les gènes d'ARNt sont liés en phase exponentielle de croissance. L'expression des ARNt est soumise à de nombreuses conditions, comme la progression du cycle cellulaire ou la transformation oncogénique (Marshall and White, 2008). Les données de ChIP-seq réalisées chez l'humain montrent que le nombre de gènes d'ARNt occupés est inférieur dans les fibroblastes, ou des cellules T, comparé aux trois autres types cellulaires transformés. Cette variation est peut-être due à la comparaison entre des cellules saines et des cellules cancéreuses, expliquant l'augmentation du nombre de transcrits de classe III. Malgré tout, il était déjà connu que l'expression des gènes d'ARNt varie fortement d'un tissu à l'autre (Dittmar et al., 2006). Cette dernière étude ne disposait pas de moyens suffisants pour conclure si cette variation était au niveau de la production ou de la dégradation des ARNt. Les études de ChIP-seq confirment que les niveaux d'expression différentielle observés sont causés par une occupation différente des gènes d'ARNt. Cette observation est difficile à accommoder avec les modèles de régulation actuels. Les gènes de classe III partagent les mêmes facteurs de transcription, et les mêmes promoteurs. La régulation de la transcription via les facteurs de transcription est déjà bien étudiée. Les voies de signalisation comme la voie MAPK, mTORC contrôlent l'accès aux promoteurs des facteurs TFIIIB et TFIIIC. L'occupation ne résulte pas non plus d'une variation de séquence promotrice. Ceci suggère que d'autres mécanismes additionnels de régulation influence l'accès au promoteur, en dehors du promoteur. L'expression différentielle peut être régulée via des séquences adjacentes, où seraient recrutés des facteurs spécifiques au type cellulaire ou à l'état de la cellule. L'expression d'un facteur de transcription en réponse à la différentiation ou à la transformation de cellules est régulée spécifiquement. Si celui-ci possède des séquences lui permettant de se lier à des gènes de classe III, il pourra alors activer ou réprimer la transcription de ces gènes. Ces protéines peuvent aussi activer des facteurs de remodelage de chromatine.

L'étude des motifs présents dans les séquences adjacentes des gènes de classe III peut nous révéler les facteurs impliqués dans la régulation de la transcription. Nous pouvons également caractériser les facteurs de remodelage de chromatine intervenant pour créer un environnement favorable à la liaison de TFIIIC, et de TFIIIB.

La forte corrélation observée chez l'humain (Listerman et al., 2007; Raha et al., 2010), entre l'occupation des gènes par la RNAP III, et la proximité de la RNAP II suggère une interaction au niveau de la régulation, mais est insuffisante pour prouver une relation fonctionnelle. Les auteurs du ChIP-seq chez l'humain proposent que la liaison de la RNAP II crée un environnement permissif pour le recrutement des facteurs de transcription de classe III. Cette proximité entre la RNAP II et la RNAP III peut être une autre explication à l'expression différentielle des gènes de classe III. Existe-t'il un lien entre l'expression spécifique des gènes de classe II, et des gènes de classe III adjacents ?

Finalement pourquoi ne pas inverser ce lien de cause à effet ? Le recrutement de la machinerie de classe III, comme chez la levure, au niveau de certains gènes de classe III, bloquerait les effets répressifs de l'hétérochromatine. TFIIIC peut acétyler les histones créant une région ouverte de chromatine. Ainsi, l'expression des gènes de classe III créerait un environnement chromatinien non répressif pour la RNAP II. La dérégulation de la transcription de classe III ou de classe II est une approche difficilement maîtrisable. L'inhibition d'une des deux classes de transcription peut provoquer de nombreuses perturbations gênant l'interprétation des effets observés.

Cependant, dans notre étude, nous n'observons le recrutement de la RNAP II, qu'à proximité des gènes dépendants de BRF2. La liaison de la RNAP II aux côtés des gènes d'ARNt reste à confirmer chez la souris.

Rôle de TFIIIC

TFIIIC est capable de bloquer la propagation de l'hétérochromatine en l'absence de recrutement de TFIIIB et de la RNAP III (Noma et al., 2006; Simms et al., 2008; Valenzuela et al., 2009). Les sites liés par TFIIIC ont à l'origine été mis en évidence chez la levure *S. cerevisiae* (Harismendy et al., 2003; Moqtaderi and Struhl, 2004; Roberts et al., 2003). Leur fonction a été caractérisée chez *S. pombe* (Noma et al., 2006).

Dans différents types cellulaires humains, de nombreux sites liés uniquement par TFIIIC ont été identifiés. CTCF se lie à proximité d'un certain nombre de ces régions (Moqtaderi et al., 2010; Oler et al., 2010). Chez la souris, nous avons identifié de nombreux sites liés par TFIIIC, dépourvus du reste de la machinerie de classe III.

La conservation de ces sites entre la levure et les mammifères implique sûrement une conservation de fonctions. Cependant, pour établir si ces régions fonctionnent en tant qu'insulateur, plusieurs vérifications s'imposent.

Approches bioinformatiques.

Une approche bioinformatique pourrait nous permettre de déterminer la localisation de ces sites, c'est-à-dire, de déterminer s'il existe un lien entre la localisation des sites liés par TFIIIC et les gènes de classe II. Peut-on imaginer que les gènes de classe II adjacents aux ETC partagent une voie métabolique commune, ou dirigent un ensemble de processus communs ? Les sites ETC chez *S. cerevisiae* sont souvent situés entre des gènes divergents (Simms et al., 2008). Il en est de même chez l'humain (Moqtaderi et al., 2010). Un tel positionnement révèlerait une fonction de barrière séparant physiquement deux domaines de chromatine régulés différemment. Ceci implique que les gènes divergents appartiendraient chacun à une voie cellulaire clairement distincte, dont la fonction n'est pas requise au même moment.

Modèle d'inactivation de séquences insulatrices.

Pour étudier *in vivo* le rôle potentiel d'insulation de TFIIIC, nous pouvons utiliser une approche d'inactivation en cellules ES. L'analyse informatique nous a permis de dégager un certain nombre de régions, présentant les caractéristiques définies d'un site ETC (absence de TFIIIB et de la RNAP III). Ces séquences, ayant un rôle putatif de barrière, peuvent être délétées par recombinaison homologue, à un locus précis. Les vecteurs pour la recombinaison homologue pourraient être générés de manière efficace par recombineering (Liu et al., 2003), comme nous l'avons fait pour l'étiquetage des protéines de la machinerie de classe III.

La plupart des insulateurs sont des régions dépourvues de nucléosomes, situées dans des régions d'hétérochomatine présentant des modifications spéciales. Cet état est maintenu via le recrutement de facteurs de remodelage de la chromatine. Ces régions peuvent recruter des protéines établissant des interactions longue distance, regroupant les insulateurs entre eux, et pouvant s'ancrer dans des structures comme le nucléole, ou à la périphérie nucléaire. Si nous pouvons vérifier certaines de ces caractéristiques en condition sauvage, l'approche de délétion nous permettra de confirmer le rôle fonctionnel de ces sites.

Organisation nucléaire et fonction insulatrice.

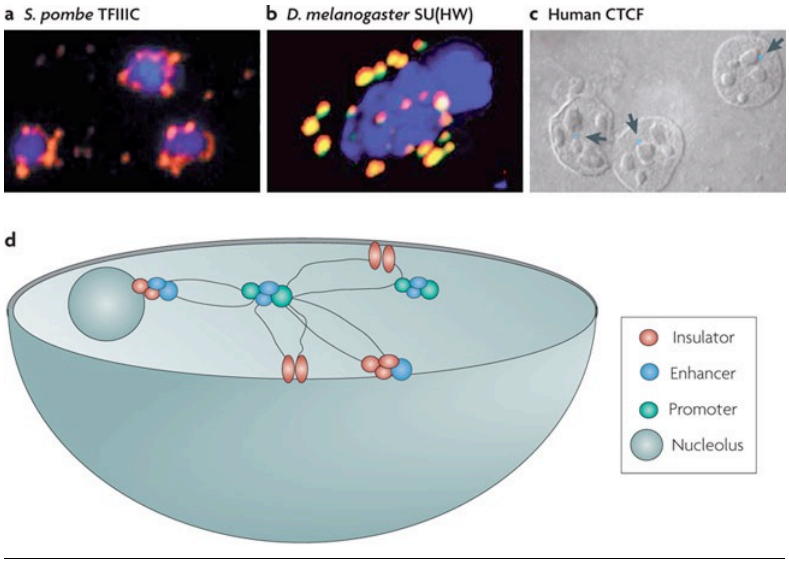


Figure 18. Les insulateurs et l'organisation tridimensionnelle du noyau (d'après Raab and Kamakaka, 2010).

Regroupement des sites liés par le facteur de transcription TFIIIC à la périphérie nucléaire chez *S. pombe* (a) (Noma et al., 2006), par la protéine Suppressor of Hairy Wings SU(HW) chez *D. melanogaster* (b) (Gerasimova et al., 2000), des sites liés par CTCF à proximité du nucléole dans les cellules humaines (c) (Yusufzai and Felsenfeld, 2004). (d) Modèle d'interaction entre les insulateurs, les promoteurs, et les structures nucléaires du noyau. Les insulateurs interagissent ensemble et avec d'autres éléments régulateurs, séparant physiquement des domaines structurels et fonctionnels (Gerasimova and Corces, 2001).

Le complexe SMC participe à la démarcation des domaines de chromatine, permettant d'organiser ces domaines dans un espace tridimensionnel dans le noyau. L'existence d'un rôle de la cohésine indépendant de la cohésion des chromatides a été illustrée chez la levure au site *HMR*. Le recrutement de la condensine (protéines SMC2 et 4) par TFIIIC au niveau de ces sites de liaison pourrait faciliter le recrutement des gènes d'ARNt au nucléole, en établissant ou en maintenant des interactions inter-chromosomiques entre les régions liées par la RNAP III. De telles interactions peuvent être générées par un seul complexe liant tous les sites, ou par l'agrégation de plusieurs complexes en une même région.

Bien que la signification de l'organisation nucléaire ne soit pas claire, elle peut permettre de regrouper les insulateurs dans un microenvironnement riche en facteurs spécifiques. Des expériences 3C suggèrent que ces éléments barrières au site *HMR* interagissent ensemble pour former une boucle, qui

contient le locus *MAT* (Valenzuela et al., 2008). Mais l'implication de la cohésine dans la stabilisation de cette structure n'a pas été éclaircie. Chez *S. pombe,* les sites ETC ou COC sont impliqués dans l'organisation de l'architecture nucléaire. Ces résultats suggèrent que l'activité insulatrice est dépendante de l'architecture chromosomique.

Deux types d'interactions peuvent donc favoriser l'action des insulateurs : l'agrégation de ces régions entre elles, et/ou l'ancrage à une structure, comme le nucléole ou la périphérie nucléaire.

En menant des expériences d'immunofluorescences avec un anticorps dirigé contre un des épitopes codé par l'étiquette d'une des sous-unités de TFIIIC, nous pourrions déterminer la localisation du complexe TFIIIC au sein du noyau. La technique de Fluorescence In situ Hybridization (FISH) a été utilisée pour analyser la localisation sous-nucléaire des loci génomiques, comme les sites ETC. Le développement de nouvelles approches comme le 3C (Chromosome Conformation Capture) couplé au séquençage haut-débit (Hi-seq, (Lieberman-Aiden et al., 2009; Tanizawa et al., 2010)) permet d'analyser les associations physiques entre ces différents loci génomiques, à l'échelle du génome, sans idée *a priori*. Ces expériences peuvent être menées après mutation d'une région ETC. La localisation de cette région au sein du noyau serait alors analysée.

Les sites liés par la cohésine coïncident souvent avec les sites liés par la protéine CCCTC (CTCF, CCCTC-binding factor), chez l'humain (Parelho et al., 2008; Rubio et al., 2008; Wendt and Peters, 2009). CTCF est impliqué l'organisation du génome, en particulier au travers de sa fonction insulatrice. La cohésine contribuerait à la fonction de CTCF par la formation de boucle ADN (Wood et al., 2010). CTCF pourrait servir à positionner la cohésine, une fois qu'elle a été chargée (Phillips and Corces, 2009; Wendt and Peters, 2009). Bien que la majorité des sites occupés par CTCF, chez les mammifères, le soit également par la cohésine, une fraction des régions liées par la cohésine le sont indépendamment de CTCF (Kagey et al., 2010).

Bien que nous ayons établi une association jusqu'à 20 kb autour des sites TFIIIC, ce qui peut sembler bien éloigné pour une action commune, deux autres études de ChIP-seq observent également cette association (Moqtaderi et al., 2010; Oler et al., 2010). Il faudrait déterminer premièrement par des expériences de ChIP-qPCR, si ces sites sont réellement liés par CTCF. Il est toujours possible d'imaginer que TFIIIC lié aux boîtes B grâce à un environnement chromatinien particulier, est capable de recruter CTCF. L'interaction avec la cohésine permettrait d'établir à son tour des boucles isolant physiquement des domaines de chromatine.

Leur ancrage à la périphérie nucléaire permet également de créer une certaine organisation sous-nucléaire. TFIIIC interagit peut-être avec des protéines nucléaires comme la lamine, ainsi qu'il a été observé pour CTCF (Guelen et al., 2008).

Les insulateurs affectent la structure locale de la chromatine.

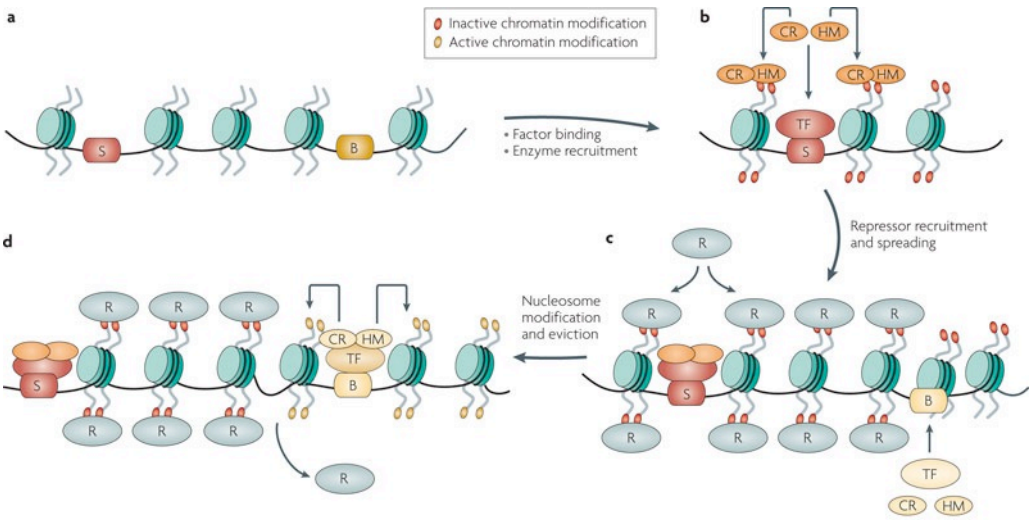


Figure 19. Établissement de l'activité de barrière (d'après Raab and Kamakaka, 2010).

S : Silencer, R : Répresseur, TF : Facteur de transcription, CR Remodeleur de Chromatine, B : élément barrière, HM : Enzymes de modification des histones.

(a, b) : les éléments silencers (S) recrutent des facteurs de transcription (TF) spécifiques. Ceux-ci, à leur tour, recrutent des remodeleurs de chromatine (CR), et des enzymes de modification des histones (HM). Ces facteurs coopèrent en modifiant la chromatine, créant ainsi des sites favorables à la liaison des protéines de répression (R).

(c, d) : l'établissement et la propagation des répresseurs le long de la fibre nucléosomale résulte en l'établissement d'un domaine d'hétérochromatine. Aux éléments barrières (B) sont recrutés un ensemble distinct de facteurs de transcription. Ceux-ci recrutent des complexes enzymatiques, qui modifient les histones, établissant des marques « actives », et déplacent les nucléosomes. Ceci crée une discontinuité dans la chaîne nucléosomale, bloquant la propagation de l'hétérochromatine.

Remodeleurs de chromatine et activité histone acétylase

La localisation particulière des sites ETC au sein du noyau ne prouve pas leur fonction insulatrice. Pour définir cette activité, il faut étudier l'influence de tels sites sur l'environnement chromatinien, et

l'expression des gènes voisins. Les éléments barrières fonctionnent en créant des régions de chromatine réfractaires à la liaison et à la propagation des complexes de répression.

Des études de localisation des nucléosomes montrent que les gènes d'ARNt sont dépourvus d'histone au niveau du corps du gène (Parnell et al., 2008; Pokholok et al., 2005). Des observations récentes montrent de plus que l'éviction des histones autour des gènes d'ARNt est réduite dans un mutant RSC (Parnell et al., 2008). De plus, une étude montre que RSC interagit avec la machinerie de transcription de classe III (Soutourina et al., 2006). Concernant les sites ETC, la fonction barrière de TFIIIC nécessite le recrutement de facteurs de remodelage de chromatine et d'histone acétylase (Valenzuela et al., 2009). Des mutations d'histones acétylase comme Sas2, Eaf3 (ESA associated factor 3), ou de remodeleurs de chromatine, Rsc2 Isw2 ISWI (chromatin-remodeling complex ATPase, ou Imitation Switch protein 2) affectent directement ou indirectement la fonction insulatrice d'un site ETC chez *S. cerevisiae*. Ces diverses données suggèrent que l'activité insulatrice de TFIIIC fonctionne via le recrutement de facteur de remodelage de la chromatine, couplé à des modifications d'histones.

Le ChIP-seq pour l'ensemble de la famille des facteurs de remodelage de la chromatine, à chromodomaine CHD (Chromodomain-helicase DNA binding Domain) et des membres de la famille SWI/SNF a été effectué dans le laboratoire de Matthieu Gérard. Un membre de la famille des CHD, CHD8 a déjà été mis en relation avec la transcription de classe III. Il est requis pour la transcription efficace du gène U6 (Yuan et al., 2007). La transcription de ce gène, dont le promoteur est de type III, ne requiert pas TFIIIC. Mais cette étude souligne tout de même que l'environnement chromatinien doit être modifié pour permettre l'accessibilité à la machinerie de classe III. Si CHD8 est requis au promoteur du gène U6, on peut tout à fait imaginer qu'il en est de même pour les autres types de promoteurs ainsi que spécifiquement au niveau des sites ETC. Nous pourrions alors mener des analyses afin de déterminer si les facteurs de remodelage sont recrutés au niveau des sites ETC.

Si une telle étude nous conduisait à l'identification de facteurs de remodelage, l'inhibition de leur activité par le mécanisme d'ARN interférence, en utilisant un shARN (small hairpin RNA) pourrait interférer avec l'activité insulatrice d'un site ETC.

Histones

Nous avons observé (données non présentées), comme Moqtadery et Oler, que les sites liés par TFIIIC présentent les modifications d'histones suivantes : H3K4me1 et H3K27ac marquant spécifiquement les enhancers actifs (Creyghton et al., 2010). TFIIIC peut fonctionner en tant qu'enhancer-blocker. Ce type d'insulateur fonctionne en interagissant avec les enhancers. Ces interactions pourraient expliquer le profil de modification d'histones que nous observons aux ETC. Ces insulateurs interagissent

également les uns avec les autres. Ce regroupement dans le noyau forme des « corps d'insulateurs », isolant les enhancers des promoteurs, bloquant ainsi leur fonction activatrice (Raab and Kamakaka, 2010)

Approches génétiques

Chez l'humain, les sites ETC varient d'un type cellulaire à un autre (Moqtaderi et al., 2010). Nos cellules peuvent se différencier en de nombreux autres types cellulaires. Il serait intéressant de vérifier si les sites ETC varient au cours de la différentiation cellulaire, et si l'on peut lier ces variations avec le profil des gènes exprimés spécifiquement. De plus, ces variations confirment que TFIIIC n'est pas le seul facteur de recrutement déterminant. D'autres facteurs liés au type cellulaire ou au stade de développement doivent être soit recruter aux loci des gènes de classe III, comme des boîtes B, ou interagir avec les facteurs de transcription de classe III, modulant leur accès au génome. Dans le cas des ETC, il est très intéressant de noter que cette modulation s'effectue au travers de TFIIIC. La plupart des études traitant de la régulation de la transcription de classe III ont mis en évidence un mode de régulation s'effectuant principalement au travers de TFIIIB. Il s'agit maintenant de se focaliser sur les facteurs pouvant interagir avec TFIIIC.

TFIIS

TFIIS et la transcription de classe III

Dans cette étude, nous avons analysé la distribution de TCEA1, sur l'ensemble du génome par ChIP-seq. Cette étude fait suite à un travail réalisé chez *S. cerevisiae*, dans notre laboratoire par Yad Ghavi-Helm (Ghavi-Helm et al., 2008). L'analyse des sites liés par TFIIS par ChIP-chip révélait l'occupation inattendue de l'ensemble des gènes de classe III. Les mécanismes de base du fonctionnement cellulaire de la levure trouvent très souvent leur équivalence chez les mammifères. Nous avons voulu savoir si TFIIS était également un facteur de transcription de classe III chez les mammifères. Trois isoformes de TFIIS existent chez les mammifères. TCEA1 est l'isoforme exprimée de façon ubiquitaire. Nous avons pu observer que TCEA1 est présent sur la plupart des gènes liés par la RNAP III, à des taux variables selon la classe du gène.

Chez la levure, des études génétiques montrent que la présence de TFIIS est indépendante de la RNAP II, mais dépend de la transcription par la RNAP III. Nous avons caractérisé la distribution de la RNAP II chez la souris, au loci des gènes de classe III. Pour les gènes de promoteurs de type II, nous n'observons pas d'enrichissement important des différentes formes de la RNAP II. Ceci suggère que TCEA1 est recruté de façon indépendante de la RNAP II. Autour des gènes dépendants de BRF2, la distribution de TCEA1 est distincte de la distribution de la RNAP II. Cependant, nous n'apportons pas

d'arguments génétiques comme chez la levure. L'inhibition spécifique de la transcription de classe III par la tagétitoxine peut par exemple être utilisé pour déterminer si chez les mammifères, le recrutement de TFIIS sur les gènes de classe III, dépend de la transcription de classe III.

Chez la levure, la délétion du gène codant TFIIS ne semble pas avoir d'impact sur la croissance de la cellule, sauf à des conditions restrictives de température. Chez les mammifères, l'importance de TFIIS au cours du développement et de la croissance des organismes n'est pas très claire. Nous ne connaissons pas en cellules ES l'impact qu'aurait l'inhibition de l'expression des isoformes de TFIIS. Pourtant, nous pourrions étudier l'impact de la dérégulation de l'expression de TFIIS par shARN sur le recrutement de la machinerie de classe III, et l'expression des gènes de classe III.

Des études biochimiques chez *S. cerevisiae* démontrent le rôle de TFIIS dans la sélection du site d'initiation des gènes de classe III, *in vitro*. *In vivo*, ces expériences n'ont pas permis de conclure quant à ce rôle, suggérant que d'autres protéines joueraient ce rôle. Chez la souris, le profil de densité de TCEA1 ne suit pas tout à fait celui de la RNAP III. Le maximum de densité de TCEA1 est plutôt localisé en 5' des gènes. Cette observation pourrait être une indication d'un rôle éventuel de TCEA1 dans l'initiation de la transcription par la RNAP III chez la souris. Les nouvelles approches de séquençage haut-débit permettant de cartographier précisément l'extrémité 5' des ARNs (Ozsolak and Milos, 2011) couplées à l'inhibition de l'expression de TFIIS chez la souris, par shARN peuvent être mises en œuvre afin de vérifier le rôle de TFIIS dans l'initiation de la transcription. Cependant, deux obstacles majeurs doivent être considérés. Les transcrits de classe III ne sont pas toujours cartographiables individuellement du fait de leur répétition. De plus, les expériences *in vivo* chez la levure n'ont pas permis de conclure à un rôle dans l'initiation de classe III. La présence des trois isoformes de TFIIS pourrait nous empêcher d'observer ce rôle dans les cellules ES de souris.

TFIIS et la transcription de classe II

Deux modèles de recrutement de TFIIS, au cours de la transcription des gènes de classe II, existent. TFIIS peut être recruté uniquement lors des arrêts de la RNAP II, ou resterait associé au cours de l'élongation, indépendamment des arrêts de transcription, induisant un changement de conformation de la RNAP II lors d'un blocage (Ghavi-Helm et al., 2008). Chez la levure, le second modèle rend probablement mieux compte de la réalité, car l'occupation des gènes par la RNAP II et TFIIS est extrêmement bien corrélée. D'autre part, l'inhibition de l'élongation par des drogues n'induit pas de fortes variations de l'occupation de TFIIS aux gènes. Chez la souris, la combinaison des données de TCEA1 et de la RNAP II montre qu'il existe également une bonne corrélation d'occupation des gènes. Le profil de la distribution de ces deux protéines sur l'ensemble des gènes est tout à fait similaire. Nos observations en accord avec celles de la levure orientent vers le deuxième modèle de recrutement. Cependant, une étude

récente indique que les pauses au cours de l'élongation sont fréquentes (Churchman and Weissman, 2011). Les profils de distribution très similaires que nous observons reflètent peut-être un état moyen. La RNAP II est régulièrement bloquée, TFIIS est recruté au niveau de chaque pause, et nous observons alors une très bonne corrélation d'occupation de la RNAP II et de TFIIS. Les études de ChIP ne permettent pas une résolution suffisante pour discriminer quel modèle reflète le mieux la réalité, puisqu'elles nous donnent une image figée des complexes.

Le contrôle précis de la transcription est critique pour la régulation de l'expression des gènes, et ainsi la différentiation, le développement et la survie de la cellule. Jusqu'il y a peu, la régulation de l'initiation de la transcription était considérée comme le point de contrôle majeur de l'expression des gènes. La régulation post-recrutement de la RNAP II est aujourd'hui considérée comme un autre mécanisme fondamental contrôlant ce processus. La modulation de la transition du passage de la pause au promoteur à l'élongation processive implique différents facteurs d'élongation comme PTEF-b. Au promoteur proximal, la RNAP II initie la transcription puis pour un très grand nombre de gènes, s'arrête après avoir transcrit 100 à 200 nucléotides. Cette pause est suivie du recul de la RNAP II, délogeant l'extrémité 3' du transcrit du site catalytique. L'activité de stimulation du clivage propre au facteur d'élongation TFIIS est alors sûrement requise pour permettre au complexe ternaire ADN-RNAP II-ARN d'être de nouveau dans une conformation propre à reprendre l'élongation. Nos résultats démontrent que TFIIS est présent dans la région proximale, aussi bien aux gènes pausés et actifs, qu'aux gènes pausés et non-productifs. Ceci semble indiquer qu'au niveau du recrutement à proprement parler, l'on ne peut différencier les gènes actifs des non-productifs. Le recrutement de TFIIS n'est pas l'événement responsable de la transition de la pause à l'élongation. Cependant la présence de TFIIS au promoteur suggère que le recrutement de TFIIS, s'il n'est pas suffisant, est nécessaire au passage de cette pause.

Le ChIP-seq permet de détecter la présence de la RNAP II liée à l'ADN. Pourtant, cette technique n'est pas très précise pour distinguer la RNAP II présente dans les complexes de pré-initiation, pausée ou engagée en transcription. Afin d'étudier plus en détail le rôle de TFIIS au cours de ces premières étapes de la transcription, nous pourrions combiner une approche de ChIP-seq permettant de déterminer l'occupation de la RNAP II, et de GRO-seq ou Global run-on sequencing (Core et al., 2008). Cette approche pourrait être mise en œuvre dans les cellules sauvages, et celles où l'expression de TFIIS serait diminuée par ARN interférence.

Rahl et al. ont utilisé le ChIP-seq pour déterminer que c-Myc régule la transition de la pause en élongation (Rahl et al., 2010). Pour caractériser l'occupation de la RNAP II, ils ont appliqué le calcul du Traveling Ratio, développé par Zeiltlinger chez la drosophile (Zeitlinger et al., 2007). Le TR est le ratio de la densité de la RNAP II dans la région du promoteur proximal, sur la densité de la RNAP II sur le corps

du gène (voir Matériel et Méthodes). L'inhition de l'expression de c-Myc par une drogue entraîne une augmentation de ce TR pour les cibles de c-Myc. Cette augmentation reflète la diminution de la densité de la RNAP II sur le corps du gène, plus que la diminution de l'occupation de la RNAP II au promoteur. Concernant TFIIS, ce type d'approche peut nous permettre de mesurer son implication lors de cette transition.

Le GRO-seq permet de quantifier au niveau activité transcriptionnelle la RNAP II engagée en transcription. Cette technique évalue directement les RNAP II pausées au promoteur. Le « pausing index » peut également être calculé. Il correspond au TR, au niveau transcriptionnelle. Le ratio de la densité des transcrits au promoteur comparé à la densité des transcrits au corps du gène permet de définir les gènes où la RNAP II est pausée au promoteur. Si TFIIS est nécessaire à la transition, alors ce ratio devrait augmenter en réponse à la dérégulation de l'activité de TFIIS.

CONCLUSION

Cette étude nous a conduit à confirmer *in vivo*, les principales caractéristiques de la transcription de classe III qui ont été découvertes par le concours de nombreux laboratoires, chez des organismes variés. Le transcriptome chez la souris est défini. Cependant, notre étude, comme celles menées chez l'humain montre que seul un sous-ensemble de gènes d'ARNt est transcrit, suggérant que ce transcriptome n'est pas figé, et qu'il nous reste à déterminer ce qui sous-tend cette expression différentielle. La liaison distincte et mutuellement exclusive de BRF1 et BRF2 est clairement établie. L'ensemble des gènes dont la transcription est dépendante de BRF2 est décrit. Nous n'avons pas identifié de nouvelles unités de transcription liées par TFIIIBα. Mais comme pour les gènes d'ARNt, la définition de nouveaux transcrits peut être dépendante du stade de développement ou de la spécificité cellulaire. De nouveaux gènes ont été identifiés, certains sont transcrits. L'environnement chromatinien établi autour des gènes de classe III est très similaire à celui caractéristique des gènes de classe II. Enfin, la liaison « extra » de TFIIIC et le recrutement du facteur TFIIS à l'ensemble des gènes de classe III ont bien été retrouvés chez la souris, confirmant la conservation des mécanismes généraux de la levure aux mammifères. Les sites ETC présente une proximité avec la cohésine, suggérant que ces sites, suite à la liaison de TFIIIC ont un rôle dans l'établissement d'une certaine architecture nucléaire. Le facteur TFIIS est bien recruté aux gènes de classe III actifs, confirmant son rôle de facteur de transcription de classe III. Aux gènes de classe II, le recrutement de TFIIS est un facteur nécessaire, pourtant peut-être pas suffisant pour provoquer la transition de la pause de la RNAP II au promoteur, en élongation.

Cette étude constitue une suite aux travaux effectués chez la levure dans notre laboratoire, et d'autres, autour. Elle transpose, en utilisant les dernières techniques disponibles, s'adaptant parfaitement aux génomes des mammifères, les études de ChIP-chip sur la machinerie de transcription, par Harismendy, et de la localisation de TFIIS, par Ghavi-Helm. Nous aurons pu, de plus, développer des méthodes bio-informatiques d'analyse des données de ChIP-seq, qui, seront utiles à un autre projet de ChIP-seq en cours au laboratoire, chez le rat.

Le ChIP-seq ne constitue en aucun cas une fin en soi. Cette technique devrait au contraire se situer en amont de toute étude cherchant à caractériser la fonction d'un facteur *in vivo*. Bien loin de répondre clairement aux questions, cette technique génère, en ouvrant tout grand les portes du génome, plus d'interrogations que celles auxquelles nous pouvons pour le moment répondre. Elle illustre tout de même un grand principe, mis en œuvre dans notre laboratoire, qui est la conservation des mécanismes de la levure à l'homme. Ce petit organisme peut encore mener à de belles découvertes, et constitue toujours un organisme de choix pour les études de génétique.

De plus, ici, elle donne un nouveau souffle aux études concernant la transcription de classe III, et pour reprendre le titre d'une revue récente : « Transcription by RNA polymerase III : more complex than we thought » (White, 2011) !

BIBLIOGRAPHIE

Adelman, K., Marr, M. T., Werner, J., Saunders, A., Ni, Z., Andrulis, E. D., and Lis, J. T. (2005). Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. Mol Cell *17*, 103-112.

Akhtar, M. S., Heidemann, M., Tietjen, J. R., Zhang, D. W., Chapman, R. D., Eick, D., and Ansari, A. Z. (2009). TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. Mol Cell *34*, 387-393.

Amouyal, M. (2010). Gene insulation. Part I: natural strategies in yeast and Drosophila. Biochem Cell Biol *88*, 875-884.

Armache, K. J., Kettenberger, H., and Cramer, P. (2003). Architecture of initiation-competent 12-subunit RNA polymerase II. Proc Natl Acad Sci U S A *100*, 6964-6968.

Armache, K. J., Mitterweger, S., Meinhart, A., and Cramer, P. (2005). Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. J Biol Chem *280*, 7131-7134.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.

Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A. B., Birch, J., Cui, K., White, R. J., and Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. Nat Struct Mol Biol *17*, 629-634.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Bernstein, B. E., Meissner, A., and Lander, E. S. (2007). The mammalian epigenome. Cell *128*, 669-681.

Bertrand, E., Houser-Scott, F., Kendall, A., Singer, R. H., and Engelke, D. R. (1998). Nucleolar localization of early tRNA processing. Genes Dev *12*, 2463-2468.

Bhargava, P., and Kassavetis, G. A. (1999). Abortive initiation by Saccharomyces cerevisiae RNA polymerase III. J Biol Chem *274*, 26550-26556.

Biswas, M., Maqani, N., Rai, R., Kumaran, S. P., Iyer, K. R., Sendinc, E., Smith, J. S., and Laloraya, S. (2009). Limiting the extent of the RDN1 heterochromatin domain by a silencing barrier and Sir2 protein levels in Saccharomyces cerevisiae. Mol Cell Biol *29*, 2889-2898.

Borchert, G. M., Lanier, W., and Davidson, B. L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol *13*, 1097-1101.

Bortolin-Cavaille, M. L., Dance, M., Weber, M., and Cavaille, J. (2009). C19MC microRNAs are processed from introns of large Pol-II, non-protein-coding transcripts. Nucleic Acids Res *37*, 3464-3473.

Boyle, A. P., Guinney, J., Crawford, G. E., and Furey, T. S. (2008). F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics *24*, 2537-2538.

Brueckner, F., Armache, K. J., Cheung, A., Damsma, G. E., Kettenberger, H., Lehmann, E., Sydow, J., and Cramer, P. (2009). Structure-function studies of the RNA polymerase II elongation complex. Acta Crystallogr D Biol Crystallogr *65*, 112-120.

Brun, I., Sentenac, A., and Werner, M. (1997). Dual role of the C34 subunit of RNA polymerase III in transcription initiation. Embo J *16*, 5730-5741.

Buratowski, S. (2005). Connections between mRNA 3' end processing and transcription termination. Curr Opin Cell Biol *17*, 257-261.

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Mol Cell *36*, 541-546.

Burnol, A. F., Margottin, F., Huet, J., Almouzni, G., Prioleau, M. N., Mechali, M., and Sentenac, A. (1993). TFIIIC relieves repression of U6 snRNA transcription by chromatin. Nature *362*, 475-477.

Bushnell, D. A., and Kornberg, R. D. (2003). Complete, 12-subunit RNA polymerase II at 4.1-A resolution: implications for the initiation of transcription. Proc Natl Acad Sci U S A *100*, 6969-6973.

Cairns, C. A., and White, R. J. (1998). p53 is a general repressor of RNA polymerase III transcription. Embo J *17*, 3112-3123.

Carter, R., and Drouin, G. (2009). Structural differentiation of the three eukaryotic RNA polymerases. Genomics *94*, 388-396.

Carter, R., and Drouin, G. (2010). The increase in the number of subunits in eukaryotic RNA polymerase III relative to RNA polymerase II is due to the permanent recruitment of general transcription factors. Mol Biol Evol *27*, 1035-1043.

Chedin, S., Riva, M., Schultz, P., Sentenac, A., and Carles, C. (1998). The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIS-like subunit and is important for transcription termination. Genes Dev *12*, 3857-3871.

Cheng, B., and Price, D. H. (2007). Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. J Biol Chem *282*, 21901-21912.

Cheung, A. C., and Cramer, P. (2011). Structural basis of RNA polymerase II backtracking, arrest and reactivation. Nature *471*, 249-253.

Christov, C. P., Gardiner, T. J., Szuts, D., and Krude, T. (2006). Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. Mol Cell Biol *26*, 6993-7004.

Churchman, L. S., and Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368-373.

Cler, E., Papai, G., Schultz, P., and Davidson, I. (2009). Recent advances in understanding the structure and function of general transcription factor TFIID. Cell Mol Life Sci *66*, 2123-2134.

Cojocaru, M., Bouchard, A., Cloutier, P., Cooper, J. J., Varzavand, K., Price, D. H., and Coulombe, B. (2011). Transcription factor IIS cooperates with the E3 ligase UBR5 to ubiquitinate the CDK9 subunit of the positive transcription elongation factor B. J Biol Chem *286*, 5012-5022.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Coughlin, D. J., Babak, T., Nihranz, C., Hughes, T. R., and Engelke, D. R. (2009). Prediction and verification of mouse tRNA gene families. RNA Biol *6*, 195-202.

Cramer, P., Bushnell, D. A., Fu, J., Gnatt, A. L., Maier-Davis, B., Thompson, N. E., Burgess, R. R., Edwards, A. M., David, P. R., and Kornberg, R. D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. Science *288*, 640-649.

Cramer, P., Bushnell, D. A., and Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science *292*, 1863-1876.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., *et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931-21936.

D'Alessio, J. A., Wright, K. J., and Tjian, R. (2009). Shifting players and paradigms in cell-specific transcription. Mol Cell *36*, 924-931.

D'Ambrosio, C., Schmidt, C. K., Katou, Y., Kelly, G., Itoh, T., Shirahige, K., and Uhlmann, F. (2008). Identification of cis-acting sites for condensin loading onto budding yeast chromosomes. Genes Dev *22*, 2215-2227.

De Dieuleveult, M. (2010). Implication des facteurs de remodelage de chromatine de la famille CHD dans les réseaux de régulation transcriptionnelle des cellules souches embryonnaires. Thesis, Paris XI.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., and Pagano, A. (2007). The expanding RNA polymerase III transcriptome. Trends Genet *23*, 614-622.

Dieci, G., Giuliodori, S., Catellani, M., Percudani, R., and Ottonello, S. (2002). Intragenic promoter adaptation and facilitated RNA polymerase III recycling in the transcription of SCR1, the 7SL RNA gene of Saccharomyces cerevisiae. J Biol Chem *277*, 6903-6914.

Dieci, G., Percudani, R., Giuliodori, S., Bottarelli, L., and Ottonello, S. (2000). TFIIIC-independent in vitro transcription of yeast tRNA genes. J Mol Biol *299*, 601-613.

Dieci, G., Preti, M., and Montanini, B. (2009). Eukaryotic snoRNAs: a paradigm for gene expression flexibility. Genomics *94*, 83-88.

Dieci, G., and Sentenac, A. (1996). Facilitated recycling pathway for RNA polymerase III. Cell *84*, 245-252.

Dieci, G., and Sentenac, A. (2003). Detours and shortcuts to transcription reinitiation. Trends Biochem Sci *28*, 202-209.

Diribarne, G., and Bensaude, O. (2009). 7SK RNA, a non-coding RNA regulating P-TEFb, a general transcription factor. RNA Biol *6*, 122-128.

Dittmar, K. A., Goodenbour, J. M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS Genet *2*, e221.

Donze, D., Adams, C. R., Rine, J., and Kamakaka, R. T. (1999). The boundaries of the silenced HMR domain in Saccharomyces cerevisiae. Genes Dev *13*, 698-708.

Donze, D., and Kamakaka, R. T. (2001). RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in Saccharomyces cerevisiae. Embo J *20*, 520-531.

Donze, D., and Kamakaka, R. T. (2002). Braking the silence: how heterochromatic gene repression is stopped in its tracks. Bioessays *24*, 344-349.

Dubey, R. N., and Gartenberg, M. R. (2007). A tDNA establishes cohesion of a neighboring silent chromatin domain. Genes Dev *21*, 2150-2160.

Dumay-Odelot, H., Marck, C., Durrieu-Gaillard, S., Lefebvre, O., Jourdain, S., Prochazkova, M., Pflieger, A., and Teichmann, M. (2007). Identification, molecular cloning, and characterization of the sixth subunit of human transcription factor TFIIIC. J Biol Chem *282*, 17179-17189.

Egloff, S., O'Reilly, D., Chapman, R. D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D., and Murphy, S. (2007). Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. Science *318*, 1777-1779.

Esakova, O., and Krasilnikov, A. S. (2010). Of proteins and RNA: the RNase P/MRP family. Rna *16*, 1725-1747.

Esnault, C., Ghavi-Helm, Y., Brun, S., Soutourina, J., Van Berkum, N., Boschiero, C., Holstege, F., and Werner, M. (2008). Mediator-dependent recruitment of TFIIH modules in preinitiation complex. Mol Cell *31*, 337-346.

Espinosa, J. M. (2010). The meaning of pausing. Mol Cell *40*, 507-508.

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. Nat Rev Genet *10*, 605-616.

Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S. J. (2008). FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics *24*, 1729-1730.

Felton-Edkins, Z. A., Fairley, J. A., Graham, E. L., Johnston, I. M., White, R. J., and Scott, P. H. (2003). The mitogen-activated protein (MAP) kinase ERK induces tRNA synthesis by phosphorylating TFIIIB. Embo J *22*, 2422-2432.

Fernandez-Tornero, C., Bottcher, B., Riva, M., Carles, C., Steuerwald, U., Ruigrok, R. W., Sentenac, A., Muller, C. W., and Schoehn, G. (2007). Insights into transcription initiation and termination from the electron microscopy structure of yeast RNA polymerase III. Mol Cell *25*, 813-823.

Ferrari, R., Rivetti, C., Acker, J., and Dieci, G. (2004). Distinct roles of transcription factors TFIIIB and TFIIIC in RNA polymerase III transcription reinitiation. Proc Natl Acad Sci U S A *101*, 13442-13447.

Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J. P., White, R. J., and Aberdam, D. (2001). Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. Nat Genet *28*, 77-81.

Fish, R. N., and Kane, C. M. (2002). Promoting elongation with transcript cleavage stimulatory factors. Biochim Biophys Acta *1577*, 287-307.

Fujinaga, K., Irwin, D., Huang, Y., Taube, R., Kurosu, T., and Peterlin, B. M. (2004). Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. Mol Cell Biol *24*, 787-795.

Fullwood, M. J., Han, Y., Wei, C. L., Ruan, X., and Ruan, Y. (2010). Chromatin interaction analysis using paired-end tag sequencing. Curr Protoc Mol Biol *Chapter 21*, Unit 21 15 21-25.

Geiduschek, E. P., and Kassavetis, G. A. (2001). The RNA polymerase III transcription apparatus. J Mol Biol *310*, 1-26.

Gerasimova, T. I., Byrd, K., and Corces, V. G. (2000). A chromatin insulator determines the nuclear localization of DNA. Mol Cell *6*, 1025-1035.

Gerasimova, T. I., and Corces, V. G. (2001). Chromatin insulators and boundaries: effects on transcription and nuclear organization. Annu Rev Genet *35*, 193-208.

Ghavi-Helm, Y., Michaut, M., Acker, J., Aude, J. C., Thuriaux, P., Werner, M., and Soutourina, J. (2008). Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription. Genes Dev *22*, 1934-1947.

Gilchrist, D. A., Dos Santos, G., Fargo, D. C., Xie, B., Gao, Y., Li, L., and Adelman, K. (2010). Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell *143*, 540-551.

Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., and Kornberg, R. D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science *292*, 1876-1882.

Gomez-Roman, N., Grandori, C., Eisenman, R. N., and White, R. J. (2003). Direct activation of RNA polymerase III transcription by c-Myc. Nature *421*, 290-294.

Goodenbour, J. M., and Pan, T. (2006). Diversity of tRNA genes in eukaryotes. Nucleic Acids Res *34*, 6137-6146.

Goodfellow, S. J., Graham, E. L., Kantidakis, T., Marshall, L., Coppins, B. A., Oficjalska-Pham, D., Gerard, M., Lefebvre, O., and White, R. J. (2008). Regulation of RNA polymerase III transcription by Maf1 in mammalian cells. J Mol Biol *378*, 481-491.

Goodfellow, S. J., Innes, F., Derblay, L. E., MacLellan, W. R., Scott, P. H., and White, R. J. (2006). Regulation of RNA polymerase III transcription during hypertrophic growth. Embo J *25*, 1522-1533.

Grewal, S. I., and Moazed, D. (2003). Heterochromatin and epigenetic control of gene expression. Science *301*, 798-802.

Grove, A., Adessa, M. S., Geiduschek, E. P., and Kassavetis, G. A. (2002). Marking the start site of RNA polymerase III transcription: the role of constraint, compaction and continuity of the transcribed DNA strand. Embo J *21*, 704-714.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948-951.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Guglielmi, B., Soutourina, J., Esnault, C., and Werner, M. (2007). TFIIS elongation factor and Mediator act in conjunction during transcription initiation in vivo. Proc Natl Acad Sci U S A *104*, 16062-16067.

Guo, H., and Price, D. H. (1993). Mechanism of DmS-II-mediated pause suppression by Drosophila RNA polymerase II. J Biol Chem *268*, 18762-18770.

Haeusler, R. A., and Engelke, D. R. (2004). Genome organization in three dimensions: thinking outside the line. Cell Cycle *3*, 273-275.

Haeusler, R. A., and Engelke, D. R. (2006). Spatial organization of transcription by RNA polymerase III. Nucleic Acids Res *34*, 4826-4836.

Haeusler, R. A., Pratt-Hyatt, M., Good, P. D., Gipson, T. A., and Engelke, D. R. (2008). Clustering of yeast tRNA genes is mediated by specific association of condensin with tRNA gene transcription complexes. Genes Dev *22*, 2204-2214.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nat Struct Mol Biol *11*, 394-403.

Hakimi, M. A., Bochar, D. A., Schmiesing, J. A., Dong, Y., Barak, O. G., Speicher, D. W., Yokomori, K., and Shiekhattar, R. (2002). A chromatin remodelling complex that loads cohesin onto human chromosomes. Nature *418*, 994-998.

Hampsey, M., and Reinberg, D. (2003). Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation. Cell *113*, 429-432.

Harismendy, O., Gendrel, C. G., Soularue, P., Gidrol, X., Sentenac, A., Werner, M., and Lefebvre, O. (2003). Genome-wide location of yeast RNA polymerase III transcription machinery. Embo J *22*, 4738-4747.

Haurie, V., Durrieu-Gaillard, S., Dumay-Odelot, H., Da Silva, D., Rey, C., Prochazkova, M., Roeder, R. G., Besser, D., and Teichmann, M. (2010). Two isoforms of human RNA polymerase III with specific functions in cell growth and transformation. Proc Natl Acad Sci U S A *107*, 4176-4181.

Hausner, W., Lange, U., and Musfeldt, M. (2000). Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase. J Biol Chem *275*, 12393-12399.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. Nat Rev Genet *11*, 476-486.

Hawley, D. K., Wiest, D. K., Holtz, M. S., and Wang, D. (1993). Transcriptional pausing, arrest, and readthrough at the adenovirus major late attenuation site. Cell Mol Biol Res *39*, 339-348.

Henry, R. W., Mittal, V., Ma, B., Kobayashi, R., and Hernandez, N. (1998). SNAP19 mediates the assembly of a functional core promoter complex (SNAPc) shared by RNA polymerases II and III. Genes Dev *12*, 2664-2672.

Hernandez, N. (1993). TBP, a universal eukaryotic transcription factor? Genes Dev *7*, 1291-1308.

Hernandez, N. (2001). Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. J Biol Chem *276*, 26733-26736.

Hon, G., Ren, B., and Wang, W. (2008). ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol *4*, e1000201.

Hsieh, Y. J., Kundu, T. K., Wang, Z., Kovelman, R., and Roeder, R. G. (1999). The TFIIIC90 subunit of TFIIIC interacts with multiple components of the RNA polymerase III machinery and contains a histone-specific acetyltransferase activity. Mol Cell Biol *19*, 7697-7704.

Huang, Y., and Maraia, R. J. (2001). Comparison of the RNA polymerase III transcription machinery in Schizosaccharomyces pombe, Saccharomyces cerevisiae and human. Nucleic Acids Res *29*, 2675-2690.

Hubbard, K., Catalano, J., Puri, R. K., and Gnatt, A. (2008). Knockdown of TFIIS by RNA silencing inhibits cancer cell proliferation and induces apoptosis. BMC Cancer *8*, 133.

Hull, M. W., Erickson, J., Johnston, M., and Engelke, D. R. (1994). tRNA genes as transcriptional repressor elements. Mol Cell Biol *14*, 1266-1277.

Isogai, Y., Takada, S., Tjian, R., and Keles, S. (2007). Novel TRF1/BRF target genes revealed by genome-wide analysis of Drosophila Pol III transcription. Embo J *26*, 79-89.

Ito, T., Arimitsu, N., Takeuchi, M., Kawamura, N., Nagata, M., Saso, K., Akimitsu, N., Hamamoto, H., Natori, S., Miyajima, A., and Sekimizu, K. (2006). Transcription elongation factor S-II is required for definitive hematopoiesis. Mol Cell Biol *26*, 3194-3203.

Ito, T., Xu, Q., Takeuchi, H., Kubo, T., and Natori, S. (1996). Spermatocyte-specific expression of the gene for mouse testis-specific transcription elongation factor S-II. FEBS Lett *385*, 21-24.

Izban, M. G., and Luse, D. S. (1992). Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. J Biol Chem *267*, 13647-13655.

Jaenisch, R. (1997). DNA methylation and imprinting: why bother? Trends Genet *13*, 323-329.

Jasiak, A. J., Armache, K. J., Martens, B., Jansen, R. P., and Cramer, P. (2006). Structural biology of RNA polymerase III: subcomplex C17/25 X-ray structure and 11 subunit enzyme model. Mol Cell *23*, 71-81.

Joazeiro, C. A., Kassavetis, G. A., and Geiduschek, E. P. (1996). Alternative outcomes in assembly of promoter complexes: the roles of TBP and a flexible linker in placing TFIIIB on tRNA genes. Genes Dev *10*, 725-739.

Johnson, S. A., Dubeau, L., and Johnson, D. L. (2008). Enhanced RNA polymerase III-dependent transcription is required for oncogenic transformation. J Biol Chem *283*, 19184-19191.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res *36*, 5221-5231.

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., *et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Kaiser, M. W., and Brow, D. A. (1995). Lethal mutations in a yeast U6 RNA gene B block promoter element identify essential contacts with transcription factor-IIIC. J Biol Chem *270*, 11398-11405.

Kassavetis, G. A., Blanco, J. A., Johnson, T. E., and Geiduschek, E. P. (1992). Formation of open and elongating transcription complexes by RNA polymerase III. J Mol Biol *226*, 47-58.

Kassavetis, G. A., Braun, B. R., Nguyen, L. H., and Geiduschek, E. P. (1990). S. cerevisiae TFIIIB is the transcription initiation factor proper of RNA polymerase III, while TFIIIA and TFIIIC are assembly factors. Cell *60*, 235-245.

Kassavetis, G. A., Han, S., Naji, S., and Geiduschek, E. P. (2003). The role of transcription initiation factor IIIB subunits in promoter opening probed by photochemical cross-linking. J Biol Chem *278*, 17912-17917.

Kassavetis, G. A., Letts, G. A., and Geiduschek, E. P. (2001). The RNA polymerase III transcription initiation factor TFIIIB participates in two steps of promoter opening. Embo J *20*, 2823-2834.

Kassavetis, G. A., Nguyen, S. T., Kobayashi, R., Kumar, A., Geiduschek, E. P., and Pisano, M. (1995). Cloning, expression, and function of TFC5, the gene encoding the B" component of the Saccharomyces cerevisiae RNA polymerase III transcription factor TFIIIB. Proc Natl Acad Sci U S A *92*, 9786-9790.

Kassavetis, G. A., Prakash, P., and Shim, E. (2010). The C53/C37 subcomplex of RNA polymerase III lies near the active site and participates in promoter opening. J Biol Chem *285*, 2695-2706.

Kendall, A., Hull, M. W., Bertrand, E., Good, P. D., Singer, R. H., and Engelke, D. R. (2000). A CBF5 mutation that disrupts nucleolar localization of early tRNA biosynthesis in yeast also suppresses tRNA gene-mediated transcriptional silencing. Proc Natl Acad Sci U S A *97*, 13108-13113.

Kenneth, N. S., Marshall, L., and White, R. J. (2008). Recruitment of RNA polymerase III in vivo. Nucleic Acids Res *36*, 3757-3764.

Kenneth, N. S., Ramsbottom, B. A., Gomez-Roman, N., Marshall, L., Cole, P. A., and White, R. J. (2007). TRRAP and GCN5 are used by c-Myc to activate RNA polymerase III transcription. Proc Natl Acad Sci U S A *104*, 14917-14922.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics *26*, 2204-2207.

Kettenberger, H., Armache, K. J., and Cramer, P. (2003). Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage. Cell *114*, 347-357.

Kettenberger, H., Armache, K. J., and Cramer, P. (2004). Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. Mol Cell *16*, 955-965.

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol *26*, 1351-1359.

Kim, B., Nesvizhskii, A. I., Rani, P. G., Hahn, S., Aebersold, R., and Ranish, J. A. (2007). The transcription elongation factor TFIIS is a component of RNA polymerase II preinitiation complexes. Proc Natl Acad Sci U S A *104*, 16068-16073.

Kim, J. K., Samaranayake, M., and Pradhan, S. (2009). Epigenetic mechanisms in mammals. Cell Mol Life Sci *66*, 596-612.

Kireeva, M. L., Hancock, B., Cremona, G. H., Walter, W., Studitsky, V. M., and Kashlev, M. (2005). Nature of the nucleosomal barrier to RNA polymerase II. Mol Cell *18*, 97-108.

Kleinschmidt, R. A., LeBlanc, K. E., and Donze, D. (2011). Autoregulation of an RNA polymerase II promoter by the RNA polymerase III transcription factor III C (TF(III)C) complex. Proc Natl Acad Sci U S A *108*, 8385-8389.

Kobayashi, S., and Anzai, K. (1998). An E-box sequence acts as a transcriptional activator for BC1 RNA expression by RNA polymerase III in the brain. Biochem Biophys Res Commun *245*, 59-63.

Koch, F., Jourquin, F., Ferrier, P., and Andrau, J. C. (2008). Genome-wide RNA polymerase II: not genes only! Trends Biochem Sci *33*, 265-273.

Koch, H. B., Zhang, R., Verdoodt, B., Bailey, A., Zhang, C. D., Yates, J. R., 3rd, Menssen, A., and Hermeking, H. (2007). Large-scale identification of c-MYC-associated proteins using a combined TAP/MudPIT approach. Cell Cycle *6*, 205-217.

Komarnitsky, P., Cho, E. J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev *14*, 2452-2460.

Koval, A. P., and Kramerov, D. A. (2009). 5'-flanking sequences can dramatically influence 4.5SH RNA gene transcription by RNA-polymerase III. Gene *446*, 75-80.

Kramerov, D. A., and Vassetzky, N. S. (2005). Short retroposons in eukaryotic genomes. Int Rev Cytol *247*, 165-221.

Krummel, B., and Chamberlin, M. J. (1989). RNA chain initiation by Escherichia coli RNA polymerase. Structural transitions of the enzyme in early ternary complexes. Biochemistry *28*, 7829-7842.

Kuhn, C. D., Geiger, S. R., Baumli, S., Gartmann, M., Gerber, J., Jennebach, S., Mielke, T., Tschochner, H., Beckmann, R., and Cramer, P. (2007). Functional architecture of RNA polymerase I. Cell *131*, 1260-1272.

Kundu, T. K., Wang, Z., and Roeder, R. G. (1999). Human TFIIIC relieves chromatin-mediated repression of RNA polymerase III transcription and contains an intrinsic histone acetyltransferase activity. Mol Cell Biol *19*, 1605-1615.

Labhart, P., and Morgan, G. T. (1998). Identification of novel genes encoding transcription elongation factor TFIIS (TCEA) in vertebrates: conservation of three distinct TFIIS isoforms in frog, mouse, and human. Genomics *52*, 278-288.

Lakdawalla, A., and VanSteenhouse. H. (2008). .Next-generation genome sequencing: towards personalized medicine. Edited by Michal Janitz.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Landrieux, E., Alic, N., Ducrot, C., Acker, J., Riva, M., and Carles, C. (2006). A subcomplex of RNA polymerase III subunits involved in transcription termination and reinitiation. Embo J *25*, 118-128.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Larminie, C. G., Cairns, C. A., Mital, R., Martin, K., Kouzarides, T., Jackson, S. P., and White, R. J. (1997). Mechanistic analysis of RNA polymerase III regulation by the retinoblastoma protein. Embo J *16*, 2061-2071.

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res *32*, 11-16.

Li, B., Carey, M., and Workman, J. L. (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res *18*, 1851-1858.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O.*, et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Lis, J. (1998). Promoter-associated pausing in promoter architecture and postinitiation transcriptional regulation. Cold Spring Harb Symp Quant Biol *63*, 347-356.

Listerman, I., Bledau, A. S., Grishina, I., and Neugebauer, K. M. (2007). Extragenic accumulation of RNA polymerase II enhances transcription by RNA polymerase III. PLoS Genet *3*, e212.

Liu, P., Jenkins, N. A., and Copeland, N. G. (2003). A highly efficient recombineering-based method for generating conditional knockout mutations. Genome Res *13*, 476-484.

Lobo, S. M., and Hernandez, N. (1989). A 7 bp mutation converts a human RNA polymerase II snRNA promoter into an RNA polymerase III promoter. Cell *58*, 55-67.

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res *25*, 955-964.

Lunyak, V. V., Prefontaine, G. G., Nunez, E., Cramer, T., Ju, B. G., Ohgi, K. A., Hutt, K., Roy, R., Garcia-Diaz, A., Zhu, X.*, et al.* (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science *317*, 248-251.

Lykke-Andersen, S., Mapendano, C. K., and Jensen, T. H. (2011). An ending is a new beginning: transcription termination supports re-initiation. Cell Cycle *10*, 863-865.

Malagon, F., Tong, A. H., Shafer, B. K., and Strathern, J. N. (2004). Genetic interactions of DST1 in Saccharomyces cerevisiae suggest a role of TFIIS in the initiation-elongation transition. Genetics *166*, 1215-1227.

Maraia, R. J. (2001). La protein and the trafficking of nascent RNA polymerase iii transcripts. J Cell Biol *153*, F13-18.

Marck, C., Kachouri-Lafond, R., Lafontaine, I., Westhof, E., Dujon, B., and Grosjean, H. (2006). The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications. Nucleic Acids Res *34*, 1816-1835.

Marshall, L., Kenneth, N. S., and White, R. J. (2008). Elevated tRNA(iMet) synthesis can drive cell proliferation and oncogenic transformation. Cell *133*, 78-89.

Marshall, L., and White, R. J. (2008). Non-coding RNA production by RNA polymerase III is implicated in cancer. Nat Rev Cancer *8*, 911-914.

Martignetti, J. A., and Brosius, J. (1995). BC1 RNA: transcriptional analysis of a neural cell-specific RNA polymerase III transcript. Mol Cell Biol *15*, 1642-1650.

Mattaj, I. W., Dathan, N. A., Parry, H. D., Carbon, P., and Krol, A. (1988). Changing the RNA polymerase specificity of U snRNA gene promoters. Cell *55*, 435-442.

Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., Qi, J., Schuster, S. C., Albert, I., and Pugh, B. F. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res *18*, 1073-1083.

Max, T., Sogaard, M., and Svejstrup, J. Q. (2007). Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. J Biol Chem *282*, 14113-14120.

McFarlane, R. J., and Whitehall, S. K. (2009). tRNA genes in eukaryotic genome organization and reorganization. Cell Cycle *8*, 3102-3106.

Mertens, C., and Roeder, R. G. (2008). Different functional modes of p300 in activation of RNA polymerase III transcription from chromatin templates. Mol Cell Biol *28*, 5764-5776.

Metzker, M. L. (2010). Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Min, I. M., Waterfall, J. J., Core, L. J., Munroe, R. J., Schimenti, J., and Lis, J. T. (2010). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev *25*, 742-754.

Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. Cell *128*, 787-800.

Moqtaderi, Z., and Struhl, K. (2004). Genome-wide occupancy profile of the RNA polymerase III machinery in Saccharomyces cerevisiae reveals loci with incomplete transcription complexes. Mol Cell Biol *24*, 4118-4127.

Moqtaderi, Z., Wang, J., Raha, D., White, R. J., Snyder, M., Weng, Z., and Struhl, K. (2010). Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nat Struct Mol Biol *17*, 635-640.

Mossink, M. H., van Zon, A., Scheper, R. J., Sonneveld, P., and Wiemer, E. A. (2003). Vaults: a ribonucleoprotein particle involved in drug resistance? Oncogene *22*, 7458-7467.

Muller, F., Zaucker, A., and Tora, L. (2010). Developmental regulation of transcription initiation: more than just changing the actors. Curr Opin Genet Dev *20*, 533-540.

Murphy, S., Yoon, J. B., Gerster, T., and Roeder, R. G. (1992). Oct-1 and Oct-2 potentiate functional interactions of a transcription factor with the proximal sequence element of small nuclear RNA genes. Mol Cell Biol *12*, 3247-3261.

Muse, G. W., Gilchrist, D. A., Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat Genet *39*, 1507-1511.

Mylona, A., Fernandez-Tornero, C., Legrand, P., Haupt, M., Sentenac, A., Acker, J., and Muller, C. W. (2006). Structure of the tau60/Delta tau91 subcomplex of yeast transcription factor IIIC: insights into preinitiation complex assembly. Mol Cell *24*, 221-232.

Natori, S., Takeuchi, K., Takahashi, K., and Mizuno, D. (1973). DNA dependent RNA polymerase from Ehrlich ascites tumor cells. II. Factors stimulating the activity of RNA polymerase II. J Biochem *73*, 879-888.

Nechaev, S., Fargo, D. C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335-338.

Noma, K., Allis, C. D., and Grewal, S. I. (2001). Transitions in distinct histone H3 methylation patterns at the heterochromatin domain boundaries. Science *293*, 1150-1155.

Noma, K., Cam, H. P., Maraia, R. J., and Grewal, S. I. (2006). A role for TFIIIC transcription factor complex in genome organization. Cell *125*, 859-872.

Oki, M., and Kamakaka, R. T. (2005). Barrier function at HMR. Mol Cell *19*, 707-716.

Oler, A. J., Alla, R. K., Roberts, D. N., Wong, A., Hollenhorst, P. C., Chandler, K. J., Cassiday, P. A., Nelson, C. A., Hagedorn, C. H., Graves, B. J., and Cairns, B. R. (2010). Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. Nat Struct Mol Biol *17*, 620-628.

Orioli, A., Pascali, C., Pagano, A., Teichmann, M., and Dieci, G. (2011a). RNA polymerase III transcription control elements: Themes and variations. Gene.

Orioli, A., Pascali, C., Quartararo, J., Diebel, K. W., Praz, V., Romascano, D., Percudani, R., van Dyk, L. F., Hernandez, N., Teichmann, M., and Dieci, G. (2011b). Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. Nucleic Acids Res *39*, 5499-5512.

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. Nat Rev Genet *12*, 87-98.

Pagano, A., Castelnuovo, M., Tortelli, F., Ferrari, R., Dieci, G., and Cancedda, R. (2007). New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. PLoS Genet *3*, e1.

Palangat, M., Renner, D. B., Price, D. H., and Landick, R. (2005). A negative elongation factor for human RNA polymerase II inhibits the anti-arrest transcript-cleavage factor TFIIS. Proc Natl Acad Sci U S A *102*, 15036-15041.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., *et al.* (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422-433.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet *10*, 669-680.

Parnell, T. J., Huff, J. T., and Cairns, B. R. (2008). RSC regulates nucleosome positioning at Pol II genes and density at Pol III genes. Embo J *27*, 100-110.

Paule, M. R., and White, R. J. (2000). Survey and summary: transcription by RNA polymerases I and III. Nucleic Acids Res *28*, 1283-1298.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nat Methods *6*, S22-32.

Peterlin, B. M., and Price, D. H. (2006). Controlling the elongation phase of transcription with P-TEFb. Mol Cell *23*, 297-305.

Phillips, J. E., and Corces, V. G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194-1211.

Phizicky, E. M., and Hopper, A. K. (2010). tRNA biology charges to the front. Genes Dev *24*, 1832-1860.

Plotkin, J. B., Robins, H., and Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A *101*, 12588-12591.

Pokholok, D. K., Hannett, N. M., and Young, R. A. (2002). Exchange of RNA polymerase II initiation and elongation factors during gene expression in vivo. Mol Cell *9*, 799-809.

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., *et al.* (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. Cell *122*, 517-527.

Prasanth, K. V., Camiolo, M., Chan, G., Tripathi, V., Denis, L., Nakamura, T., Hubner, M. R., and Spector, D. L. (2010). Nuclear organization and dynamics of 7SK RNA in regulating gene expression. Mol Biol Cell *21*, 4184-4196.

Prather, D. M., Larschan, E., and Winston, F. (2005). Evidence that the elongation factor TFIIS plays a role in transcription initiation at GAL1 in Saccharomyces cerevisiae. Mol Cell Biol *25*, 2650-2659.

Raab, J. R., and Kamakaka, R. T. (2010). Insulators and promoters: closer than we think. Nat Rev Genet *11*, 439-446.

Radonjic, M., Andrau, J. C., Lijnzaad, P., Kemmeren, P., Kockelkorn, T. T., van Leenen, D., van Berkum, N. L., and Holstege, F. C. (2005). Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit. Mol Cell *18*, 171-183.

Raha, D., Wang, Z., Moqtaderi, Z., Wu, L., Zhong, G., Gerstein, M., Struhl, K., and Snyder, M. (2010). Close association of RNA polymerase II and many transcription factors with Pol III genes. Proc Natl Acad Sci U S A *107*, 3639-3644.

Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A., and Young, R. A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432-445.

Reina, J. H., and Hernandez, N. (2007). On a roll for new TRF targets. Genes Dev *21*, 2855-2860.

Roberts, D. N., Stewart, A. J., Huff, J. T., and Cairns, B. R. (2003). The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. Proc Natl Acad Sci U S A *100*, 14695-14700.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods *4*, 651-657.

Roeder, R. G., and Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224*, 234-237.

Rougvie, A. E., and Lis, J. T. (1988). The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell *54*, 795-804.

Roy, A. M., West, N. C., Rao, A., Adhikari, P., Aleman, C., Barnes, A. P., and Deininger, P. L. (2000). Upstream flanking sequences and transcription of SINEs. J Mol Biol *302*, 17-25.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol *27*, 66-75.

Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., Aebersold, R., Ranish, J. A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. Proc Natl Acad Sci U S A *105*, 8309-8314.

Ruth, J., Conesa, C., Dieci, G., Lefebvre, O., Dusterhoft, A., Ottonello, S., and Sentenac, A. (1996). A suppressor of mutations in the class III transcription system encodes a component of yeast TFIIIB. Embo J *15*, 1941-1949.

Saunders, A., Core, L. J., and Lis, J. T. (2006). Breaking barriers to transcription elongation. Nat Rev Mol Cell Biol *7*, 557-567.

Sawadogo, M., Huet, J., and Fromageot, P. (1980). Similar binding site for P37 factor on yeast RNA polymerases A and B. Biochem Biophys Res Commun *96*, 258-264.

Sawadogo, M., Lescure, B., Sentenac, A., and Fromageot, P. (1981). Native deoxyribonucleic acid transcription by yeast RNA polymerase--P37 complex. Biochemistry *20*, 3542-3547.

Schaub, M., Krol, A., and Carbon, P. (2000). Structural organization of Staf-DNA complexes. Nucleic Acids Res *28*, 2114-2121.

Schramm, L., and Hernandez, N. (2002). Recruitment of RNA polymerase III to its target promoters. Genes Dev *16*, 2593-2620.

Schramm, L., Pendergrast, P. S., Sun, Y., and Hernandez, N. (2000). Different human TFIIIB activities direct RNA polymerase III transcription from TATA-containing and TATA-less promoters. Genes Dev *14*, 2650-2663.

Scott, K. C., Merrett, S. L., and Willard, H. F. (2006). A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains. Curr Biol *16*, 119-129.

Scott, P. H., Cairns, C. A., Sutcliffe, J. E., Alzuherri, H. M., McLees, A., Winter, A. G., and White, R. J. (2001). Regulation of RNA polymerase III transcription during cell cycle entry. J Biol Chem *276*, 1005-1014.

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A., and Sharp, P. A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.

Seila, A. C., Core, L. J., Lis, J. T., and Sharp, P. A. (2009). Divergent transcription: a new feature of active promoters. Cell Cycle *8*, 2557-2564.

Sigurdsson, S., Dirac-Svejstrup, A. B., and Svejstrup, J. Q. (2010). Evidence that transcript cleavage is essential for RNA polymerase II transcription and cell viability. Mol Cell *38*, 202-210.

Simms, T. A., Dugas, S. L., Gremillion, J. C., Ibos, M. E., Dandurand, M. N., Toliver, T. T., Edwards, D. J., and Donze, D. (2008). TFIIIC binding sites function as both heterochromatin barriers and chromatin insulators in Saccharomyces cerevisiae. Eukaryot Cell *7*, 2078-2086.

Simms, T. A., Miller, E. C., Buisson, N. P., Jambunathan, N., and Donze, D. (2004). The Saccharomyces cerevisiae TRT2 tRNAThr gene upstream of STE6 is a barrier to repression in MATalpha cells and exerts a potential tRNA position effect in MATa cells. Nucleic Acids Res *32*, 5206-5213.

Soutourina, J., Bordas-Le Floch, V., Gendrel, G., Flores, A., Ducrot, C., Dumay-Odelot, H., Soularue, P., Navarro, F., Cairns, B. R., Lefebvre, O., and Werner, M. (2006). Rsc4 connects the chromatin remodeler RSC to RNA polymerases. Mol Cell Biol *26*, 4920-4933.

Studitsky, V. M., Kassavetis, G. A., Geiduschek, E. P., and Felsenfeld, G. (1997). Mechanism of transcription through the nucleosome by eukaryotic RNA polymerase. Science *278*, 1960-1963.

Studitsky, V. M., Walter, W., Kireeva, M., Kashlev, M., and Felsenfeld, G. (2004). Chromatin remodeling by RNA polymerases. Trends Biochem Sci *29*, 127-135.

Su, L. K., Kinzler, K. W., Vogelstein, B., Preisinger, A. C., Moser, A. R., Luongo, C., Gould, K. A., and Dove, W. F. (1992). Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene. Science *256*, 668-670.

Sun, H., Yuan, Y., Wu, Y., Liu, H., Liu, J. S., and Xie, H. (2010). Tmod: toolbox of motif discovery. Bioinformatics *26*, 405-407.

Sutcliffe, J. E., Brown, T. R., Allison, S. J., Scott, P. H., and White, R. J. (2000). Retinoblastoma protein disrupts interactions required for RNA polymerase III transcription. Mol Cell Biol *20*, 9192-9202.

Tanaka, A., Watanabe, T., Iida, Y., Hanaoka, F., and Ohkuma, Y. (2009). Central forkhead domain of human TFIIE beta plays a primary role in binding double-stranded DNA at transcription initiation. Genes Cells *14*, 395-405.

Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., Fu, Z., and Noma, K. (2010). Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res *38*, 8164-8177.

Teichmann, M., Dieci, G., Pascali, C., and Boldina, G. (2010). General transcription factors and subunits of RNA polymerase III: Paralogs for promoter- and cell type-specific transcription in multicellular eukaryotes. Transcr *1*, 130-135.

Teichmann, M., and Seifart, K. H. (1995). Physical separation of two different forms of human TFIIIB active in the transcription of the U6 or the VAI gene in vitro. Embo J *14*, 5974-5983.

Teichmann, M., Wang, Z., and Roeder, R. G. (2000). A stable complex of a novel transcription factor IIB-related factor, human TFIIIB50, and associated proteins mediate selective transcription by RNA polymerase III of genes with upstream promoter elements. Proc Natl Acad Sci U S A *97*, 14200-14205.

Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S., and van Helden, J. (2008). RSAT: regulatory sequence analysis tools. Nucleic Acids Res *36*, W119-127.

Trapnell, C., and Salzberg, S. L. (2009). How to map billions of short reads onto genomes. Nat Biotechnol *27*, 455-457.

Valenzuela, L., Dhillon, N., Dubey, R. N., Gartenberg, M. R., and Kamakaka, R. T. (2008). Long-range communication between the silencers of HMR. Mol Cell Biol *28*, 1924-1935.

Valenzuela, L., Dhillon, N., and Kamakaka, R. T. (2009). Transcription independent insulation at TFIIIC-dependent insulators. Genetics *183*, 131-148.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods *5*, 829-834.

Vannini, A., Ringel, R., Kusser, A. G., Berninghausen, O., Kassavetis, G. A., and Cramer, P. (2010). Molecular basis of RNA polymerase III transcription repression by Maf1. Cell *143*, 59-70.

Wade, J. T., and Struhl, K. (2008). The transition from transcriptional initiation to elongation. Curr Opin Genet Dev *18*, 130-136.

Waldschmidt, R., Wanandi, I., and Seifart, K. H. (1991). Identification of transcription factors required for the expression of mammalian U6 genes in vitro. Embo J *10*, 2595-2603.

Wang, D., and Hawley, D. K. (1993). Identification of a 3'-->5' exonuclease activity associated with human RNA polymerase II. Proc Natl Acad Sci U S A *90*, 843-847.

Wendt, K. S., and Peters, J. M. (2009). How cohesin and CTCF cooperate in regulating gene expression. Chromosome Res *17*, 201-214.

Werner, M., Thuriaux, P., and Soutourina, J. (2009). Structure-function analysis of RNA polymerases I and III. Curr Opin Struct Biol *19*, 740-745.

White, R.J. (1998). RNA Polymerase III Transcription (2nd edn), Springer-Verlag.

White, R. J. (2004). RNA polymerase III transcription and cancer. Oncogene *23*, 3208-3216.

White, R. J. (2005). RNA polymerases I and III, growth control and cancer. Nat Rev Mol Cell Biol *6*, 69-78.

White, R. J. (2008). RNA polymerases I and III, non-coding RNAs and cancer. Trends Genet *24*, 622-629.

White, R. J. (2011). Transcription by RNA polymerase III: more complex than we thought. Nat Rev Genet *12*, 459-463.

Willoughby, D. A., Vilalta, A., and Oshima, R. G. (2000). An Alu element from the K18 gene confers position-independent expression in transgenic mice. J Biol Chem *275*, 759-768.

Woiwode, A., Johnson, S. A., Zhong, S., Zhang, C., Roeder, R. G., Teichmann, M., and Johnson, D. L. (2008). PTEN represses RNA polymerase III-dependent transcription by targeting the TFIIIB complex. Mol Cell Biol *28*, 4204-4214.

Wood, A. J., Severson, A. F., and Meyer, B. J. (2010). Condensin and cohesin complexity: the expanding repertoire of functions. Nat Rev Genet *11*, 391-404.

Wu, C. H., Yamaguchi, Y., Benjamin, L. R., Horvat-Gordon, M., Washinsky, J., Enerly, E., Larsson, J., Lambertsson, A., Handa, H., and Gilmour, D. (2003). NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. Genes Dev *17*, 1402-1414.

Ye, T., Krebs, A. R., Choukrallah, M. A., Keime, C., Plewniak, F., Davidson, I., and Tora, L. (2011). seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res *39*, e35.

Ying, Q. L., Stavridis, M., Griffiths, D., Li, M., and Smith, A. (2003). Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. Nat Biotechnol *21*, 183-186.

Yoon, J. B., Murphy, S., Bai, L., Wang, Z., and Roeder, R. G. (1995). Proximal sequence element-binding transcription factor (PTF) is a multisubunit complex required for transcription of both RNA polymerase II- and RNA polymerase III-dependent small nuclear RNA genes. Mol Cell Biol *15*, 2019-2027.

Yuan, C. C., Zhao, X., Florens, L., Swanson, S. K., Washburn, M. P., and Hernandez, N. (2007). CHD8 associates with human Staf and contributes to efficient U6 RNA polymerase III transcription. Mol Cell Biol *27*, 8729-8738.

Yudkovsky, N., Ranish, J. A., and Hahn, S. (2000). A transcription reinitiation intermediate that is stabilized by activator. Nature *408*, 225-229.

Yusufzai, T. M., and Felsenfeld, G. (2004). The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. Proc Natl Acad Sci U S A *101*, 8620-8624.

Zaratiegui, M., Irvine, D. V., and Martienssen, R. A. (2007). Noncoding RNAs and gene silencing. Cell *128*, 763-776.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J. W., Nechaev, S., Adelman, K., Levine, M., and Young, R. A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet *39*, 1512-1516.

Zhang, J., Chiodini, R., Badr, A., and Zhang, G. (2011). The impact of next-generation sequencing on genomics. J Genet Genomics *38*, 95-109.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhao, X., Pendergrast, P. S., and Hernandez, N. (2001). A positioned nucleosome on the human U6 promoter allows recruitment of SNAPc by the Oct-1 POU domain. Mol Cell *7*, 539-549.

METHODES

Principe du Recombineering et Culture des cellules ES

Le protocole de Recombineering a été adapté de Liu et al (Liu et al., 2003), par Fayçal Boussouar, Maud de Dieuleveult du laboratoire de Matthieu Gérard, et Sébastien Graziani, du laboratoire de Michel Werner.

Afin d'optimiser les procédures d'immunoprécipitation, nous avons introduit une cassette codant une étiquette 6 histidines-Flag-HA au locus des gènes.

Cette étiquette est introduite en N-terminale pour TCEA1 (Yad Ghavi-Helm), car l'ajout d'une étiquette en C-terminal chez S. cerevisiae conduit à l'expression d'une protéine non fonctionnelle. Pour les protéines de la RNAP III, TFIIIB, et TFIIIC, la cassette codant l'étiquette est introduite après le dernier exon (Sébastien Graziani).

Principe général (voir figure 20)

Le développement de la technique de recombineering a révolutionné l'approche de la construction de vecteur de recombinaison homologue. Cette technique réalisée dans *E. coli*, est rendue possible grâce à l'utilisation de protéines de phage, comme celles codés par les gènes *Red* du bactériophage λ. Ces protéines permettent le transfert de fragments d'ADN double-brin et linéaires (dsDNA, double strand DNA) dans des plasmides par recombinaison homologue, au niveau d'une séquence ADN complémentaire clonée préalablement. Les deux gènes *Red* nécessaires à la recombinaison sont : *exo*, qui encode une exonucléase 5'-3' (Exo), et *bet*, qui code pour une protéine d'appariement (Beta). Exo produit une extrémité 3' simple brin débordante à partir de du côté 5' du dsDNA. Beta se lie à l'extrémité 3' simple brin débordante, crée par Exo, et stimule son hybridation avec le brin complémentaire du fragment cloné d'ADN. Les fonctions de recombinaison d'Exo et de Beta sont assistées par la protéine Gam. Celle-ci inhibe l'activité exonucléase de la protéine RecBCD, codée par *E. coli.* L'expression de ces gènes est sous la dépendance d'un répresseur sensible à la température. A la température de croissance de la souche SW106 d'*E. coli*, les gènes ne s'expriment pas. Leur expression est induite après un choc thermique à 42°C pendant une quinzaine de minutes.

Deux étapes de recombinaison homologues sont nécessaires pour construire le vecteur de recombinaison final, qui sera électroporé dans les cellules ES (figure 20). La première étape consiste en l'introduction de la région génomique d'environ 10 kb couvrant la région d'intérêt dans le plasmide pl253. Ce fragment génomique est obtenu à partir d'un BAC (Bacterial Artificial Chromosome). Lors de la

deuxième étape, une séquence codant l'étiquette est introduite dans le premier plasmide, juste après le codon d'initiation de la transcription, ou juste avant le codon STOP, selon le facteur étudié.

Deux séquences, les miniarms 5' et 3', aux deux extrémités de la région génomique d'intérêt sont obtenues par PCR, puis introduites dans le plasmide pl253, générant le Retrievel Plasmid. Le BAC contenant le région d'intérêt est introduit dans les bactéries *E. coli*, porteuse des gènes Red du phage λ. Par recombinaison homologue la région de 10 kb contenu dans le BAC, est insérée dans le Retrieval Plasmid, c'est l'étape de construction du Gap Repair Plasmid. Pour sélectionner les évènements de recombinaison homologue, et ainsi contrer tout intégration à un locus non désiré, le plasmide pl253 contient une cassette codant le gène de la Thymidine Kinase (TK).

En parallèle, deux séquences, le 5' et 3' bras d'homologie, encadrant le codon STOP, sont obtenues par PCR et clonées dans le pl452, contenant la séquence codant l'étiquette, générant le Mini-targeting Vector. Le 5' bras d'homologie est homologue à la séquence du dernier exon du facteur étudié. Le 3' bras d'homolgie est homologue à la séquence en aval du codon STOP. Le codon STOP doit être localisé dans le 3 bras d'homologie, pour que lors de la traduction, l'étiquette soit incorporée dans la séquence protéique du facteur. Ce plasmide contient en plus un gène de sélection de la Neomycine, encadrée de deux sites LoxP, sous la dépendance d'un promoteur fort CMV (CytomégaloVirus). Cette cassette permettra de sélectionner en cellules ES les clones qui ont intégré la construction. Enfin, une dernière étape de recombinaison homologue entre le MTV et le GRP permet d'insérer la séquence codant l'étiquette au niveau du codon STOP. Le plasmide final obtenu est le Neo-Targetd Plamid. Il contient les 10 kb de séquence génomique, la séquence codant l'étiquette, et la cassette Neomycine, et la cassette TK.

Dans le cas des gènes TCEA, l'étiquette a été incorporée au niveau du codon d'initiation, car chez la levure, la présence d'une étiquette en C-terminal de TFIIS altère son activité.

Le NTP est ensuite linéarisé par l'enzyme de restriction Not1, et le fragment dsDNA généré est électroporé en cellules ES.

Les cellules ES sont transformées par électroporation. Les événements de recombinaison sont sélectionnés grâce à la résistance au G418, apportée par la cassette Neomycine, et au gancyclovir (contre l'intégration de la cassette TK, la TK est une enzyme qui modifie le gancyclovir en produit toxique). Une centaine de clones sont repiqués, amplifiés puis génotypés par Southern Blot pour vérifier l'intégration correcte de la construction. Si la protéine est bien exprimée, la deuxième étape consiste à enlever la cassette Neomycine, qui peut gêner l'expression de la protéine. La cassette Neomycine flanquée de deux sites LoxP est excisée après élctroporation d'un vecteur codant la CRE recombinase, et d'un plasmide de résistance à la puromycine. Les clones sont sélectionnés suivant leur résistance à leur puromycine,

repiqués, amplifiés. La délétion de la cassette est vérifiée par Southern Blot, l'expression de la protéine étiquettée par Western Blot. L'intégrité de leur matériel génomique est vérifiée par caryotypage.

Le protocole détaillé du recombineering est en annexe.

Cellules ES

Deux lignées de cellules ES sont disponibles dans le laboratoire, les cellules 46C, et les AT1. Elles sont cultivées sur des fibroblastes nourriciers (MEF, Mouse Embryonic Fibroblast), dont la croissance a été stoppée par la mitomycine C.

La lignée 46C a été établie dans le laboratoire d'Austin Smith (Ying et al., 2003). Cette lignée a été modifiée afin de pouvoir suivre la différentiation en neurones. Le gène Sox1 est fusionné à la séquence codant la GFP, et au gène de résistance à la puromycine en aval d'une séquence IRES (Internal Ribosome Entry Site). Le gène Sox1 est un marqueur de la différentiation neuronale. Son expression apparaît au cours de la différentiation en neurones. Le marqueur de fluorescence GFP permet alors de suivre cette différentiation, tandis que la puromycine permet de sélectionner les cellules engagées dans la différentiation neuronale.

Les cellules AT1 sont utilisées pour la transmission germinale. Elles peuvent réinjecter dans les blastocystes, eux-mêmes réintroduits dans des souris-porteuses.

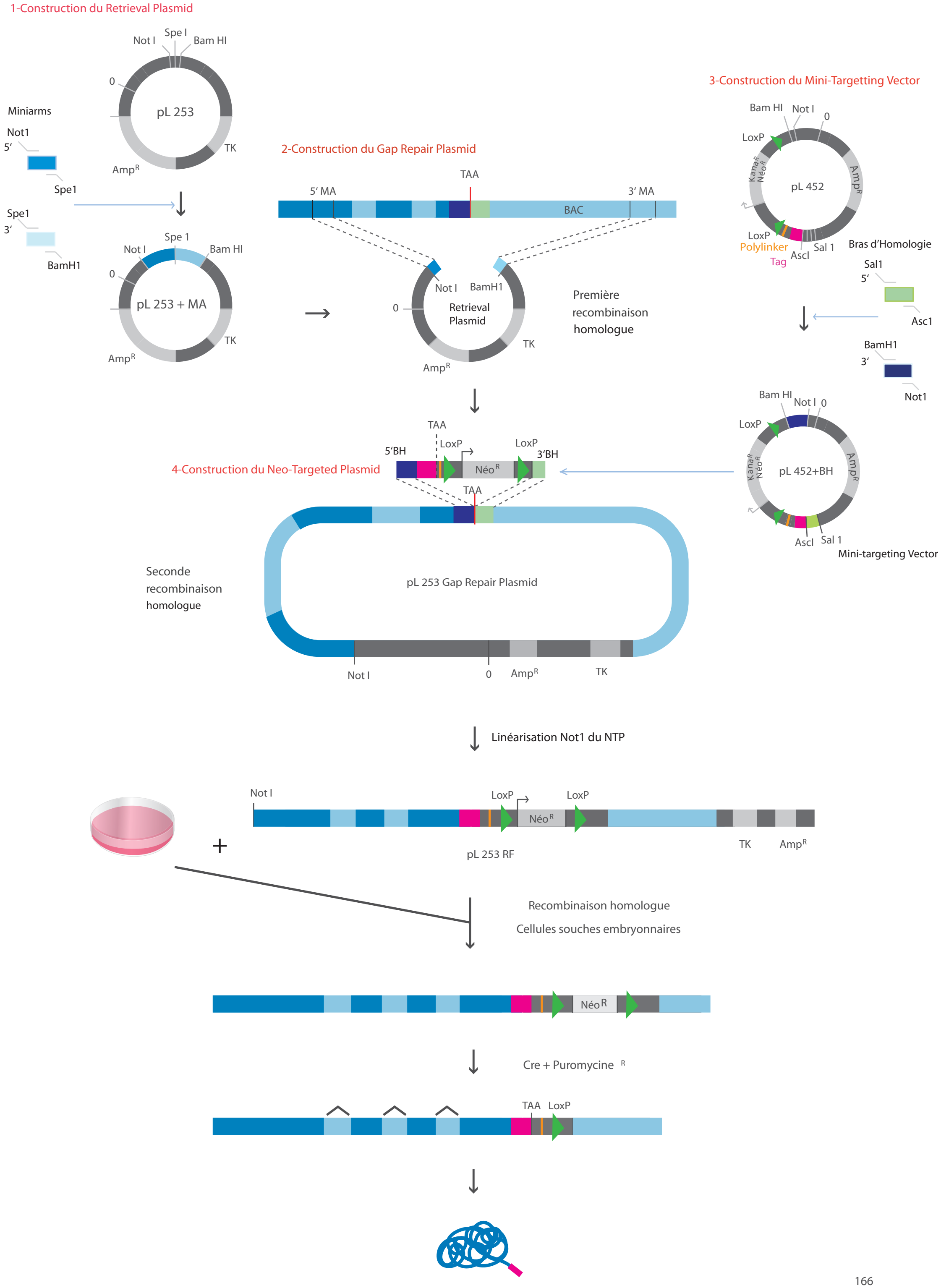


Figure 20. Schéma décrivant le protocole du Recombineering (d'après de Dieuleveult, 2010). Se référer au texte.

Bio-informatique : du bon usage des scripts…

A. Format des Données

Un des problèmes majeurs auquel se trouve confrontée la bioinformatique est la profusion des formats de fichiers rencontrés, parfois peu ou mal définis. Les fichiers stockant les données sont structurées selon le type d'information qu'ils véhiculent. A chaque type de fichier est associée une extension bien particulière, permettant de savoir de quel genre de données l'on dispose.
Les informations décrites ici sont extraites du site UCSC : http://genome.ucsc.edu/FAQ/

1. Fichier bed

Extension : .bed
Format : Fichier texte tabulé, **B**rowser **E**xtensible **D**ata

Ce type de fichier a été développé par UCSC (University of California Santa Cruz). Il est utilisé pour définir des éléments du génome, comme les tables de gènes, de transcrits ou les séquences. Ce format est modulable, il contient au minimum trois colonnes indiquant les coordonnées des éléments (chrom, start, end), les neufs autres colonnes sont optionnelles.

Ce format permet de représenter graphiquement les éléments du génome.

Commentaire :

UCSC numérote la première position sur un génome à 0 et non à 1. Les fichiers BED peuvent donc avoir un décalage de 1 dans les positions 'start' et 'end' selon le logiciel qui l'aura généré. Il faut toujours vérifier le système de numérotation !

Structure générale :
Les trois colonnes BED requises sont :

- **chrom** - Le nom du chromosome (e.g. chr3, chrY, chr2_random).
- **chromStart** – La position de départ de l'élément sur le chromosome. La première base du chromosome est numérotée 0.
- **chromEnd** – La position de fin de l'élément sur le chromosome. La position de fin n'est pas comprise dans l'élément. Par exemple, les 100 premières bases d'un chromosome seront définies comme *chromStart=0, chromEnd=100*, et recouvrent les bases de 0-99.

Les neufs autres colonnes additionnelles sont :

- **name** –nom de l'élément décrit.
- **score** – Un score entre 0 et 1000.
- **strand** – Défini le brin sur lequel est aligné l'élément -'+' or '-'.
- thickStart –
- thickEnd -
- **itemRgb** – Une valeur RGB (**Rouge vert bleu**, abrégé **RVB,** ou **RGB** de l'anglais *red green blue*, format de codage des couleurs, par exemple 255,0,0). Ce code permet de représenter sur le Browser l'élément selon la couleur définie.
- blockCount -
- blockSizes -
- blockStarts -

```
chr1  34491654    34491728    tRNA_Glu_TTC      2701  -     tRNA  type 2
chr1  74863318    74863391    tRNA_Gly_GCC      2703  -     tRNA  type 2
chr1  134930539   134930614   tRNA_Lys_TTT      3709  +     tRNA  type 2
chr1  134930900   134930977   tRNA_Lys_TTT      3708  -     tRNA  type 2
chr1  167571765   167571839   tRNA_Pro_AGG      2714  +     tRNA  type 2
```

Figure 21. Exemple de fichier type BED, table des gènes codant les ARNs de transfert, identifiés chez la souris.

2. Format Wiggle

Extension : .wig
Format : Fichier texte tabulé, **Wig**gle

Le format wiggle est utilisé pour décrire des données de mesures quantitatives, cartographiées sur le génome. Il permet de représenter des données de densité, d'occupation d'une protéine par exemple, ou de la conservation de séquences entre espèces.

Structure générale :

Le format Wiggle est composé de lignes de déclaration et de lignes comportant les données. Il existe deux options pour formater les données : **variableStep** et **fixedStep**. L'option **variableStep** est utilisée pour les données recouvrant des intervalles irréguliers. Ce format commence avec une ligne de déclaration, et suivie de deux colonnes contenant la position sur le chromosome et une valeur (de densité ou autre).

Ce fichier comporte en plus, s'il est destiné à être visualisé dans le browser UCSC, une ligne définissant la piste (track), avec des options variables.

```
variableStep  chrom=chrN  [span=windowSize]
chromStartA  dataValueA
chromStartB  dataValueB
... etc ...  ... etc ...
```

- **La ligne de déclaration** commence avec le mot variableStep, et est suivie du chromosome. Le paramètre « span » indique le pas entre des positions spécifiées, continues pour un intervalle particulier. Le même pas est utilisé pour l'ensemble du fichier Wig.

- **Data Values** – la position du chromosome est spécifié 1-relative (voir fichier bed, la première position du chromosome est numérotée 1). Les positions n'ayant aucune valeur associée sont exclues.

```
track type=wiggle_0 name=chr1   description="Shifted Merged tag counts for
every 1 bp from chip-seq data (by WigMaker)"
variableStep chrom=chr1 span=1
3003581     1
3003582     1
3003583     1
3003584     1
```

Figure 22. Exemple d'un fichier Wig crée à partir des données de séquençage de la lignée non-étiquetées 46C, avec le programme WigMaker, en utilisant un pas de 1.

3. Formats indexés de type binaire : bigBed et bigWig

Ces deux formats sont créés à partir des fichiers BED et WIG, respectivement (Kent et al., 2010). Les fichiers résultants sont indexés et de type binaire. L'intérêt de ces formats réside tout d'abord dans leur taille, beaucoup moins volumineux qu'un fichier de type texte (BED ou WIG). De plus, étant indexés, les informations requises sont extraites plus rapidement.

- Format bigBed

Extension : .bb
Format : Fichier indexé binaire

Le format bigBed est très semblable au format BED, car il contient le même type de données, il est crée à partir d'un fichier BED en utilisant le script UCSC bedToBigBed.

a. Créer un fichier BED
b. Le fichier BED doit être trié selon les chromosomes et la position sur le génome :

Synopsis: sort -k1,1 -k2,2n unsorted.bed > input.bed

c. Créer une table du génome sur lequel vous travaillez. Ce fichier décrit pour chaque chromosome sa taille en bases. Utiliser le script UCSC fetchChromSizes.

Synopsis: fetchChromSizes mm9 > mm9.chrom.sizes

```
chr1 197195432
chr2 181748697
chr3 159599783
chr4 155630120
chr5 152537259
```

Figure 23. Exemple de la table du génome de la souris mm9 des cinq premiers chromosomes.

d. Créer le fichier bigBed en utilisant le script bedToBigBed.

Synopsis: bedToBigBed input.bed chrom.sizes myBigBed.bb

e. Pour visualiser le fichier bigBed, transférer le fichier sur un serveur http, ou ftp.
f. Créer un nouveau custom track, et ajouter cette ligne de description dans la parie configuration :

```
track type=bigBed name="My Big Bed" description="A Graph of Data from My Lab"
bigDataUrl=http://myorg.edu/mylab/myBigBed.bb
```

Attention :

Le programme d'alignement Eland se focalise sur un nombre x de bases, à partir de la gauche de la séquence de la lecture, x étant défini suivant la longueur de la lecture. Cette séquence est appelée « seed ». Par exemple, sur une lecture de 36b, les 32 premières bases constitueront le seed. Lors de l'alignement, Eland cherche à aligner cette séquence. Les paramètres définis par l'utilisateur, comme le nombre maximal d'erreurs autorisées seront appliqué à cette séquence.

Ainsi Eland peut aligner des lectures en fin de chromosomes, et produire des coordonnées aberrantes, car en dehors du chromosome. Ceci a normalement peu d'effet, sauf lorsque de la création des fichiers bigBED, ou bigWIG. Les scripts renvoient un message d'erreurs, indiquant que le fichier contient des coordonnées inexistantes.

Il faut alors nettoyer le fichier d'entrée, en utilisant le script bedClip, celui-ci éliminera toutes les lectures dont les coordonnées sont supérieures à la taille du chromosome.

Synopsis: bedClip input.bed chrom.sizes output.bed

- Format bigWig

Extension : .bw
Format : Fichier indexé binaire

Le format bigWig est, à l'image des fichiers BED et bigBED, semblable au fichier WIG. Il est utilisé pour décrire des données de densité continues. Il est créé à un partir d'un fichier WIG en utilisant le script wigToBigWig.

Pour créer un fichier bigWig :

a. Créer un fichier WIG.
b. Utiliser la même table du génome, que celle créée pour convertir les fichiers bigBed.
c. Créer le fichier bigWig en utilisant le script wigToBigWig :

Synopsis: wigToBigWig input.wig chrom.sizes myBigWig.bw

d. Pour visualiser le fichier bigWig, transférer le fichier sur un serveur http, ou ftp.
e. Créer un nouveau custom track, et ajouter cette ligne de description dans la partie configuration :

```
track type=bigWig name="My Big Wig" description="A Graph of Data from My Lab"
bigDataUrl=http://myorg.edu/mylab/myBigWig.bw
```

4. Formats SAM/BAM

Le format SAM ou Sequence Alignment Map est un format générique donnant tous les détails de l'alignement. Le format BAM est la représentation binaire du format SAM, et contient les mêmes informations.

```
GA8-EAS671_0017_FC:2:1:4630:934#0/1 0     chr17 32355485    255   36M   *     0
      0     NTGGGAGACAAGTATAACAGCATGGAAGACGCCAAG
      BKIIFMNNMM___bbb_b_____b__b____bbb__      XA:i:1      MD:Z:0C35   NM:i:1
GA8-EAS671_0017_FC:2:1:4652:940#0/1 16    chr9  12617496    255   36M   *     0
      0     TCTAGTCAGGAAGCAAAGAGAGGCGAGTGTTGGTGN
      _______b___b_b_b___b__bbb_LLNMLHJKIB      XA:i:1      MD:Z:35G0   NM:i:1
GA8-EAS671_0017_FC:2:1:4692:936#0/1 0     chr3  112860094   255   36M   *     0
      0     NGAAGCAGATCGACTTGTTTCGTATTCTTAAAAACC
      BOKKKQQTQQ_________QQ_________________      XA:i:1      MD:Z:0G35   NM:i:1
```

Figure 24. Exemple d'un fichier de type SAM, contenant les données de d'alignement issues du ChIP de l'ARN Polymérase II, chez le rat.

5. Format fastq

Extension : il n'y a pas d'extension standard, on peut utiliser .fastq.
Format : Fichier texte

Ce format est adapté du format FASTA, utilisé pour décrire des séquences. Il est émis par le programme GERALD, contenu dans la pipeline d'analyse d'Illumina. Ce format contient des informations sur le séquenceur, la séquence des lectures, et des indications sur leur qualité.

```
@HWI-EAS341_3_303N2AAXX:2:1:902:1664
GTGTGGGGGACTTTTGGGATAGGATA
+HWI-EAS341_3_303N2AAXX:2:1:902:1664
ZZZZYZZZYZZZZZZZYYZZYYYZZY
```

Figure 25. Exemple d'un fichier de type fastQ, contenant les données de séquençage de la protéine RPC4 (lecture de 25 bases).

Les données issues de séquençage Paired-end contiennent en plus des informations sur les paires de lectures. Les lectures sont présentées dans l'orientation dans laquelle elles sont alignées sur le génome de référence (5'-3'_3'-5'), qui est en fait l'orientation dans laquelle elles ont été séquencées.

```
@GA8-EAS671_0008_FC:7:1:1186:944#0/1
NGACATTTGCTCTTTGTATGTTACTATAAATTTCCATGATAAAACTTGAAAACCCCATGCATGACAGGGTTGGCGAAA
GACATTACAAAGTCCATTGTGCT
+GA8-EAS671_0008_FC:7:1:1186:944#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBB
```

Figure 26. Exemple d'un fichier de type fastQ, contenant les données de séquençage de la protéine TCEA1 (lecture de 101 bases, en paired-end).

Ligne 1 : débute par @. Elle donne l'identifiant de la séquence, et diverses descriptions comme :

GA8-EAS671_0008_FC	Nom unique de l'instrument
7	Ligne sur la Flowcell
1	Numéro du « tile » ou région au sein de la ligne de la flowcell
1186	'coordonnée-x' du cluster à l'intérieur du tile
944	'coordonnée-y' du cluster à l'intérieur du tile
#0	Numéro d'index si séquençage multiplex (0 sinon)
/1	Numéro du lecture /1 ou /2 si séquençage paired-end

Tableau 6. Identifiant Illumina des séquences de TCEA1.

Ligne 2 : séquence de la lecture.
Ligne 3 : débute normalement par +, suivi de l'identifiant donné à la ligne 1.
Ligne 4 : code donnant la qualité de la séquence, elle contient le même nombre de signes que la séquence.

Qualité des lectures.

La qualité des lectures codée en caractères ASCII, est définie suivant le score dit « Phred ». Ce score diffère selon les versions du Genome Analyser, et de la pipeline Casava (Illumina) utilisées (http://en.wikipedia.org/wiki/Phred_quality_score). Il indique la probabilité que la base définie au cours du séquençage soit incorrecte.

B. Traitement des Données

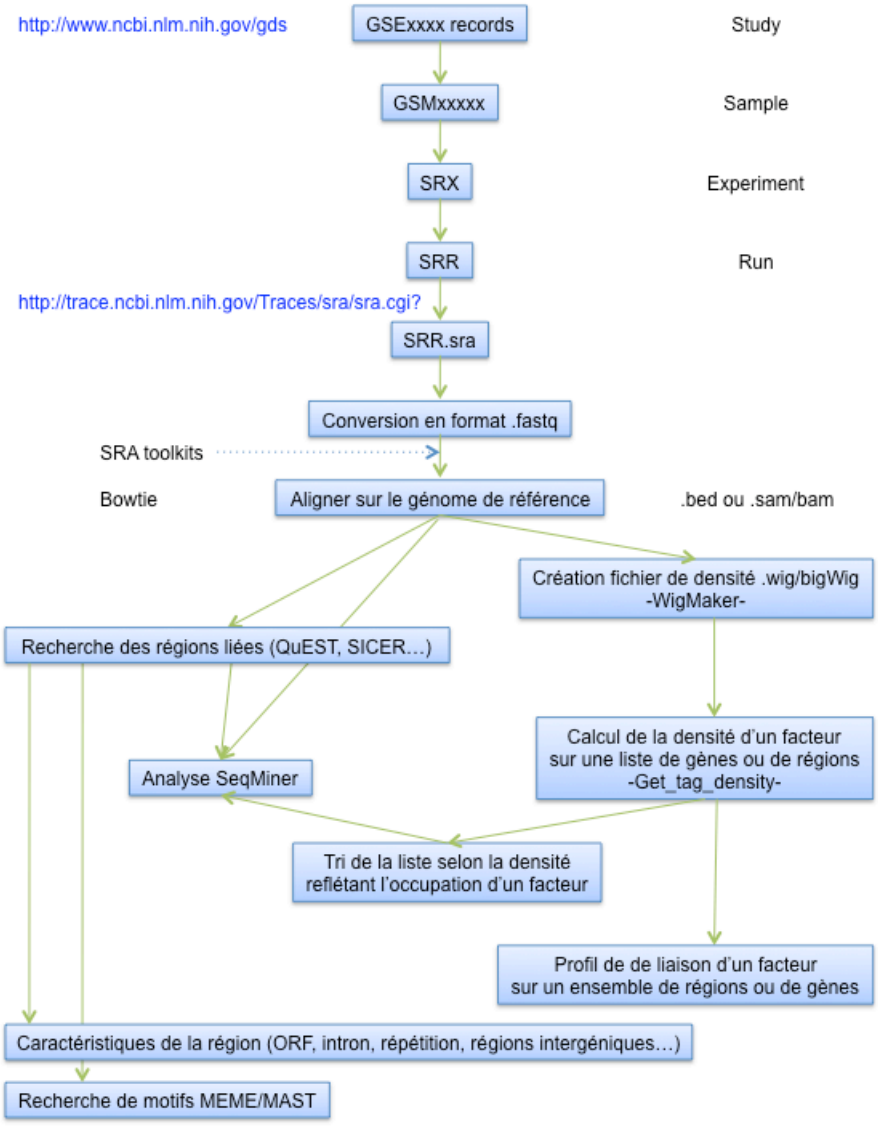


Figure 27. Description d'une analyse classique de ChIP-seq.
Cet exemple présente les différentes étapes réalisées lors d'une analyse des données de ChIP-seq, issues du laboratoire, ou disponible dans la base de données SRA ou GEOdatasets.

L'analyse standard des données de ChIP-seq se décompose brièvement en deux parties. Premièrement, les sites génomiques présentant un enrichissement significatif en lectures sont identifiés et annotés selon les caractéristiques des régions liées, déjà connues (classe du gène, introns, éléments répétés, régions inter-géniques…). Puis, afin d'aborder les règles biologiques définissant les sites de liaison, l'approche commune consiste à rechercher les sites de co-occurrence avec d'autres facteurs. En définissant les partenaires potentiels du facteur, c'est-à-dire ayant les mêmes sites d'occupation, on cherche à mieux cerner la fonction de la protéine.

Ce problème peut être abordé sous trois angles, par la recherche de corrélation existante entre deux protéines à des sites bien définis, en estimant la proximité des sites de liaison entre différents facteurs, ou encore, par seqMINER (Ye et al., 2011), en distinguant les sites de liaison selon leur co-occurrence avec d'autres facteurs.

Afin de mieux définir le rôle des protéines étudiées, et qualifier les sites d'occupation, il faut rechercher quels autres facteurs peuvent se lier aux mêmes régions. Il faudra alors cumuler d'autres données issues de ChIP-seq. Comme il n'est pas possible de réaliser au laboratoire toutes ces expériences, la littérature permet d'accéder à une multitude de données. Il va sans dire, qu'il faut utiliser des résultats de séquençage effectués dans le même organisme, et le même type cellulaire.

1. Partage des données

Deux possibilités sont offertes ; soit passer directement par la base de données SRA regroupant l'ensemble des données de séquençage à haut-débit, ou utiliser la base de données Geodataset.

- Base de Données et format SRA
(NCBI handbook http://www.ncbi.nlm.nih.gov/books/NBK7522/)

Le NCBI maintient la base de données SRA ou Short Lecture Archive (www.ncbi.nlm.nih.gov/Traces/sra/) comme dépôt des données issues des technologies de séquençage à haut-débit (454 de Roche, Illumina, SOLiD d'Applied Biosystems…).

Les données de SRA sont classifiées, comme dans GeoDataset, selon une hiérarchie d'études (Studies, SRP), d'expérience (Experiment, SRX) et d'échantillons et leurs « runs » correspondants (Samples, SRR, R pour Run).

Une étude regroupe une ou plusieurs expériences ayant une finalité commune. Une expérience décrit spécifiquement ce qui a été séquencé, et détaille la méthode utilisée. Chaque expérience est composée de plusieurs runs. Un run contient les résultats, ou lecture pour chaque échantillon, un échantillon pouvant être associé à plusieurs runs s'il a été séquencé plusieurs fois. Un numéro unique d'accession est attribué à chacun.

Au moment de la soumission, les données ont été converties en format SRA. SRA est un format d'archive de l'INSDC (International Nucleotide Sequence Databases Collaboration, entre DDBJ,, DNA DataBank of Japan, ENA, European Nucleotide Archive, and GenBank). Les données accessibles sur GeoDataset et sur SRA doivent donc être converties en un format utilisable, comme le format fastq. Il existe un ensemble d'outils, SRAtoolkits, permettant ces conversions (SRA handbook **http://www.ncbi.nlm.nih.gov/books/NBK47528/)**

Convertir le format.SRA en .fatsq

Programme : fastq-dump

Synopsis : fastq-dump -A <SRR_accession> <Path_to_SRR_Directory>

Récupérer les données sur internet

Ici est décrite la méthode pour télécharger les données à partir de Geodatasets.

a. Aller sur le site du NCBI : http://www.ncbi.nlm.nih.gov/
b. Dans le menu déroulant, sélectionner GEO (Gene Expression Omnibus) datasets, et rechercher les données issues d'expérience de ChIP-seq (mot-clé = ChIP-seq).
c. Sélectionner l'organisme qui vous intéresse (par exemple, *Mus musculus*)
d. Rechercher les données d'intérêt, et télécharger le format SRA.
e. Convertir le format SRA en fastq.
f. Aligner les nouvelles données en utilisant Bowtie.

2. Aligner les lectures sur le génome de référence

Qu'il s'agisse d'aligner ses propres données ou celles récupérés sur Internet, via geodataset ou SRA (NCBI), Eland n'est pas accessible, car c'est un outil commercial, mis en place par Illumina. Nous avons choisi d'utiliser Bowtie (Langmead et al., 2009), même s'il existe d'autres programmes tels que MAQ.

Bowtie est un outil d'alignement de lectures de petites tailles (issus de séquençage à haut-débit) sur de larges génomes. Bowtie aligne les lectures suivant une combinaison de différentes options déterminant quels alignement sont autorisés, et lesquels doivent être reportés.

Synopsis : bowtie -a -m1 --best -strata -v2 -p3 /home/BP_DATA/genomes/mm9/bowtie/mm9 file.fastq file.align

Paramètres utilisés :

- **/home/BP_DATA/genomes/mm9/bowtie/mm9** : chemin d'accès au génome version NCBI37 (ou mm9) de la souris, indexé selon Burrow-Wheelers utilisable par Bowtie.

- **a** :Bowtie reporte tous les alignements valides.
- **v** : <int > nombre maximal d'erreurs autorisées sur l'ensemble du lecture. Le mode –v ne prend pas en compte la qualité de la lecture. Il faut bien distinguer ce mode du paramètre –n, qui comme Eland ne considère qu'une partie de la séquence dite « seed » pour l'alignement. En général, deux erreurs maximales sur l'ensemble de la séquence sont autorisées. Ce paramètre ne prend pas en compte la qualité de la séquence.
- **strata** : les alignements sont classés suivant les critères imposés comme le nombre d'erreurs. Si --strata est spécifié, alors -- best doit également être ajouté.
- **best** : Bowtie reporte le meilleur alignement suivant les critères imposés.
- **m** : <int> supprime les alignements reportés s'il y en a plus que le <int> défini. Nous avons fixé -m1, seuls les lectures s'alignant de façon unique seront retenus.
- **p** : <int> nombre de processeurs utilisés en parallèle par le programme, ici -p3.

Combinaison des différents paramètres :

En clair, la combinaison de ces différents paramètres permet de reporter les alignements des lectures s'alignant de manière unique, avec 0, 1 ou 2 erreurs. Le meilleur alignement, contenant le plus faible nombre d'erreur sera retenu.

- Pour les séquençages Paired-end :

Le séquençage en paired-end (par opposition au séquençage « single read») génère des lectures issues des deux extrémités de chaque molécule d'ADN de la librairie. Les séquences obtenues sont appelées « paired-end read». Après séquençage de la première lecture, les matrices sont régénérées *in situ*, afin de séquencer l'extrémité opposée de la matrice. Les brins qui viennent d'être synthétisés, au cours de l'étape de séquençage sont éliminés, et le brin « reverse » complémentaire à la matrice d'origine (brin « forward ») ayant servi au premier séquençage est amplifié, toujours par « bridge-PCR », pour former des clusters. Les matrices d'origine sont ensuite clivées et éliminées, le brin reverse subit à son tour l'étape de séquençage.

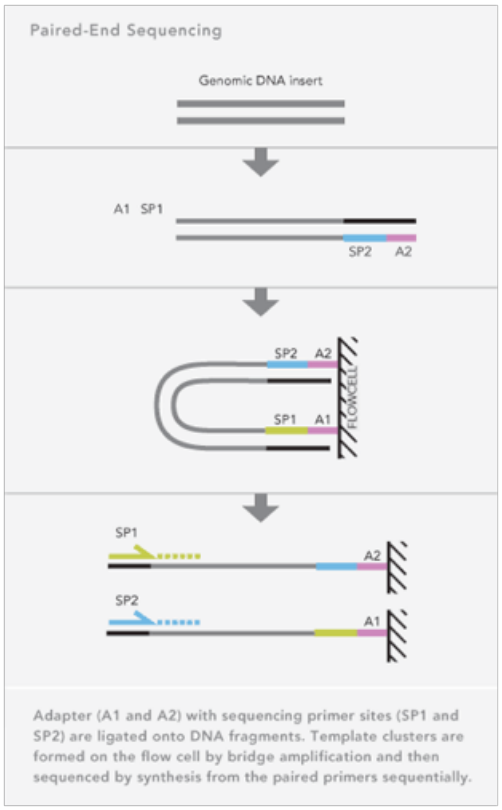


Figure 28. Schéma décrivant la méthodologie du « paired-end » par Illumina.

En plus des informations liées aux deux séquences partenaires, la taille de l'insert ou « linker », partie non séquencée entre ces deux lectures peut être estimée. Cette information est très importante, car elle permet de positionner les deux lectures l'une par rapport à l'autre, suivant la taille de cet insert, lors de l'alignement,.

Deux fichiers de séquences sont obtenus à la fin du séquençage, contenant les lectures 1 et 2, chacune issue du séquençage d'une des deux extrémités.

Bowtie peut aligner les lectures « paired-end », si des paramètres spécifiques du Paired-end sont fournis.

- Chaque lecture est alignée séparément suivant les paramètres –v (ou –n)
- L'orientation relative et la position des deux lectures doivent ensuite satisfaire des contraintes fixées par les paramètres –I/-X.

`-I <int>` : la taille minimale de l'insert entre les deux lectures.
`-X <int>` : la taille maximale de l'insert.

Relative Orientation Statistics

F-: > R2 R1 >	F+: > R1 R2 >	R-: < R2 R1 >	R+: > R1 R2 <	Total
44574 (0.3%)	42792 (0.3%)	773135 (4.9%)	**14880865 (94.5%)**	15741366

R = Lecture

Ce tableau donne les alignements réalisés avec les orientations possibles des deux lectures pour TCEA1. Seul l'orientation R+ (>R1 R2<) est correcte.

Insert Size Statistics
(for relative orientation R+)

Median	Below-median SD	Above-median SD	Low thresh.	High thresh.
140	12	19	104	197

Ce tableau résume les caractéristiques moyennes de taille de l'insert. Ces caractéristiques sont utilisées pour l'alignement par Bowtie, I = 104, et X = 197. La taille médiane peut également être utilisée.

Insert Statistics
(% of individually uniquely alignable pairs)

Too small	Too large	Orientation and size OK
75788 (0.5%)	688667 (4.4%)	**14116410 (89.7%)**

Figure 29. Statistiques issues de l'analyse par Eland-Casava du séquençage de TCEA1.

A l'issue de l'alignement, Bowtie renvoie un certain nombre de statistiques :

```
bowtie -a -m1 --best --strata -p3 -v2 /home/BP_DATA/genomes/mm9/bowtie/mm9
C53-chanel_2.txt C53-chanel_2.align
# lectures processed: 6113915
# lectures with at least one reported alignment: 4297020 (70.28%)
# lectures that failed to align: 365601 (5.98%)
# lectures with alignments suppressed due to -m: 1451294 (23.74%)
Reported 4297020 alignments to 1 output stream(s)
```

Figure 30. Statistiques à l'issue de l'alignement des séquences de RPC4.

Si aucun format de sortie n'a été précisé, le fichier d'alignement contient, par défaut, les colonnes suivantes :

a. le nom de la lecture alignée.
b. brin + (forward) ou – (reverse) sur lequel est aligné la lecture.

c. chromosome.
d. coordonné en 0-basé de l'extrémité gauche de la lecture, indiqué sur le brin référence forward.
e. séquence de la lecture.
f. séquence version qualité de la lecture.
g. description des alignements avec 0, 1 ou 2 erreurs. Cette colonne est vide si l'alignement reporté ne contient pas d'erreur.

3. Création d'un fichier BED

Suite à l'alignement des lectures, il est nécessaire de créer un fichier de type BED, qui est utilisé en entré de nombreux programmes, tels seqMINER (Ye et al., 2011), ou WigMaker.

A l'heure actuelle, une certaine harmonisation des formats tend à se mettre en place, le format SAM est de plus en plus utilisé. Ce format peut être obtenu directement à l'issue de l'alignement en spécifiant simplement -S, lors du lancement de Bowtie. Ce format est utilisable par seqMINER, mais pas par WigMaker.

Le fichier BED est constitué des colonnes -c (chromosome) et -d (coordonnée de l'extrémité gauche de la lecture, donnée par rapport au brin référence forward). Artificiellement, la coordonnée de l'extrémité droite de la lecture est créée en ajoutant la longueur du lecture à la coordonnée gauche. De même, les colonnes du nom et du score sont à la discrétion de l'utilisateur. Par exemple, à partir d'un fichier .align (sortie de Bowtie),

Synopsis : awk 'BEGIN {OFS="\t"} {print $4,$5,$5+36,"name","score",$3}' file.align > file.bed

Comparer l'occupation ou dessiner le profil de liaison de différents facteurs nécessitent de quantifier leur taux d'occupation en un site défini. Une manière d'estimer cette intensité de liaison est de calculer la densité des lectures. Cette densité correspond en réalité à un simple décompte des lectures, par intervalle ou « bin ». Ceci passe par la création d'un fichier de densité WIG.

4. Création d'un fichier de densités WIG.

Le nombre de lectures à la position près ou dans un intervalle, selon le bin choisi est calculé. Il ne faut cependant pas oublier le paramètre d'extension des lectures. À l'origine, la chromatine a été fragmentée jusqu'à ce que la majorité des fragments ait une certaine taille. Par la suite, l'ADN a été sélectionné suivant cette taille. Or seule une extrémité de ce même fragment d'ADN sera séquencée. En étendant la lecture, on cherche à recréer de manière artificielle ce fragment d'origine. La lecture est étendue de x bases à son extrémité 3' ou « end », selon le brin sur lequel il est aligné.

Pour les données de littérature, ce paramètre est indiqué dans le protocole du ChIP utilisé. Parfois il est précisé dans la partie traitant de l'analyse informatique, de combien les auteurs estiment devoir étendre leurs lectures.

Le programme utilisé pour créer ces fichiers WIG a été écrit par Tao Ye de l'IGBMC à Strasbourg. D'autres outils sont disponibles comme le script pileup (emplilement) de SAMtools (http://samtools.sourceforge.net/). Il est donc nécessaire de passer par le format SAM (possible à l'issue de l'alignement par Bowtie).

Synopsis : Java -Xmx3000m -jar WigMaker.jar input.bed 300 25

-Xmx3000m : mémoire allouée
300 : Extension des lectures (les lectures sont étendues à partir de leur coordonnée « start »).
25 : bin ou pas, intervalle regroupant les densités.

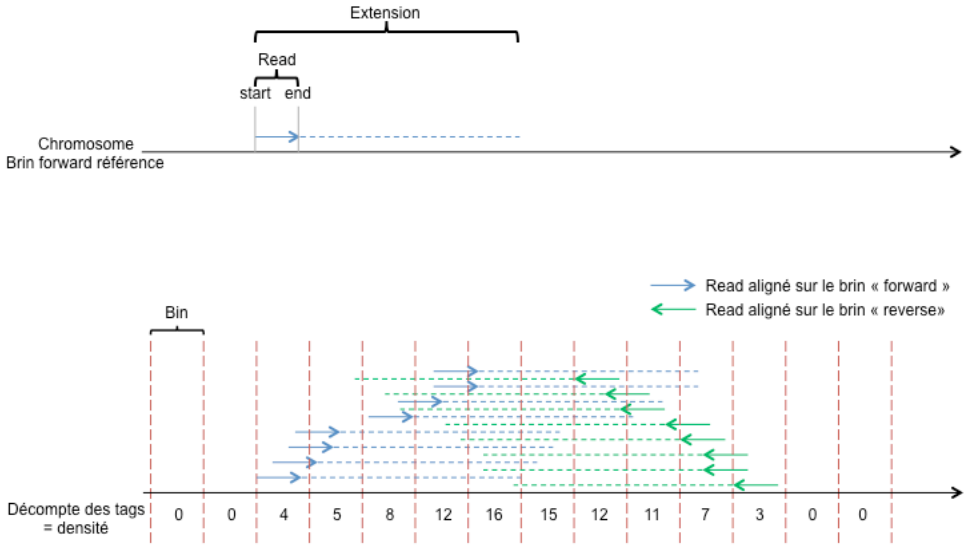


Figure 31. Description de la création d'un fichier de densité WIG.

5. Déterminer les régions liées avec QuEST (Valouev et al., 2008)

http://www.stanford.edu/~valouev/QuEST/QuEST.html

L'analyse des résultats des expériences de ChIP-seq conduit à la recherche des régions enrichies en lectures issues du séquençage de l'ADN immunoprécipité. Ces régions présentant une forte densité en lectures par rapport à un contrôle, de type input ou « mock immunoprécipitation » sont considérées comme liées par le facteur étudié. Différentes méthodes de recherche de ces régions ont été développées, le programme utilisé dans cette étude est QuEST ou Quantitative Enrichment of Sequence Tags.

QuEST utilise les coordonnées génomiques et l'orientation des lectures cartographiées. Dans le cas du séquençage simple ou « single read», les lectures « forward » ou « reverse » proviennent des extrémités opposées du fragment d'ADN immunoprécipité. Le séquençage étant orienté de l'extrémité du fragment vers le milieu, ceci crée une sous-représentation des lectures au centre, site de liaison du facteur.

QuEST construit tout d'abord un profil pour chaque orientation « forward » ou « reverse ». Ces profils reflétant la densité, quantifient l'enrichissement en lectures.

A partir d'un ensemble de profils, considérés comme robustes, car présentant la plus forte densité de lectures, QuEST calcule la distance existante entre les sommets des profils « reverse » et « forward ». La moitié de cette distance correspond au « peak shift ». Cette distance doit être estimée afin de combiner au mieux les deux profils opposés en un seul, le CDP ou Combine Density Profil. QuEST recherche ensuite pour chaque locus enrichi en lecture, ce CDP. Si ces régions sont suffisamment enrichies par rapport au contrôle, alors QuEST note cette région comme potentiellement liée par le facteur étudié.

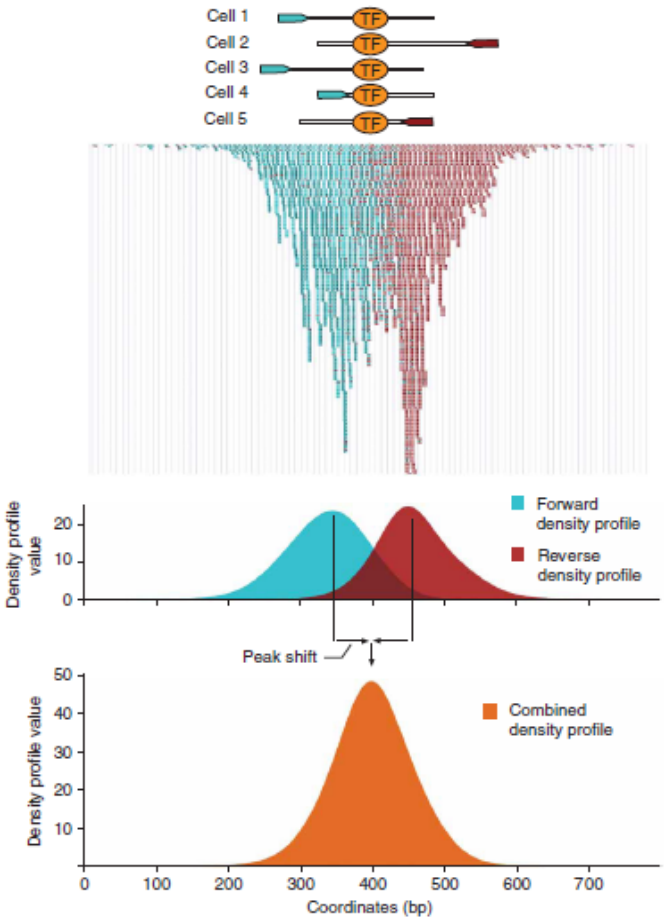


Figure 32. Illustration de la méthodologie de QuEST pour déterminer les régions liées.

Note : Il peut donc exister un décalage entre les sommets des deux profils, dans notre cas, ce décalage est aggravé du fait de la nature répétée des gènes de classe III. Les lectures ne pouvant être cartographiées sur ces sites, car non-uniques sont éliminées, créant ainsi une zone dépourvue de lectures.

Synopsis : generate_QuEST_parameters.pl -QuEST_align_ChIP data.txt -QuEST_align_RX_noIP control_ChIP.txt -gt ./genome_table_mm9 -ap nom_du_répertoire

Définitions :

Format du fichier fourni en entrée, listant les coordonnées génomiques et l'orientation des lectures.
<chromosome> <coordinate> <+/->

Chaque ligne du fichier d'entrée donne l'alignement d'une lecture. Les colonnes sont séparées par des espaces. La coordonnée de la lecture est indiquée par la position de la base située à l'extrémité 5'. +/- spécifie le si l'alignement est réalisé sur le brin sens ou anti-sens. Les coordonnées sont en 0-basée.

-QuEST_align_ChIP <QuEST_align_file> -Lectures du facteur étudié-
-QuEST_align_RX_noIP <QuEST_align_RX_noIP_file> -Lectures du contrôle-
Dans notre cas, le contrôle est l'ADN immunoprécipité des cellules 46C, cellules ne contenant pas de protéine étiquetée.
-gt <genome_table>

-ap <analysis_path> répertoire où les résultats issus de l'analyse QuEST seront reports.

Configuration des paramètres

Au cours de l'analyse par QuEST, le programme demande à l'utilisateur de spécifier des paramètres pour la suite de l'analyse.

```
----------- QuEST parameter configuration -------------

What kind of ChIP experiment is this?

1. Transcription factor with defined motif and narrow (punctate) binding site
   resulting in regions of enrichment 100-300 bp wide.
   (bandwidth = 30 bp, region_size = 300 bp)

2. PolII-like factor resulting in regions of enrichment 300-1000 bp with
   some narrow binding sites and some wide sites possibly occupying
   the entire gene length.
   (bandwidth = 60 bp, region_size = 600 bp)

3. Histone-type ChIP resulting in wide regions of enrichment
   1Kb and up possibly occupying multiple genes.
   (bandwidth = 100 bp, region_size = 1000 bp)

4. Neither, I would like to configure individual parametrs myself.
```

<bandwidth> : ce paramètre servira à la création des profils de densité, suivant l'approche de la densité de Kernel.
<region size> : fenêtre utilisée pour rechercher les loci enrichis en lectures.

```
Let's now determine QuEST peak calling parameters.(Parameters below are
optimized for human and mouse data when default stack collapsing is
performed).

Choose one of the following options:

1. Stringent peak calling parameters.
   ChIP enrichment                 = 50
   ChIP to background enrichment   = 3
   ChIP extension enrichment       = 3

2. Recommended peak calling parameters.
   ChIP enrichment                 = 30
   ChIP to background enrichment   = 3
   ChIP extension enrichment       = 3

3. Relaxed peak calling parameters.²
   ChIP enrichment                 = 10
   ChIP to background enrichment   = 3
   ChIP extension enrichment       = 3

4. Neither, I want to specify peak calling parameters myself.
```

Brève description des résultats

QuEST fournit en sortie une liste de régions définies suivant les critères imposés. A chaque région, sont associés un ou plusieurs pics P, donnant la position de chaque sommet du profil de densité. Différents résultats décrivant l'enrichissement des lectures par rapport au contrôle sont fournis afin d'estimer la validité de la région.

```
R-1 chr12 70458925-70462343 ChIP: 723.328 control: 9.50439 max_pos: 70461842
  ef: 76.1046 ChIP_tags: 941 background_tags: 28 tag_ef: 27.6282 ps: -84 cor:
  0.167549 -log10_qv: 10476.4 qv_rank: 3
P-1-1 chr12 70459382 ChIP: 12.6099 control: 0 region: 70458925-70462344 ef:
  12.6099 ps: -92 cor: 0.32825 -log10_qv: 8.01625 qv_rank: 11374

R-<region_id> <chromosome> <region start>-<region end> ChIP: <ChIP enrichment
  intensity of the strongest peak> control: <control enrichment intensity at
  the position of the strongest ChIP peak> max_pos: <coordinate of the
  highest peak in the region> ef: <normalized enrichment fold at the position
  of the strongest peak>
P-<region_id>-<peak_id> <chromosome> <peak coordinate> ChIP: <ChIP enrichment
  intensity> control: <control enrichment intensity> region: <region start>-
  <region end> ef: <normalized_enrichment fold>
```

Exemple de fichier listant les régions obtenues à l'issue d'une analyse QuEST, pour le facteur TFIIIC220, et explication des données.

6. Extraction des données de densité

Programme : Get-tag-density

Ce script a été écrit par Olivier Alibert de l'IRCM, à Evry. Il permet d'extraire les données de densités d'une région, par intervalle, et de calculer la moyenne des densités sur cette région.

Ces informations de densité comme la moyenne seront utilisées pour comparer l'occupation de plusieurs facteurs. Les densités par intervalle permettront de dessiner le profil d'occupation d'un facteur sur les sites qu'il occupe (voir profil de liaison).
Synopsis : get_tag_density [-h] [-g <GENOME>] [-c <ChIP>] [-d <BIGWIG_DIRECTORY>] <INPUT_FILE>

<BIGWIG_DIRECTORY> : fichiers de densité de type bigWig. Ce programme est très rapide car il utilise le format indexé des fichiers de densité WIG, pour récupérer uniquement les données qui nous intéressent.
Les fichiers qui ont été créés par WigMaker doivent être transformés en bigWig (voir wigTobigWig).
<INPUT_FILE> : Un fichier de type BED contenant les coordonnées des régions pour lesquelles on souhaite récupérer les données de densité.

```
#!/usr/bin/perl

=pod

=head1 NAME

get_tag_density
```

```
=head1 DESCRIPTION

Renvoie la densite des tags (a 1bp) pour une ChIP et une liste de régions
données (au format bed).

chrom   start   end     name    score   strand

=head1 SYNOPSYS

get_tag_density  [-h]  [-g  <GENOME>]  [-c  <ChIP>]  [-d  <BIGWIG_DIRECTORY>]
<INPUT_FILE>


<GENOME> nom du genome : mm9  hg18  hg19 ...

<ChIP> nom d'un expérience de ChIP pour laquelle des données de densités a 1bp
sont disponibles sous la forme de fichiers bigWig organisés par chromosome
(chr1.bw , chr2.bw ...etc) et rassemblés dans un même dossier lui-même situé
dans l'arborescence reserve au <GENOME> (par defaut $BP_DATA/<GENOME>/genome).

<BIGWIG_DIRECTORY>  le nom du répertoire contenant les fichiers bigWig peut
aussi être fourni à la place de <GENOME> et de <ChIP>.


=head1 EXAMPLES

Pour afficher le contenu, les densités en tags C53 des régions du fichier
toto.bed :

   get_tag_density -g mm9 -c C53 toto.bed

ou

   get_tag_density -d $BP_DATA/mm9/genome/C53_1bp toto.bed


=cut

use Getopt::Std;
#use strict;
use POSIX qw(ceil floor);

sub help {
        system ("pod2text $0");
}


my $bwdir='';
my $genome='mm9';
my $chip='';

getopts('hc:g:d:l');

if ($opt_l)
{
        system("ls -l $ENV{BP_DATA}/genomes/${genome}/density");
      exit(1);
```

```
}

if ($opt_h || $#ARGV<0)
{
        &help;
        exit(1);
}

if ($opt_g)
{
      $genome=$opt_g;
}

if ($opt_c)
{
      $chip=$opt_c;
}

if ($opt_d)
{
      $bwdir=$opt_d;
}

if ($bwdir)
{
      if (! -d $bwdir)
      {
            die "ERROR: no directory $bwdir\n";
      }
}
else
{
      if (! $genome )
      {
            die "ERROR: no genome\n";
      }
      elsif (! -d "$ENV{BP_DATA}/genomes/${genome}")
      {
            die "ERROR: unknown genome $genome\n";
      }

      if (! $chip )
      {
            die "ERROR: no ChIP\n";
      }
      elsif (! -d "$ENV{BP_DATA}/genomes/${genome}/density/${chip}_1bp")
      {
            die "ERROR: unknown ChIP $chip for genome $genome\n";
      }
      else
      {
            $bwdir="$ENV{BP_DATA}/genomes/${genome}/density/${chip}_1bp";
      }
}


my $chrom;
my $start;
```

```
my $end;
my $name;
my $score;
my $strand;
my $cpt;

my $file=shift @ARGV;

if (! -f $file)
{
    die "ERROR: no file $file";
}
elsif ($file=~/\.gz$/)
{
    open(INPUT,"zcat $file |");
}
else
{
    open(INPUT,"$file");
}


my $tmpfile="/tmp/get_tag_density.".$$;

@fields=("chr","start","end","q_id","score","strand","density_values","average
","sum","minimum","maximum","max_positions","median_max_position");
printf "#%s\n",join("\t",map {uc($_)} @fields);

while (<INPUT>)
{
      s/(\r|\n)//g;
        ($chrom, $start, $end, $name, $score, $strand)=split(/\t/,$_);

        # Si nécessaire on remplit les colonnes name et score
        $name=++$cpt if (!$name);
        $score=0 if (!$score);

        printf join("\t",$chrom,$start,$end,$name,$score,$strand);

        # ATTENTION bigWigToWig attend un start "0 based" (comme dans le
.bed)!!!!!
      my  $cmd=sprintf("bigWigToWig -chrom=%s -start=%s -end=%s %s/%s.bw
%s",$chrom,$start,$end,$bwdir,$chrom,$tmpfile);

      #print STDERR $cmd."\n";

        open(BIGWIG,"$cmd |");
      while (<BIGWIG>)
      {
            print $_;
      }
      close(BIGWIG);

      my @density=();
      my $max=0;
      my $min=0;
      my $sum=0;
      my @maxpos=();
```

```
      #print STDERR $tmpfile."\n";


        # On initialise les valeurs de la densité a 0

      my $nbp=$end-$start;

        foreach my $p (0..$nbp-1)
      {
            $density[$p]=0;
      }

      open(WIG,"$tmpfile");
      $first=1;
      while (<WIG>)
      {
            if (/^variableStep/)
            {
                  if (! /span=1(\s|$)/ )
                  {
                        die    "ERROR:    bad    span    for    bigWig    file
${bwdir}/${chrom}.bw \n";
                  }
            }
            elsif (/\S+/)
            {
                  s/(\n|\r)//g;
                  @c=split(/\t/,$_);

                  if ( $first || ($max<$c[1]) )
                  {
                        $max=$c[1];
                        @maxpos=();
                        push(@maxpos,$c[0]);
                  }
                  elsif ($c[1] == $max)
                  {
                        push(@maxpos,$c[0]);
                  }
                  $min = ( (!$first) && ($min<$c[1]) ) ? $min : $c[1];
                  $sum = $sum + $c[1];
                  $density[$c[0]-($start+1)]=$c[1];
                  $first=0;
            }
      }
      close(WIG);
      if ($strand eq '+')
            {
            printf
"\t%s\t%.2f\t%s\t%s\t%s\t%s\t%s\n",join(";",@density),$sum/$nbp,$sum,$min,$max
,join(';',@maxpos),$maxpos[floor(($#maxpos+1)/2)];
            }
      else
            {
             printf    "\t%s\t%.2f\t%s\t%s\t%s\t%s\t%s\n",join(";",    reverse
@density),$sum/$nbp,$sum,$min,$max,join(';',reverse
@maxpos),$maxpos[floor(($#maxpos+1)/2)];
            }
```

```
}
close(INPUT);

unlink "$tmpfile";
```

```
chr6  128749416   128749609   4.5SRNA     2093  -
      442;444;447;450;451;451;451;451;451;451;451;451;451;451;451;451;451;
451;451;451;451;451;451;451;451;452;452;453;453;453;453;454;454;454;456;457;45
8;458;458;460;460;461;461;461;461;463;463;463;463;466;466;466;466;466;466;466;
466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;46
6;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;
466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;466;469;46
9;468;469;469;469;469;469;470;470;470;470;469;469;469;469;469;469;469;469;
468;467;467;467;467;464;462;460;460;456;450;445;441;435;431;431;430;427;427;42
7;427;427;427;427;427;427;427;427;427;425;419;419;413;411;408;408;408;408;
407;407;407;406;406;405;405;402;401;399;399;397;392;387;386;383;380;373;372
      449.98      86847 372   470   128749487;128749486;128749485;128749484
      128749486
```

Exemple de fichier tabulé obtenu à la sortie de get_tag_density.
Ici, les densités de la protéine RPC4 sont reportées pour le gène codant l'ARN 4,5 S, dont les bornes ont été étendues de 50 bp de part et d'autres.

`chr6 128749416 128749609 4.5SRNA 2093 -` : coordonnées extraites du fichier BED, indiquant les coordonnées du gène de l'ARN 4,5S.
`442;444;447;450…` : densité de RPC4 par nucléotide, sur l'ensemble de la région définie en entrée (bornes du gène 4,5S étendues de 50 bp de part et d'autres).
`449.98` : moyenne des densité sur la région.
`86847` : somme des densités sur la région.
`372` : densité minimale.
`470` : densité maximale.
`128749487;128749486;128749485…` : positions génomiques où la densité est maximale.
`128749486` : moyenne des positions de densité maximale.

7. Profil de liaison

Pour dessiner le profil de densité d'une protéine sur une région, il faut récupérer les données densités données pour chaque position (colonne `@densité`). Les densités sont calculées sur un intervalle autour de la coordonnée « start » (TSS) du gène.

Synopsis : calculate_density_profil file.bed file.bw

```
#!/bin/bash

variable=${1%.bed}

##définition des coordonnées +/- 500bp

awk 'BEGIN {OFS="\t"} ($6 ~ /+/) {print $1,$2-1000,$2+1000,$4,$5,$6} ($6 ~ /-
/) {print $1,$3-1000,$3+1000,$4,$5,$6}' "$variable".bed > "$variable"_tss

for i in $2

do
```

```
get_tag_density_oriente -c "${i}" -g mm9 "$variable"_tss | sed '1d' > Density_for_"${i}"_"$variable"_tss

##récupérer les valeurs de densité

get_density_values    Density_for_"${i}"_"$variable"_tss | cut -f4,7 | sed 's/\;/\t/g' > density_values_for_"${i}"_"$variable"_tss

done
```

$variable : fichier BED définissant les régions.
$2 : fichier bigWIG (file.bw) du facteur étudié, mis en deuxième position.

Programme : get_density_value

```
#!/usr/bin/perl

my $chr;
my $start;
my $end;
my $q_id;
my $score;
my $strand;
my $density_values;
my $average;
my $sum;
my $minimum;
my $maximum;
my $max_positions;
my $median_max_position;

while (<>)
{
        ($chr, $start, $end, $q_id, $score, $strand,$density_values, $average, $sum, $minimum, $maximum, $max_positions,$median_max_position)=split(/\t/,$_);
            printf join("\t",$chr,$start,$end,$q_id,$score,$strand,$density_values,"\n");
```

Une fois les données de densités stockées dans un fichier, la représentation graphique de ces données est effectué sous R.

```
#1-charger les fichiers de densités

c53<-lecture.table('density_values_for_C53_tRNA_positifs_tss', sep='\t')
c155<-lecture.table('density_values_for_C155_tRNA_positifs_tss',sep='\t')
brf1<-lecture.table('density_values_for_Brf1_tRNA_positifs_tss', sep='\t')
brf2<-lecture.table('density_values_for_Brf2_tRNA_positifs_tss',sep='\t')
tbp<-lecture.table('density_values_for_TBP_tRNA_positifs_tss', sep='\t')

#2-calculer la moyenne des densités pour l'ensemble de la liste, par position (ou intervalle)

c53m=mean(c53[,2:1001])
c155m=mean(c155[,2:1001])
brf1m=mean(brf1[,2:1001])
```

```
brf2m=mean(brf2[,2:1001])
tbpm=mean(tbp[,2:1001])

#3-définir les positions de l'axe des abscisses

x=c(-500:-1,1:500)

#4-sauvegarder le (ou les) plot dans un fichier .pdf

pdf('bound_tRNA.pdf')

#5-Définir les paramètres graphiques, et « ploter » les valeurs de densité de
chaque facteur

par(cex.axis=1.5,cex.lab=1.5,cex.main=1.5,oma=c(0,2,0,0))
plot(x,c53m,type='l',col='gold',xlab="RNA                                    5'
end",ylab='',ylim=c(0,700),yaxt='n')
lines(x,c155m,col='blue')
lines(x,brf1m,col='red')
lines(x,brf2m,col='green3')
lines(x,tbpm,col='brown')
axis(2,las=2)
mtext(side=2,line=4,'Tag density',cex=1.5)
title('Bound tRNA genes',font.main=2)

legend("topright",  legend  =  c("RPC4","RPC1","BRF1","BRF2","TBP"),  col  =
c("gold","blue","red","green3","brown"),lwd=c(1.5),bty="o",cex=1)

dev.off()
```

8. Site de liaison (ou maximum de densité)

Pour déterminer la position représentant le pic de densité de la protéine, ou le site de liaison de la protéine, la moyenne de la distance de la position représentent ce maximum de densité par rapport à la coordonnée « start », est calculé pour un ensemble représentatif de gènes. Puis en utilisant la densité de Kernel, sous R, la position représentative est identifiée.

```
#!/bin/bash

#1-Fichier BED définissant les régions
variable=${1%.bed}

#2-Extension des bornes de 50bp autour de la coordonnée "start"
awk 'BEGIN {OFS="\t"} {print $1,$2-50,$2+50,$4,$5,$6}' "$variable".bed >
"$variable"_50nt-extended

for i in "$2"

do

get_tag_density_oriente -c "${i}" -g mm9 "$variable"_50nt-extended | sed '1d'
> Density_for_"${i}"_"$variable"_50bp_extended


#3-Calcul de la position par rapport au TSS
```

```
get_max_position Density_for_"${i}"_"$variable"_50bp_extended  >
max_postion_for_"${i}"_"$variable"_50bp_extended.intermediate

paste "$variable".bed
max_postion_for_"${i}"_"$variable"_50bp_extended.intermediate | cut -f1-6,15>
max_postion_for_"${i}"_"$variable".txt

awk 'BEGIN {OFS="\t"} ($6 ~ /+/) {print $1,$2,$3,$4,$5,$6,($7-$2)} ($6 ~ /-/)
{print $1,$2,$3,$4,$5,$6,($3-$7)}' max_postion_for_"${i}"_"$variable".txt  >
"${i}"_peak_relative_to_TSS_of_"$variable
".txt

done
```

Programme : get_max_position
But : récupérer la colonne donnant la position où la densité est maximale (@median_max_position).

```
#!/usr/bin/perl

my $chr;
my $start;
my $end;
my $q_id;
my $score;
my $strand;
my $density_values;
my $average;
my $sum;
my $minimum;
my $maximum;
my $max_positions;
my $median_max_position;

while (<>)
{
        ($chr, $start, $end, $q_id, $score, $strand,$density_values, $average,
$sum, $minimum, $maximum, $max_positions,$median_max_position)=split(/\t/,$_);
        if ($sum==0)
           {
           printf join("\t",$chr,$start,$end,$q_id,$score,$strand,"NA","\n") ;
           }
      elsif (!$median_max_position)
         {
         printf
join("\t",$chr,$start,$end,$q_id,$score,$strand,$max_positions) ;
         }
      else
           {
           printf
join("\t",$chr,$start,$end,$q_id,$score,$strand,$median_max_position)
         }
}
```

```
chr8  97227900    97227987    tRNA_Leu_CAG      2118  -     -4
```

```
chr8  113154450   113154523   tRNA_Gly_GCC      1293  -     -25
chr8  113586461   113586534   tRNA_Gly_GCC      1295  +     -27
chr8  124103126   124103201   tRNA_Met_CAT      1300  -     -10
```

Exemple de fichier récapitulant la distance moyenne de BRF1 par rapport à la coordonnée « start » des gènes d'ARN de transfert (colonne 7).

C. Bases de Données

Les bases de données utilisées pour réaliser l'annotation des régions liées par la machinerie de classe III ont été téléchargées depuis le site ftp de l'UCSC. UCSC intègre les informations provenant de différentes bases ou outils d'annotation, et les repositionne sur le génome. Ces données sont regroupées en tables, et sont accessibles sur le site ftp : ftp://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/, ou via l'interface http://genome.ucsc.edu/cgi-bin/hgTables?command=start.
Sur le site ftp, à une table, correspondent deux fichiers, le .sql, décrivant la structure de la table, et le .txt qui est la table en elle-même.

Plusieurs tables ont été téléchargées puis compilées, afin d'obtenir un panel d'annotations du génome le plus complet possible. Chaque outil d'annotation fonctionne selon une approche différente. Les annotations ne sont pas décrites par toutes les sources, ainsi pour augmenter les chances de trouver une annotation, différentes tables ont été combinées.

Les tables utilisées sont les suivantes. Pour avoir une description complète du contenu de ces tables, le lecteur pourra se référer au site de l'UCSC, et trouver les descriptions correspondant à chaque piste.

Table	Brève description
ENSgene_ ENSgtp	Prédiction réalisée par ENSEMBL http://www.ensembl.org/info/docs/genebuild/index.html
VEGA	Annotations réalisées par Vertebrate Genome Annotation (Vega)
MGC	Alignements d'ARNm murins du MGC (Mammalian Gene Collection), ayant une ORF complète.
MIRNA	microRNAs de la miRBase au The Wellcome Trust Sanger Institute.
Refseq	Gènes codant des protéines ou gènes non codant de la base NCBI RNA Reference Sequence Collection (RefSeq).
JaxRep	Transcrits les plus représentatifs issus de la base MGI (Mouse Genome Informatics).
Aceview	Les modèles de gènes AceView sont construits à partir des cDNAs, et de vérifications expérimentales, en utilisant le programme Acembly.
UCSC	Prédiction de gènes basée sur les données de Refseq, GenBank et Uniprot.
RepeatMasker	Annotations réalisées en utilisant le programme RepeatMasker, à partir des motifs consensus de répétitions recensés par Repbase. (A.F.A. Smit, R. Hubley, and P. Green, http://www.repeatmasker.org)
GtRNAdb	Base de données Genomic tRNA database utilisant le programme ARNtcan-SE*.
Coughlin study	ARNt identifiés par Coughlin et al, réalignés sur le génome mm9 (programme liftOver).

Tableau 7. Table regroupant les différentes sources utilisées pour l'annotation des gènes liés par la machinerie de classe III.
*tRNAscan-SE est un programme intégré utilisant en premier filtrage, l'algorithme tRNAscan (Fichant), et la recherche de boîte A et B (Pavesi). Les ARNt candidats sont ensuite sélectionnés selon une recherche basée sur des modèles (COVE, Eddy). Cette base de données regroupe un ensemble de gènes candidats, représentant 99-100% de vrais ARNt.

D. Définitions des régions liées par l'ARN Polymérase III

Les régions considérées comme liées par la RNAP III sont issues de l'intersection des régions RPC1-RPC4, présentant un enrichissement suffisant de chaque sous-unité par rapport au contrôle 46C.

Les régions liées par chaque sous-unité de la RNAP III ont tout d'abord été recherchées en utilisant un programme dérivé de QuEST, Skeletton. Puis seules les régions RPC1-RPC4 chevauchantes sont gardées, les bornes minimales de l'intersection définissent les coordonnées de la nouvelle région.

Les régions ont ensuite été filtrées selon leur enrichissement par rapport au contrôle. Un rapport RPC4/46C, RPC1/46C a été calculé, seules les régions dont le rapport est supérieur à 5 ont été retenues.
Le rapport BRF1/BRF2 devait également être supérieur à 5, car leur présence est normalement exclusive au locus. Ce filtrage permet d'éliminer les régions non-spécifiques, où l'on observe la liaison simultanée de tous les facteurs.

Ces régions ont ensuite été annotées selon la base de données crée. Un score définissant la qualité de l'annotation est calculé à chaque fois. Il tient compte du chevauchement de la région de classe III, avec l'annotation candidate, le tout rapporté à la taille du gène candidat. Ce score permet de discriminer entre les différentes annotations, si une région de classe III est associée à plusieurs gènes candidats.
Parfois, les bornes des régions de classe III ont mal été définies lors du peak calling, et peuvent être associées à deux annotations d'ARN non-codant de scores équivalents. Ces régions sont inspectées visuellement et séparées en deux si besoin. De même, deux régions correspondant à une même annotation peuvent être fusionnées.

E. Définition des nouveaux gènes

Les régions liées par la RNAP III, auxquelles ne correspondent aucune annotation, sont examinées plus en détail, car elles constituent de nouveaux gènes potentiels transcrits par la RNAP III. Assez souvent, ce sont des sites faiblement liés par RPC4 et RPC1. Ainsi, ces régions seront validées après inspection visuelle de la répartition des lectures de ces deux protéines, et du contrôle 46C, sur le Browser de l'UCSC.

Pour délimiter les coordonnées de ces nouvelles régions, une procédure standard a été appliquée. Les bornes initiales définies par Skeletton ont été repoussées de part et d'autres de 100 pb. La densité de RPC4 a ensuite été calculée sur ce site ainsi étendu. Puis, centrée sur la position représentant la densité maximale de RPC4, une nouvelle région de 200 pb a été définie.

F. Défintion des régions liées par le complexe TFIIIC

Le ChIP-seq a été réalisé pour trois sous-unités du complexe TFIIIC. Les séquences ont été analysées indépendamment ; les régions ont été définies par QuEST, en utilisant une fenêtre de 600 pb et le seuil dit relâché ou seuil 3. Les sites liés par TFIIIC220 sont pris comme base de référence, les régions retenues seront celles communes aux trois protéines, définies suivant les bornes des régions TFIIIC220.

Programme : bed_intersect.py
Synopsis : `/share/apps/bx-python/scripts/bed_intersect.py régions_TFIIIC220.bed régions_TFIIIC.bed > TFIIIC_régions_communes.bed`

1. Défintions des sites ETC Extra-TFIIIC-loci

Les sites ETC sont les régions liées par le complexe TFIIIC indépendamment de l'ARN polymérase III. Il faut isoler les régions qui ne sont pas liées par une des sous-unités de la RNAP III ; le ChIP-seq de la sous-unité RPC4 ayant été le plus efficace, les régions liées par RPC4 ont servi de référence pour distinguer les sites ETC des sites liés par la RNAP III.

Programme : bed_intersect.py
Synopsis :
Sites communs avec ceux de la RNAP III:

```
/share/apps/bx-python/scripts/bed_intersect.py
TFIIIC_régions_communes.bed Regions_RPC4.bed > sites_TFIIIC_PolIII.bed
```

Sites ETC :

```
/share/apps/bx-python/scripts/bed_intersect.py -v
TFIIIC_régions_communes.bed Regions_RPC4.bed > sites_TFIIIC_ETC.bed
```

2. Organisation des régions TFIIIC-ETC par rapport aux régions CTCF.

Pour déterminer si les sites ETC ont un lien particulier avec CTCF, nous avons souhaité savoir s'il existait une proximité physique entre les sites TFIIIC-ETC et les sites CTCF. Après avoir défini les régions CTCF par QuEST (fenêtre 600pb, seuil 3), les régions TFIIIC-ETC sont organisées selon leur distance calculée entre les milieux des régions TFIIIC-ETC et CTCF.

Le script est exécuté sous R.
Les régions ainsi organisées sont soumises à une analyse par SeqMINER.

Programme : distanceTFIIIC_2_CTCF

```
tf3cReg<-
lecture.table('TF3C_ETC.bed',sep='\t',header=FALSE,col.names=c('chr','start','
end','score'))

ctcfReg<-lecture.table('CTCF_quest_2-
3.bed',sep='\t',header=FALSE,col.names=c('chr','start','end'))

ctcfReg<-ctcfReg[ (ctcfReg$chr != 'chrY') ,]

tf3cReg$mid = (tf3cReg$start+tf3cReg$end)/2
```

```
tf3cReg<-tf3cReg[ order(tf3cReg$chr,tf3cReg$mid), ]

ctcfReg$mid = (ctcfReg$start+ctcfReg$end)/2
ctcfReg<-ctcfReg[ order(ctcfReg$chr,ctcfReg$mid) , ]


tf3cReg$ctcfDist=NA
tf3cReg$ctcfChr=''
tf3cReg$ctcfMid=NA
tf3cReg$ctcfStart=NA
tf3cReg$ctcfEnd=NA


chroms=levels(tf3cReg$chr)
#chroms=c('chr19')

for (c in chroms)
{
  cpt=0
  #print(c)
  for (i in which(tf3cReg$chr == c) )
  {
    cpt=cpt+1
    minDist=10^10
    info1=''
    info2=''
    info3=''
    info4=''

    for (j in  which(ctcfReg$chr == c) )
    {
       dist = ctcfReg$mid[j] - tf3cReg$mid[i]
       #print(dist)
       if ( abs(minDist) >  abs(dist) )
       {
       minDist=dist;
         info1 = ctcfReg$chr[j]
         info2 = ctcfReg$mid[j]
         info3 = ctcfReg$start[j]
         info4 = ctcfReg$end[j]
       }
    }
    #print(i)

    if (minDist < 10^10)
    {
      tf3cReg$ctcfDist[i]=minDist;
      tf3cReg$ctcfChr[i]=info1;
      tf3cReg$ctcfMid[i]=info2;
      tf3cReg$ctcfStart[i]=info3;
      tf3cReg$ctcfEnd[i]=info4;
    }
  }
  print(paste(cpt,"regions for ",c))
}

tf3cRegOrd<-tf3cReg[ order(tf3cReg$ctcfDist), ]
```

```
write.table(tf3cRegOrd, file='TFIIIC_ETC_sort_by_CTCF_distance.tsv' ,
sep='\t', quote=FALSE,row.names=FALSE)
```

G. Unicité des régions

A chaque base du génome, a été attribué un score d'unicité. Une fenêtre glissante base par base, permet de désigner des séquences de 26 nucléotides. Si cette séquence est unique, c'est à dire, si elle ne peut s'aligner à aucun autre endroit dans le génome, alors la base de l'extrémité 5' de cette séquence, aura un score de 1. Si cette séquence n'est pas unique alors, un score de 0 sera attribué à cette position dans le génome. Ceci est réalisé pour chaque position du génome.

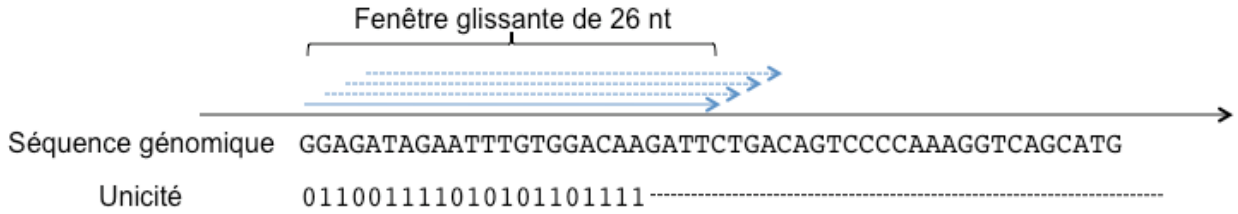


Figure 33. Description de la méthode permettant de calculer l'unicité de chaque position du génome.

1. Mappabilité des gènes d'ARN de transfert

- Une sélection des ARNt décrits par GtRNAdb et/ou Coughlin donne 526 ARNt prédits (Selection des ARNt).
- La mappabilité moyenne sur une fenêtre de plus ou moins 150 nt autour des ARNt a été calculée.

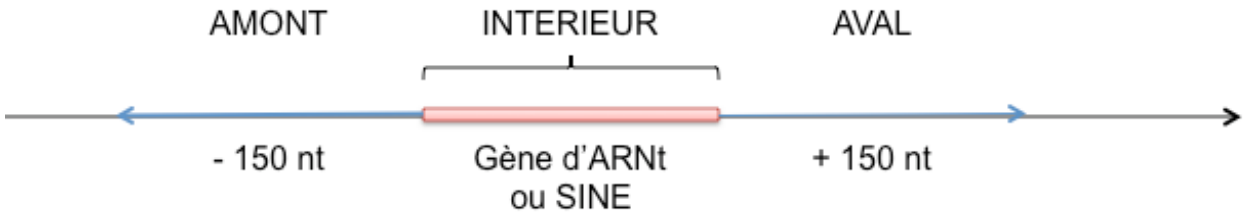


Figure 34. Calcul de la mappabilité des gènes d'ARN de transfert, ou des SINEs.

- Différentes sous-listes ont été définies selon la liaison par la RNAP III et la source de la description (GtRNAdb et/ou Coughlin):
  - Mappabilité des 284 ARNt liés par la RNAP III
  - Mappabilité des 242 ARNt NON liés par RNAP III et prédits par GtRNAdb ou Coughlin

Dans le cas des ARNt liés par la RNAP III, les plus mauvaises mappabilités observées en amont et en aval d'un ARNt sont obtenues pour All_tRNA.63 avec 0.12 et 0.23 (c'est à dire 12% et 23% de positions mappables). On observe des mappabilités inférieures (voir nulles) d'un côté ou de l'autre mais toujours compensées par une mappabilité supérieure de l'autre côté de l'ARNt.

Nous considérons la mappabilité comme "BAD" lorsque la mappabilité moyenne est strictement inférieure à 25% en AMONT, à l'INTERIEUR et en AVAL des ARNt. Avec ces critères nous obtenons les résultats suivants:

	GOOD	BAD	% BAD
liés par RNAP III	283	1	0,4%
non liés par RNAP III mais prédits par Coughlin OU GtRNAdb	210	32	13,2%

Tableau 8. Classement des gènes d'ARNt prédits dans la base de GtRNAdb et l'étude de Coughlin, selon les critères de mappabilité fixés.

2. Divergence et Mappabilité des SINES

La distribution des SINEs selon le pourcentage de divergence est bimodale, avec des maxima vers 7% et 27% de divergence. La séparation entre les deux sous-populations se fait vers 13%.

Cette valeur de 13% a été utilisée comme seuil pour distinguer deux sous-populations de SINEs faiblement (<13%) et fortement (>=13%) divergents. Ces deux sous-populations sont respectivement désignées par LD (Low Divergent, 282388 SINEs ~ 19%) et HD (High divergent, 1223225 SINEs ~ 81%).

La mappabilité a été calculée sur les régions flanquantes (150nt) et internes de chaque SINE prédit par RepeatMasker (UCSC track / mm9). Cette mappabilité correspond au pourcentage de positions donnant une séquence de 26 nucléotides UNIQUE dans le génome (sur les deux brins). L'analyse de cette mappabilité pour les SINEs liés par la RNAP III permet d'établir des valeurs minimales pour la mappabilité sur les trois régions (150nt amont, interne, 150nt aval) dans les deux catégories LD et HD:

- LD: 35%
- HD: 50%

En utilisant ces seuils, on peut classer les SINEs comme BAD (= NON détectable) si aucune des trois régions (150nt amont, interne, 150nt aval) n'atteint le seuil de mappabilité. On obtient ainsi, les effectifs suivants:

Seuil à 35% minimum sur AU MOINS 1 des 3 régions	GOOD	BAD	% BAD
Low Divergent Bound SINEs	190	1	0.5
High Divergent Bound SINEs	49	0	0
Low divergent not bound SINEs	274131	8257	2.9
High divergent not bound SINEs	1201726	21499	1.8

Seuil à 50% minimum sur AU MOINS 1 des 3 régions	GOOD	BAD	% BAD
Low Divergent Bound SINEs	190	1	0.5
High Divergent Bound SINEs	48	1	2

Low divergent not bound SINEs	268864	13524	4.8
High divergent not bound SINES:	1197021	26204	2.1

Tableau 9. Classement des SINEs « Low divergent », et « High divergent » selon les critères de mappabilité fixés à 35% pour une des trois régions, amont, aval ou corps du gène.

Estimation du nombre de SINEs liés et non mappés (les faux négatifs):

soit P le nombre de SINEs liés rapporté au nombre de SINEs mappables (GOOD pour seuil de mappabilité à 50%):

Low Divergent :
P = Nb Bound / Nb GOOD
= 190 / (190 + 268864)
= 190 / 269054 ~ 7 e-4

High Divergent :
Nb Bound / Nb GOOD
= 48 / (48 + 1197021)
= 48 / 1197069 ~ 4 e-5

Si on fait l'hypothèse que ce rapport P est le même pour la population des SINEs NON mappables (BAD), on peut dire:

LD :
P = Nb Bound / Nb BAD
Nb Bound = P * Nb BAD
Nb Bound = 7 e-4 * 13524
Nb Bound ~ 1

HD :
Nb Bound = 4 e-5 * 26204
Nb Bound ~ 0.1

`

H. Base de données des gènes de classe II

1. Création de la table des gènes RefSeq

La table des gènes RefSeq a été téléchargée depuis UCSC (http://genome.ucsc.edu/cgi-bin/hgTables?command=start), le 15 décembre 2010. Cette table contient l'ensemble des gènes connus, codant des protéines, issus de la collection des séquences d'ARN messagers de référence (RefSeq) du NCBI. Un même gène est souvent associé à plusieurs transcrits, possédant un TSS ou un TES alternatifs, ou issus d'épissage alternatifs. Ceci crée une certaine redondance dans la table. La première étape a consisté à créer une table, où un gène ne serait représenté qu'une seule fois. Le tri s'effectue au niveau des transcrits. La méthode est inspirée d'un papier de Zeitlinger (Zeitlinger et al., 2007). La densité de l'ARN Polymérase II (RNAP II) est calculée en différents endroits : au Promoteur-Proximal (PP), et au corps du gène (GB, « Gene-Body »).

Les gènes sont ensuite classés (1) selon la densité au PP, (2) selon la densité au GB. Les transcrits retenus sont ceux présentant la plus forte densité au PP, si celle-ci est égale, alors selon la densité la plus

forte au GB. Lorsqu'il est impossible de différencier les transcrits suivant la densité, alors un classement alphabétique permet de sélectionner celui apparaissant le premier.

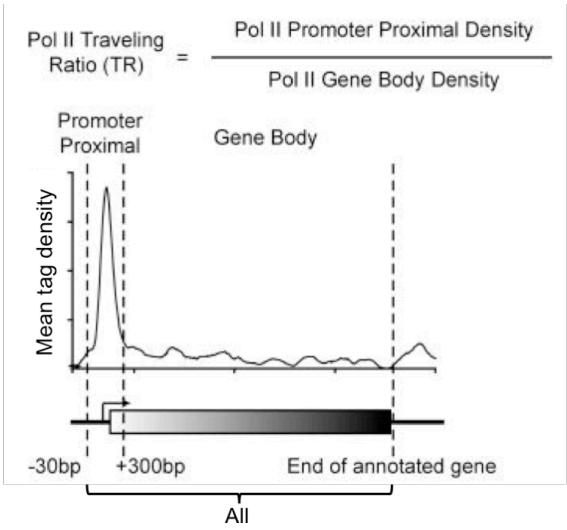


Figure 35. Schéma décrivant les différents intervalles utilisés pour le calcul des corrélations RNAP II-TCEA1 (d'après Rahl et al, 2010).
La région Promoter-proximal PP est définie comme la fenêtre de -30bp à +300bp autour du TSS du gène. Le corps du gène GB, s'étend de +300bp après le TSS, jusqu'au TES. La région « All » couvre l'ensemble du gène de -30bp jusqu'au TES.

Nous possédons une liste de gènes triée selon leur expression mesurée par RNA-seq, données issues du laboratoire de Matthieu Gérard. Cette liste a donc été triée selon la méthode décrite ci-dessus. De plus, les gènes dont la taille est inférieure à 300bp ont été éliminés, car la densité au PP est faussée, puisqu'elle recouvre l'ensemble du gène, au lieu de se limiter à la région du promoteur. Les chromosomes de type Random ont également été éliminés.

TRI	Nombre de gènes	Description de la table
1er	28046	Table téléchargée sur UCSC.
2ème	21191	Table contenant une liste de gènes uniques.
3ème	16867	Croisement entre la table avec un gène unique et les données RNA-seq.
4ème	16855	Elimination des transcrits de moins de 300bp.
5ème	16802	Elimination des chromosomes Random.

Tableau 10. Tableau résumant le nombre de gènes présents dans les listes à l'issue des différents tris.

Programme : handling_of_alternative_transcript

```
#!/bin/bash

variable=${1%.tsv}
bigWig="$2"
```

```
##définition des coordonnées Promoteur Proximal PP (-30 +300bp)
awk 'BEGIN {OFS="\t"} ($6 ~ /+/) {print $1,$2-30,$2+300,$4,$5,$6} ($6 ~ /-/)
{print $1,$3-300,$3+30,$4,$5,$6}' "$variable".tsv >
"$variable"_promoterProximal.bed

##définition des coordonnées  Gene Body GB (+300bp to TES)
awk 'BEGIN {OFS="\t"} ($6 ~ /+/) {print $1,$2+300,$3,$4,$5,$6} ($6 ~ /-/)
{print $1,$2,$3-300,$4,$5,$6}' "$variable".tsv >  "$variable"_geneBody.bed

##densité pour Pol II sur la région PP ou GB

get_tag_density -c "$bigWig" -g mm9 "$variable"_geneBody.bed >
Density_for_"$bigWig"_"$variable"_geneBody.bed

get_average_density Density_for_"$bigWig"_"$variable"_geneBody.bed >
average_density_for_"$bigWig"_"$variable"_geneBody.bed

get_tag_density -c "$bigWig" -g mm9 "$variable"_promoterProximal.bed >
Density_for_"$bigWig"_"$variable"_promoterProximal.bed

get_average_density Density_for_"$bigWig"_"$variable"_promoterProximal.bed  >
average_density_for_"$bigWig"_"$variable"_promoterProximal.bed

paste "$variable".tsv
average_density_for_"$bigWig"_"$variable"_promoterProximal.bed
average_density_for_"$bigWig"_"$variable"_geneBody.bed  | sed '1d' | cut -f1-
6,13,21 > "$variable"_"$bigWig".density

cat "$variable"_"$bigWig".density | sort -k5,5 -k7,7nr -k8,8nr  >
"$variable"_"$bigWig"_density.sort

get_uniq_alternative-transcript "$variable"_"$bigWig"_density.sort >
"$variable"_"$bigWig"_density.uniq
```

get_average_density

```
#!/usr/bin/perl

my $chr;
my $start;
my $end;
my $q_id;
my $score;
my $strand;
my $density_values;
my $average;
my $sum;
my $minimum;
my $maximum;
my $max_positions;
my $median_max_position;

while (<>)
{
        ($chr, $start, $end, $q_id, $score, $strand,$density_values, $average,
$sum, $minimum, $maximum, $m
ax_positions,$median_max_position)=split(/\r|\t/,$_);
        if ($sum==0.00)
```

```
           {
           printf join("\t",$chr,$start,$end,$q_id,$score,$strand,"NA","\n") ;
           }
        else
           {
           printf
join("\t",$chr,$start,$end,$q_id,$score,$strand,$average,"\n")
         }
}
```

get_uniq_alternative_transcript

```
#!/usr/bin/perl

my $alternatif='';

while (<>)
{
@c=split(/\t/,$_);
if ($c[4] ne $alternatif)
{
print $_;
}
$alternatif = $c[4];
}
```

2. Définition des gènes actifs et des gènes non-productifs

Les gènes sont triés selon leur expression. Les 10000 premiers ont été utilisés pour l'analyse, car ils étaient liés par la RNAP II au moins au promoteur, et présentaient la marque H3K4me3, modification associée aux promoteurs actifs (Barski et al., 2007; Mikkelsen et al., 2007). Deux listes en ont été extraites, les gènes dits actifs, les 2000 premiers, normalement activement transcrits, et ainsi associés à la modification H3K79me2, en aval du TSS. Les promoteurs des gènes dits non-productifs sont occupés par la RNAP II, mais ne présentent pas de marque H3K79me significativement associée. Ces gènes ne seraient pas transcrits, mais auraient une RNAP II pausée au promoteur. Les gènes de la position 6001 à 8000 présentent la marque H3K4me3 au promoteur, mais n'étant pas transcrit, la marque H3K79me2 n'est pas présente sur le corps du gène.

```
library(plyr,lib.loc="~/myRlibrary_tmp")
library(proto,lib.loc="~/myRlibrary_tmp")
library(reshape,lib.loc="~/myRlibrary_tmp")
library(ggplot2,lib.loc="~/myRlibrary_tmp")
library(splines)

if(!exists("tcea"))

pos =
lecture.table(file="refGeneReduit_RNAP2_hypoP_density_uniq_tri_rpkm_size_rando
m_occupancy_RNAP2_hypoP_TCEA1_density_all_pp_gb_TR.txt", header=FALSE,
stringsAsFactors=FALSE)

pos = pos[!is.na(pos$V7),]     # remove V7 NA's
pos = pos[!pos$V7 == 0,]     # remove V7 ZERO's
```

```
pos = pos[!is.na(pos$V11),]     # remove V11 NA's

#Etude sur l'ensemble des gènes

pos.df = data.frame(V7=pos$V7, V11=pos$V11)

pos.df = cbind(pos.df, V7.log=log(pos.df$V7), V11.log=log(pos.df$V11))



pdf('correlation_TCEA1_RNAP2.pdf')
## PLOT....

p = ggplot(pos.df, aes(V7.log,V11.log)) + geom_point(alpha=0.05)
p= p + stat_smooth(method="lm",formula=y~x, colour="red")
p = p + labs(x = 'Average enrichment of Pol II',y = 'Average enrichment of
TCEA1')
p=p+ geom_text(aes(x=-3,y=c(4,3.5),label=c("y = 0.555x +
0.541","R2=0.6267")),size=4)
print(p)


## Fit Linear model...
fit = lm(V11.log ~ V7.log, data=pos.df)
print(summary(fit))


cor.test(pos.df$V7,pos.df$V11)
cor.test(pos.df$V7,pos.df$V11,method="spearman")


#10000 premiers gènes de la liste
M=pos[1:10000,]

#Gènes actifs == 2000 plus forts
act=pos[1:2000,]

#Gènes non-productifs == 6001:8000
np=pos[6001:8000,]

dev.off()
```

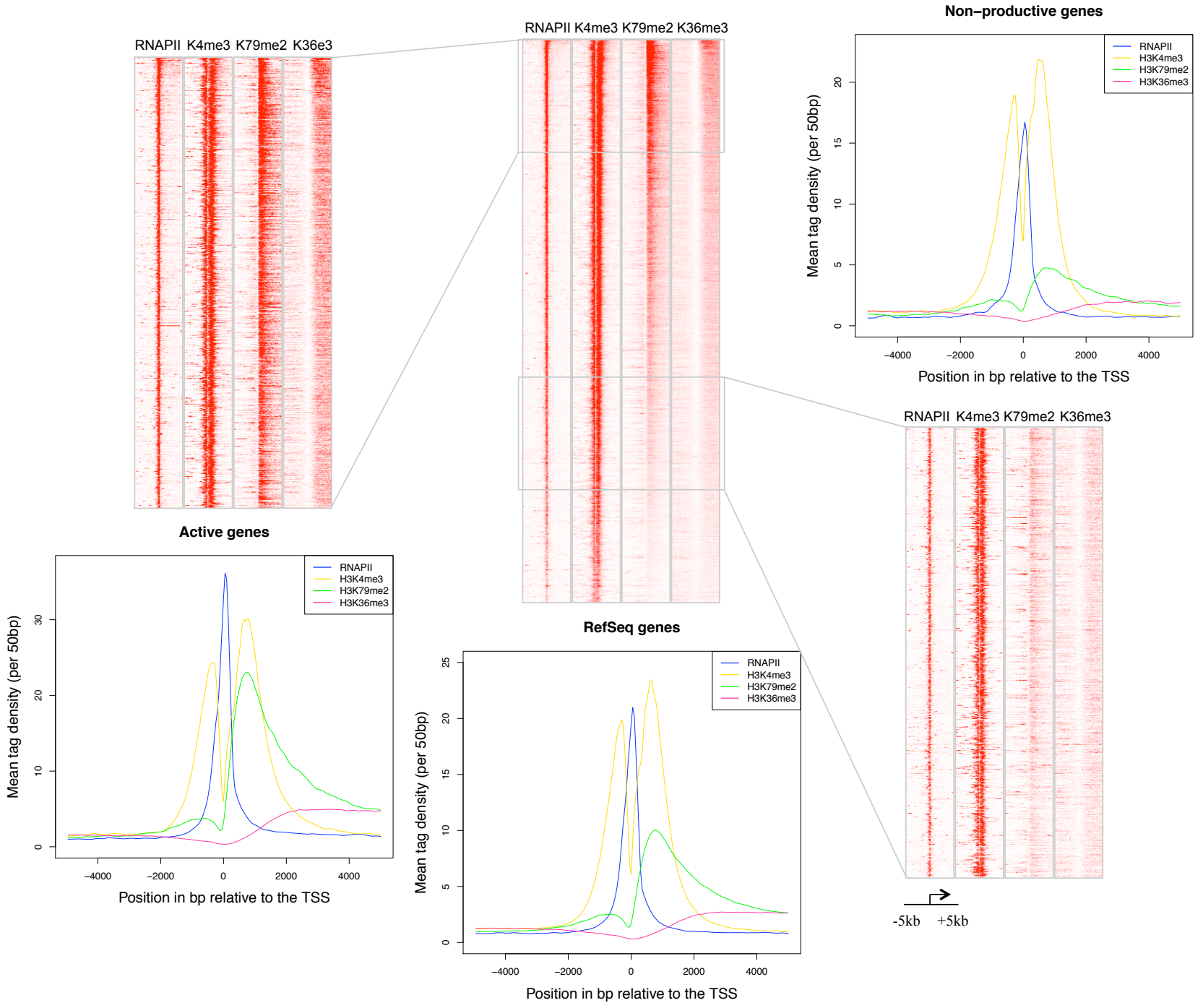

RNAPII K4me3 K79me2 K36e3
RNAPII K4me3 K79me2 K36me3
Non−productive genes
RNAPII
H3K4me3
H3K79me2
H3K36me3
Mean tag density (per 50bp)
Position in bp relative to the TSS
RNAPII K4me3 K79me2 K36me3
Active genes
RefSeq genes
-5kb +5kb

Figure 36. Schéma décrivant les différentes classes de gènes. La densité des différents facteurs est calculée plus ou moins 5 kb autour du TSS des gènes RefSeq. Chaque ligne correspond à un gène. L'ensemble des gènes a été trié selon le niveau de leur expression déterminée par RNAseq. Les heat-map sont générés par seqMINER (bin : 50 pb, extension des lectures :200 pb).
Le profil de densité est généré sous R, à partir des données de densité calculées par seqMINER.
En bleu, RNAP II. Jaune, H3K4me3 (K4me3). Vert, H3K79me2 (K79me2). Rose, H3K36me3 (K36me3).

I. SeqMINER

SeqMINER est une plateforme d'analyse des données de ChIP-seq mise en place par Tao Ye et Arnaud Krebs, de l'IGBMC, à Strasbourg (Ye et al., 2011).

SeqMINER permet de réaliser des comparaisons entre différents jeux de données de séquençage, et d'extraire des informations qualitatives (quel autre facteur se lie aux mêmes sites, par exemple) comme quantitatives (estimer la corrélation existante entre deux facteurs pour un ensemble de sites, notamment).
Une brève description de SeqMINER est donnée ici. Une information complète est disponible dans la publication de Ye et al.

seqMINER travaille sur un ensemble de régions (sites identifiés pour un facteur, gènes RefSeq, …), sur lesquelles sera calculé l'occupation des différents facteurs fournis au cours de l'analyse. Selon l'option spécifiée, les régions seront étendues de part et d'autres de la coordonnée centre, et découpées en intervalle (50pb par défaut), où la densité des lectures sera estimée sur l'ensemble du gène. Les données peuvent ensuite être visualisées par une heatmap.

Qu'est-ce qu'une Heatmap ?

Une heatMap est une représentation graphique où les valeurs sont représentées par des couleurs. Les HeatMaps sont, à l'origine, une matrice à deux-dimensions où sont regroupées les valeurs. Plus une valeur est forte, plus la couleur utilisée pour illustrer les valeurs sera foncée, et inversement, à une faible valeur sera associée une intensité de couleur plus faible.

Ici les matrices sont constituées de lignes indiquant la région étudiée (nom, coordonnées, et autres renseignements), les colonnes donnent les valeurs de densité calculées par bin, sur une fenêtre déterminée autour de la région.

```
chr17:45704726-45710210|NM_008302   2     0     0     1     1     1     2     1
      3     3     2     5     3     13
```

Figure 37. Exemple d'une matrice regroupant des valeurs de densité, pouvant être représentées par une heatMap.

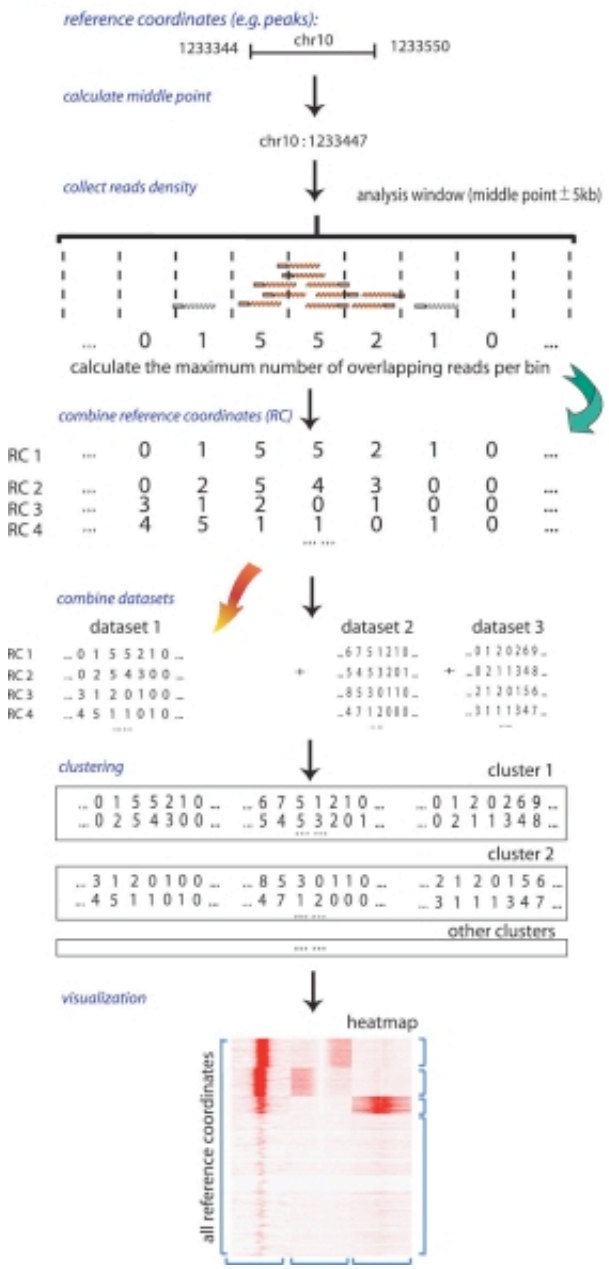


Figure 38. Représentation schématique décrivant la méthode de matrice de densité (density array) implémentée dans seqMINER (d'après Ye et al, 2010).
L'utilisateur fournit une liste de régions ou de gènes. Avant la quantification, les lectures sont étendues, comme lors de la création d'un fichier Wig, selon une valeur définie par l'utilisateur (par défaut, 200 pb). Le nombre de bins est également choisi dans une fenêtre fixe, définie autour de la coordonnée du milieu de la région. Les lectures sont décomptées par bin. Les valeurs sont ensuite collectées et les matrices créées peuvent être soumises au clustering, regroupant les régions présentant un profil de densité similaire. La visualisation directe de la densité peut aussi être utilisée. Les valeurs sont ensuite visionnées comme une heatmap.

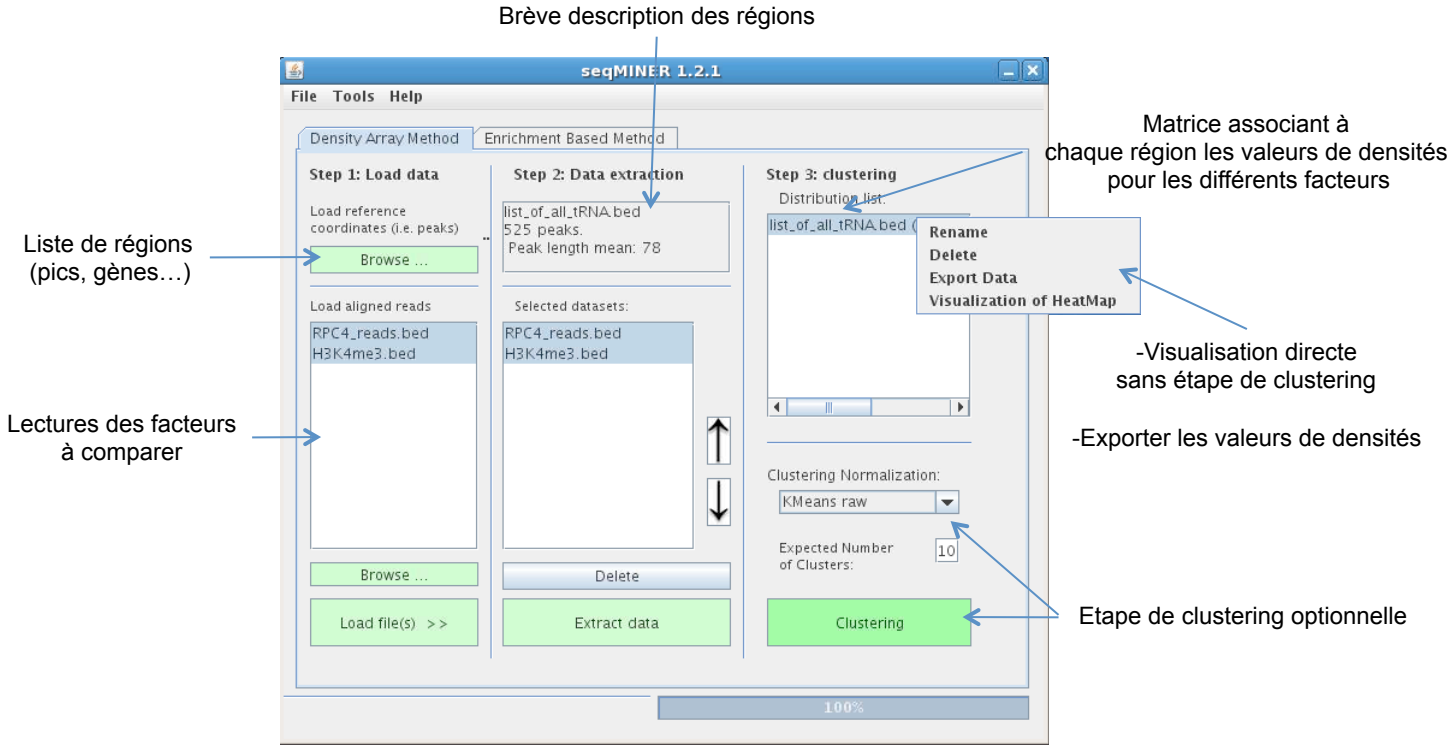


Figure 39. Schéma représentant l'interface graphique de seqMINER.

J. Matrices consensus

Les motifs de liaisons de la machinerie de classe III ont été identifiés avec le logiciel MEME 4.5, installé en local. MEME donne en sortie, un fichier indiquant sous plusieurs formats le ou les motifs identifiés dans un ensemble de séquences. La première ligne, quelque soit le format utilisé pour décrire le motif identifié, indique la taille de l'alphabet utilisé, la E-value du motif, la taille (width) et le nombre d'occurrence de ce motif. Cette E-value (Expect value) décrit le nombre de fois où un motif peut apparaître purement par chance dans une base de donnée d'une taille particulière. Ce paramètre permet d'évaluer la valeur statistique du motif identifié.

1. Format PSPM ou Position-Specific Probability Matrix

Ce motif donne pour chaque position la fréquence observée ou probabilité de chaque lettre (nucléotide dans notre cas) possible. Cette matrice de probabilité donne, par ligne une lettre du motif, dans l'ordre, la première ligne étant la première lettre du motif. A cette lettre, correspond autant de colonne que de lettre possible (quatre pour les nucléotides), et à chaque lettre la probabilité associée. La première ligne de description commence ici par « letter-probability matrix ».

2. Format PSSM ou Position-Specific Scoring Matrix

Ce format est toujours organisé en lignes pour les lettres du motif, et en colonnes, les différentes lettres pouvant être retrouvées dans le motif et un score associé. Cette matrice est une matrice log-odds, calculé à partir du ratio 100 * log2 p/f.

-p : probabilité d'occurrence d'une lettre à cette position dans le motif

-f : fréquence « bruit de fond » d'une lettre toujours à cette position.

La première ligne de description commence ici par « log-odds matrix ».

Ce format est utilisable par les programmes MAST ou FIMO, qui peuvent rechercher un motif prédéfini, dans un ensemble de séquences, selon une certaine p-value.

Dans le cas des gènes ARNt, j'ai directement recherché le motif dans l'ensemble des séquences d'ARNt retrouvés liés. Ceci m'a permis de créer un fichier contenant les motifs en format PSSM, utilisé ensuite pour les nouveaux gènes ou les SINEs.

Motif PSPM

```
MEME version 4.5

ALPHABET= ACGT

strands: +

Background letter frequencies (from
A 0.213 C 0.287 G 0.287 T 0.213

MOTIF A_box
letter-probability matrix: alength= 4 w= 11 nsites= 283 E= 0
 0.000000  0.000000  0.000000  1.000000
 0.390071  0.028369  0.581560  0.000000
 0.000000  0.000000  1.000000  0.000000
 0.000000  0.652482  0.000000  0.347518
 0.081560  0.280142  0.283688  0.354610
 0.007092  0.464539  0.141844  0.386525
 1.000000  0.000000  0.000000  0.000000
 0.241135  0.000000  0.758865  0.000000
 0.028369  0.159574  0.067376  0.744681
 0.021277  0.106383  0.702128  0.170213
 0.000000  0.000000  1.000000  0.000000
##Box A t[ag]g[ct]nnannng

MOTIF B_box
letter-probability matrix: alength= 4 w= 9 nsites= 283 E= 0
 0.000000  0.000000  1.000000  0.000000
 0.056537  0.000000  0.000000  0.943463
 0.000000  0.000000  0.000000  1.000000
 0.000000  0.992933  0.000000  0.007067
 0.229682  0.000000  0.770318  0.000000
 0.996466  0.000000  0.003534  0.000000
 0.498233  0.049470  0.183746  0.268551
 0.084806  0.233216  0.028269  0.653710
 0.000000  1.000000  0.000000  0.000000
##Box B g[at]tc[ag]annc
```

Motif PSSM

```
MEME version 4.5

ALPHABET= ACGT

strands: +

Background letter frequencies (from
A 0.213 C 0.287 G 0.287 T 0.213

MOTIF A_box_T[AG]G[CT]T[CT]A[AG]TGG
##TRGYTYARTGG
log-odds matrix: alength= 4 w= 11
 -1478  -1478  -1478    223
    88   -334    102  -1478
 -1478  -1478    180  -1478
 -1478    118  -1478     71
  -138     -4     -2     74
  -490     69   -102     86
   223  -1478  -1478  -1478
    18  -1478    140  -1478
  -291    -85   -209    181
  -332   -143    129    -32
 -1478  -1478    180  -1478

MOTIF B_box_GTTC[AG]A[AT]TC
##GTTCRAWTC
log-odds matrix: alength= 4 w= 9
 -1479  -1479    180  -1479
  -191  -1479  -1479    215
 -1479  -1479  -1479    223
 -1479    179  -1479   -491
    11  -1479    142  -1479
   223  -1479   -634  -1479
   123   -254    -65     34
  -133    -30   -335    162
 -1479    180  -1479  -1479
```

Cependant les motifs extraits à partir des séquences d'ARNt sont très stricts. Les motifs des nouveaux gènes pourraient, comme pour ceux de l'ARN 7SL, avoir dégénéré. A partir de ces motifs identifiés, j'ai donc recréé une matrice PSSM, en « allégeant » les scores pour certaines positions qui peuvent, selon la littérature être plus flexible (nucléotide N), tout en maintenant un score élevé pour les positions dites fondamentales. Le programme MEME est capable à partir d'une séquence de type IUPAC de créer une matrice PSSM ou PSPM (script iupac2meme).

Boîte A

```
MEME version 3.0

ALPHABET= ACGT

strands: + -

Background letter frequencies (from
```

```
A 0.213 C 0.287 G 0.287 T 0.213

##A box

MOTIF TRGYTYARTGG

BL   MOTIF TRGYTYARTGG width=11 seqs=20
log-odds matrix: alength= 4 w= 11 n= 0 bayes= 0 E= 0
-100.000000 -100.000000 -100.000000   2.000000
  1.000000  -100.000000   1.000000  -100.000000
-100.000000 -100.000000   2.000000  -100.000000
-100.000000   1.000000  -100.000000   1.000000
-100.000000 -100.000000 -100.000000   2.000000
-100.000000   1.000000  -100.000000   1.000000
  2.000000  -100.000000 -100.000000 -100.000000
  1.000000  -100.000000   1.000000  -100.000000
-100.000000 -100.000000 -100.000000   2.000000
-100.000000 -100.000000   2.000000  -100.000000
-100.000000 -100.000000   2.000000  -100.000000

letter-probability matrix: alength= 4 w= 11 nsites= 20 E= 0
  0.000000    0.000000    0.000000    1.000000
  0.500000    0.000000    0.500000    0.000000
  0.000000    0.000000    1.000000    0.000000
  0.000000    0.500000    0.000000    0.500000
  0.000000    0.000000    0.000000    1.000000
  0.000000    0.500000    0.000000    0.500000
  1.000000    0.000000    0.000000    0.000000
  0.500000    0.000000    0.500000    0.000000
  0.000000    0.000000    0.000000    1.000000
  0.000000    0.000000    1.000000    0.000000
  0.000000    0.000000    1.000000    0.000000
```

Boîte B

```
MEME version 3.0

ALPHABET= ACGT

strands: + -

Background letter frequencies (from
A 0.213 C 0.287 G 0.287 T 0.213

##B box
MOTIF GTTCRAWTC

BL   MOTIF GTTCRAWTC width=9 seqs=20
log-odds matrix: alength= 4 w= 9 n= 0 bayes= 0 E= 0
-100.000000 -100.000000   2.000000  -100.000000
-100.000000 -100.000000 -100.000000   2.000000
-100.000000 -100.000000 -100.000000   2.000000
-100.000000   2.000000  -100.000000 -100.000000
  1.000000  -100.000000   1.000000  -100.000000
  2.000000  -100.000000 -100.000000 -100.000000
  1.000000  -100.000000 -100.000000   1.000000
```

```
-100.000000 -100.000000 -100.000000   2.000000
-100.000000   2.000000  -100.000000 -100.000000

letter-probability matrix: alength= 4 w= 9 nsites= 20 E= 0
  0.000000    0.000000    1.000000    0.000000
  0.000000    0.000000    0.000000    1.000000
  0.000000    0.000000    0.000000    1.000000
  0.000000    1.000000    0.000000    0.000000
  0.500000    0.000000    0.500000    0.000000
  1.000000    0.000000    0.000000    0.000000
  0.500000    0.000000    0.000000    0.500000
  0.000000    0.000000    0.000000    1.000000
  0.000000    1.000000    0.000000    0.000000
```

Un peu de statistiques…

A. Etudes de Corrélation

En probabilités et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques numériques, signifie étudier l'intensité de la liaison qui peut exister entre ces variables. La liaison recherchée peut-être une relation affine. Dans le cas de deux variables numériques, il s'agit de la régression linéaire.

Qu'est qu'une relation linéaire ?

La régression linéaire consiste à déterminer une estimation des valeurs a et b et à quantifier la validité de cette relation grâce au coefficient de corrélation linéaire. On peut alors proposer un modèle linéaire, c'est-à-dire chercher la droite dont l'équation est yi = axi + b et qui passe au plus près des points du graphe. Comment calculer les caractéristiques de cette droite ? En faisant en sorte que l'erreur que l'on commet en représentant la liaison entre nos variables par une droite soit la plus petite possible. Le critère formel le plus souvent utilisé, mais pas le seul possible, est de minimiser la somme de toutes les erreurs effectivement commises au carré. On parle alors d'ajustement selon la méthode des moindres carrés ordinaires. La droite résultant de cet ajustement s'appelle une droite de régression. Plus la qualité globale de représentation de la liaison entre nos variables par cette droite est bonne, et plus le coefficient de corrélation linéaire associé l'est également. Il existe une équivalence formelle entre les deux concepts.

Coefficient de corrélation linéaire, ou de Pearson.

Le coefficient de corrélation ou coefficient de corrélation linéaire ou coefficient de corrélation Bravais-Pearson mesure l'intensité de liaison existant entre deux séries d'observations, pour autant que cette relation soit linéaire ou approximativement linéaire. Ce coefficient est égal au rapport de leur covariance et du produit non nul de leurs écarts types. Le coefficient de corrélation est compris entre -1 et +1.

Si r vaut 0, les deux courbes ne sont pas corrélées. Les deux courbes sont d'autant mieux corrélées que r est loin de 0 (proche de -1 ou +1).

Il est égal à 1 dans le cas où l'une des variables est fonction affine croissante de l'autre variable, à -1 dans le cas où la fonction affine est décroissante. Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte ; on emploie simplement l'expression «fortement corrélées» pour qualifier les deux variables. Une corrélation égale à 0 signifie que les variables sont linéairement indépendantes.

En revanche, ce coefficient de corrélation est extrêmement sensible à la présence de valeurs aberrantes ou extrêmes (ces valeurs sont appelées des "déviants") dans notre ensemble de données (valeurs très éloignées de la majorité des autres, pouvant être considérées comme des exceptions).

Attention, il est toujours possible de calculer un coefficient de corrélation (sauf cas très particulier) mais un tel coefficient n'arrive pas toujours à rendre compte de la relation qui existe en réalité entre les variables étudiées. En effet, il suppose que l'on essaye de juger de l'existence d'une relation linéaire entre nos variables. Il n'est donc pas adapté pour juger de corrélations qui ne seraient pas linéaires et non linéarisables.

Il n'existe pas de relation linéaire claire entre les données d'enrichissement de la RNAP II et de TCEA1 ; nous ne pouvions donc pas utiliser le coefficient de corrélation de Pearson. La transformation des données en fonction log-log permettait d'obtenir un coefficient fort de corrélation entre nos données, mais quelle était l'explication biologique ?

Nous avons donc opté pour le coefficient de Spearman. Il permet de rechercher une corrélation entre des rangs. Ceci amène à s'affranchir des biais créés par l'expérience de ChIP-seq. Nous cherchons à comparer des données provenant de deux protocoles totalement différents, l'un avec un anticorps endogène, réalisé avec une étape de précipitation et une fragmentation par sonication, l'autre résultant d'une double immunoprécipitation, et d'une fragmentation par la MNase.

Corrélation de Spearman.

Le coefficient de Spearman, (de Charles Spearman), noté ρ (rho) ou rs, est étudiée lorsque deux variables statistiques semblent corrélées sans que la relation entre les deux variables soit de type affine, contrairement au coefficient de Bravais-Pearson. Elle consiste à trouver un coefficient de corrélation, non pas entre les valeurs prises par les deux variables mais entre les rangs de ces valeurs. Elle permet de repérer des corrélations monotones. Il faut également souligner que la corrélation de Spearman utilise les rangs plutôt que les valeurs exactes. L'interprétation est identique à celle de la corrélation par rangs de Pearson.

Formule utilisée dans R

```
cor(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))
```

B. Analyse de la proximité des sites liés par CTCF des Sites ETC

Nous avons défini aléatoirement autant de régions sur le génome, que de sites ETC identifiés, en utilisant une loi uniforme, runif dans R.

```
genomeTab <- lecture.table(file='mm9.chrom.sizes',sep=' ',
col.names=c("chrom","size"),row.names='chrom')



initGenomeTab <- function()
{
    genomeSize=0
    genomeTab$start=0
    genomeTab$end=0

    for (i in row.names(genomeTab))
    {
        genomeTab[i,'start']=genomeSize+1
       genomeSize=genomeSize+genomeTab[i,'size']
        genomeTab[i,'end']=genomeSize
      #print(i)
    }
    return(genomeTab)
}

genome2Chrom <- function(genomePos)
{
   chromPos=data.frame()
   for (p in 1:length(genomePos))
   {
     for (i in row.names(genomeTab))
     {
      if (genomePos[p]>=genomeTab[i,'start'] &
genomePos[p]<=genomeTab[i,'end'])
      {
            chromPos[p,'chrom']=i
                chromPos[p,'pos']=genomePos[p]-genomeTab[i,'start']+1
      }
    }
  }
 return(chromPos)
}




genomeTab<- initGenomeTab()


randomPos=runif(2233,genomeTab$start[1],genomeTab$end[20])


randomChromPos<-genome2Chrom(randomPos)
randomChromPos$pos=as.integer(randomChromPos$pos)
```

```
exportPos=data.frame(randomChromPos)
exportPos$start=randomChromPos$pos

write.table(exportPos,file="exportRandom_unif_pos.tsv",sep='\t',row.names=FALS
E,col.names=FALSE)


#2è partie

tf3cReg<-
lecture.table('exportRandom_unif_pos.tsv',sep='\t',header=FALSE,col.names=c('c
hr','start','end'))
```

Données utilisées au cours de l'analyse.

Auteurs	Facteur	Numéro d'accession SRA	Année
Mikkelsen et al.	H3K4me1	SRR002255	2008
	H3K4me2	SRR002253 SRR002254	
	H3K36me3	SRR007433 SRR007434	
Marson et al.	H3K79me2	SRR015144 SRR015145	2008
Bilodeau et al.	H3K9me3_Upstate	SRR031675 SRR031676	2009
	H3K9me3_Abcam	SRR031672, SRR031673 SRR031674	
Bilodeau et al	TBP-B2	SRR057606	2010
Kagey et al	Smc1A	SRR058981 SRR058982	2010
	Smc3	SRR058983 SRR058984	
Chen et al	CTCF	SRR001985 SRR001986 SRR001987	2008
Seila et al	RNAPII hypophosporylé	SRR014267	2008
Rahl et al	RNAPII ser2P	SRR036734 SRR036735 SRR036736	2010
	RNAPII ser5P	SRR036733	
	RNAPII ser7P	SRR051928	
Cox et al	H3K4me3	SRR038974 SRR038984	2010
	H3K27me3	SRR038975 SRR038976	

Tableau 11. Récapitulatif des données de ChIP-seq utilisées dans l'étude.

ANNEXES

Protocole Recombineering

But : Insérer par recombineering un TAG en C-ter d'un gène dans un locus de souris.

I-Constructions in silico

Localiser le gène d'intérêt avec ENSEMBL.
Dans Detailed view, sélectionner dans DAS source l'option 129S7/AB2.2 clones pour commander le BAC. Comme le TAG doit être inséré en C-ter choisir deux BAC qui pourront contenir 10 kb de part et d'autre du C-ter. Les commander (Geneservice) (figure).
Extraire sous forme FASTA la séquence -10 kb - « C-ter (codon stop) »- +10 kb et la mettre dans strider ou ApE.
Repérer précisément le codon stop et adopter une stratégie pour le génotypage des cellules ES (figure). Ce génotypage se fera par Southern Blot. Il faut choisir une enzyme de restriction qui coupe idéalement 2 fois (une à l'extérieur des miniarms, et l'autre à l'intérieur). Cette enzyme doit être une de celles utilisées par le labo pour les Southern Blot (ne doit pas être inhibée par CpG méthylation).
BamHI, NcoI, Eco RI, Xho I, et Xba I sont apportés par la construction. En plus de ces sites le plasmide pL452-FHH (Fayçal) comprend, après le stop du TAG, les sites de restriction pour Eco RV et Hind III. Le plasmide pL452-FHHseb (Sébastien) contient les sites Kpn I et Xba I (figure):
Placer la sonde à l'extérieur des miniarms 5' ou 3'. Cette sonde ne doit pas être disposée dans des régions répétées. Pour le vérifier utiliser le programme *Repeat Masker* et sélectionner le génome de la souris.
Choisir les amorces PCR adéquates avec Primer3 Input pour générer une sonde de 600-700 pb (inclure de part et d'autre un site Eco RI pour le clonage futur dans bluescript : le but est de fabriquer la sonde en purifiant le plasmide la possédant en insert).

IMPORTANT : Si possible choisir 2 stratégies, une avec une sonde en 5' et l'autre avec une sonde en 3' pour pouvoir voir un shift de 2 kb quand la K7 Néo sera excisée.
Tenir compte du placement des miniarms (la sonde doit être à l'extérieur des miniarms.
En fonction de l'enzyme choisie, définir le plasmide pL452 qui sera utilisé (Fayçal ou Sébastien).
Le profil entre le WT et le mutant en southern devra être assez différent :
Ex : Si pour le WT on obtient une bande à 10 kb, dans l'idéal on doit avoir une bande à 5 kb pour le mutant.

Une fois la sonde placée, disposer les miniarms 5' et 3'. Ils doivent être idéalement placés à respectivement -5 kb et +5 kb du Stop et comprendre entre 250-500 pb. Dans tous les cas s'assurer qu'il y a 10 kb entre les miniarms. On peut avoir un bras court à 3kb et un bras long à 7kb.
Choisir les amorces adéquates avec *Primer3 input.*
Comme les miniarms seront amplifiés par PCR et clonés dans le « retrieval plasmid » (en Not I / Spe I pour le miniarm 5' et Spe I / Bam HI pour le miniarm 3') vérifier qu'il n'y a pas de sites de restriction Not I / Spe I dans le 5' et Spe I / Bam HI dans le 3'.
Ajouter dans l'amorce FW du miniarm 5' le site Not I, et dans l'amorce Rev le site SpeI.
Ajouter dans l'amorce FW du miniarm 3' le site Spe I, et dans l'amorce Rev le site Bam HI.
(voir tableau et figure)
Les bras d'homologie doivent être compris entre 250-500 pb. L'amorce Rev du bras d'homologie 5' comprendra les 20 nucléotides avant le codon stop. L'amorce FW du bras d'homologie 3' comprendra les 20 nucléotides suivants codon stop compris (figure).
Ajouter dans l'amorce FW du miniarm 5' le site Sal I, et dans l'amorce Rev le site Asc I.

Ajouter dans l'amorce FW du miniarm 3' le site Bam HI, et dans l'amorce Rev le site Not I (figure et tableau).

IMPORTANT : Le bras d'homologie 5' ne devra pas contenir de sites Sal I et Asc I. Le 3' ne devra pas posséder de sites Bam HI et Not I.

Rapatrier dans strider les séquences des miniarms, des bras d'homologie, des sondes pour toutes les constructions, et les séquences de 20 kb flanquant le stop du gène d'intérêt.
Construire dans strider la carte du « Gap Repaired Plasmid », du « Retrieval Plasmid », du « Minitargeting Vector, du « Retrieval Plasmid » et du « Neotargeted Plasmid ».

II- Insertion du BAC dans la souche SW102 et Vérifications

Préculture des bactéries renfermant les BAC :
5 mL LB/ Chloramphénicol 12,5µg/ml final.

Préparer un stock glycérol : Mettre 1,2 mL de préculture avec 0,6 mL de glycérol 75% stérile.

Purifier les deux BAC sélectionnés.

-Miniprep d'ADN plasmidique:

1. Culture ON à 37°C sur 3 ml LB liquide + Chloramphénicol.
2. Récolter les cellules par centrifugation 2 min dans un eppendorf 2 ml.
3. Resuspension dans 100µl de solution I puis vortexer.
4. Ajouter 200µl de solution II puis inverser (5-6 fois).
5. Ajouter 150µl d'Acétate d'ammonium 3M pH 5,2 puis inverser (5-6 fois).
6. Centrifuger 10 min puis récupérer le surnageant dans un eppendorf 1,5 ml.
7. Ajouter 1ml d'Ethanol 100% (-20°C) puis laisser précipiter 10 min à -20°C.
8. Centrifuger 10 min à 13000 rpm et enlever le surnageant.
9. Laver le culot avec 1ml d'éthanol 70% puis centrifuger 10 min à 13000 rpm.
10. Enlever le surnageant et sécher le culot 15 min dans l'étuve à 37°C.
11. Resuspendre le culot dans 50µl d'H2O dd stérile.

On doit obtenir environ 10µg pour une culture en LB/ Chloramphénicol (digestion sur 2µl)

Solution I :

Solution II à préparer extemporanément :
NaOH 200mM
SDS 1%

Doser l'ADN pour vérifier sa pureté et la quantité. (Faire une dilution au 1/100 et doser à 260 nm).
Vérifier sur gel le profil des BAC préalablement (Digestion par une, deux, ou trois enzymes de restriction (une, deux ou 3 digestions indépendantes).
PS : j'ai vérifié mes BAC avec Kpn I et Eco RI.

Mix réactionnel :
10µl ADN BAC (miniprep)
3µL Tampon 10X
16µl H2O dd
1µl Enzyme

Incubation à 37°C pendant 3hrs au bain marie.

Migration sur gel agarose 1% ON à 40 V. (de 18h30 à 9h00). Le lendemain, il est parfois utile d'incuber le gel dans un bain TBE 1X + BET, avant de prendre une photo.

- Transformation par électroporation des cellules SW102

1. Pré-culture de 5 ml LB ON à 30-32 °C.
2. Mettre de l'H2O dd, des falcons 50 mL des eppendorfs et des cuvettes à 4°C.
3. Diluer au 50ème (500µL dans 25 mL) et incuber 3-5h jusqu'à atteindre une DO_{600}=0,6. 25 mL de culture correspond à 2 tubes de cellules compétentes.
4. Incuber les cellules dans la glace pendant 2 min puis les transférer (12,5 ml) dans 2 falcons 50 mL préalablement à 4°C.
5. Centrifuger à 4°C, 5000 rpm pendant 5 min. Enlever le surnageant et égoutter sur papier, puis ajouter 5 ml d'eau à 4°C en gardant le tube dans la glace. Resuspendre le culot en agitant doucement (en faisant un cercle) le falcon placé dans la glace pendant 5 min. Quand les cellules sont resuspendues, compléter à 15 mL avec de l'eau glacée puis inverser quelques fois.
6. Centrifuger à 4°C, 5000 rpm pendant 5 min.
7. Refaire les étapes 6 et 7 (la resuspension est plus facile cette fois). Enlever le surnageant en inversant le tube sur du papier (ne pas perdre le culot). Resuspendre le culot dans le volume résiduel du tube et stocker les cellules compétentes à 4°C (environ 200µL par falcon).
8. Transférer 50µL de cellules électrocompétentes dans un eppendorf préalablement mis à 4°C et mélanger 2 µL d'ADN. Pour un BAC utiliser 1-5µg d'ADN (ici environ 10 µg). Transformer par électroporation (1750 V). Vérifier le time constant. Il doit être aux alentours de 5-6mv.
9. Transférer les cellules dans 1 mL de LB puis incuber à 30°C pendant 1h.
10. Etaler les bactéries (30 µL, 150µL, et le reste) sur le milieu de sélection ici : Chloramphénicol. Centrifuger à 5000 rpm et enlever le surnageant pour le reste. Resuspendre dans 150µl et étaler.
11. Incuber à 30-32°C pour 16-24h.

NB : Inclure un contrôle sans BAC et l'étaler sur une boîte LB et une LB/Chloramphénicol.

Le lendemain :
Les colonies sont très petites. Il est préférable de les laisser une bonne partie de la journée.
Repiquer 5 clones dans 3 ml de LB/ Chloramphénicol ON à 30°C.

Préparer un stock glycérol (1 échantillon par BAC. Mettre 1,2 mL de préculture avec 0,6 mL de glycérol 75% stérile).
Purifier les BAC à partir de la souche SW102

Vérifier sur gel le profil des BAC préalablement digérés par une, deux voire 3 enzymes de restriction (une, deux ou 3 digestions indépendantes). PS : j'ai vérifié mes BAC avec Kpn I et Eco RI.
Digérer le BAC original (issu de la souche E. coli initiale) et 2 BAC issus de la souche SW102.
Comparer les profils et valider le BAC.
Vérifier les BAC par PCR.

III- Sonde, miniarms et bras d'homologie.

1- Vérification des couples d'amorces.

Mix réactionnel :

1µl ADN BAC
1µl Amorces FW 10µM
1µl Amorces Rev 10µM

2,5µl Tampon 10X
0,2µl Taq
1µl $MgCl_2$ 50 mM
1µl dNTPs 10 mM
17,3 µl H20 dd

Programme PCR :

2' à 94°C
15'' à 94°C
30'' à 60°C } 30 cycles
45'' à 72°C
4' à 72°C

Faire migrer 5-10 µl de la PCR dans un gel 1% agarose. Stocker à 4°C le reste de la PCR.

2- Amplification par PCR des miniarms, bras d'homologie et sonde pour clonage.

Refaire les PCR mais cette fois faire 4 tubes pour chaque produit de PCR (pour générer suffisamment de matériel pour le clonage). Faire un contrôle négatif pour chaque PCR (chaque couple d'amorces).

Faire migrer les quatre PCR avec à coté 5-10µl de la PCR ayant servi à contrôler les amorces (pour vérifier la taille de l'amplicon et ainsi voir s'il n'y a pas eu d'inversions).

Purifier les produits PCR (96µl restant) sur colonne Nucleospin (Kit Nucléospin extract II Macherey-Nagel). Reprendre les échantillons dans 30µl d'H2O dd.

Vérification des BmQ dans la souche SW102 :

- Par PCR

Prendre par exemple les amorces pour amplifier la sonde et tester 3 BmQ par construction.
Mettre en contrôle positif le reste de la PCR de vérification des amorces.

- Par digestion

IV- Construction du Retrieval Plasmid
Clonage des miniarms 5' et 3' dans pL253

1- Digestion du miniarm 5' par Not I / SpeI

Mix réactionnel :

30µL PCR
1µl Not I
1µl Spe I
3,5µl NEB2

2- Digestion du miniarm 3' par Spe I / Bam HI

Mix réactionnel :

30µL PCR
1µl Spe I
1µl Bam HI
3,5µl React 4

3- Digestion de pL253

Deux options :

- si on veut cloner les miniarms 5' et 3' en même temps, il faut digérer pL253 par Not I / Bam HI.
- Si on veut procéder par étape, il faut commencer par digérer pL253 avec Not I / Spe I et cloner le miniarm 5'. Ensuite digérer pL253-miniarm 5' par Spe I / Bam HI et y cloner ensuite le miniarm 3'.

Dans tous les cas digérer 20µg de plasmide dans 30µl final.
Utiliser le tampon React 3 pour Not I / Bam HI.

En contrôle pour vérifier que les enzymes de restriction sont fonctionnelles digérer séparément 2µg de pL253 avec Spe I, Bam HI et Not I.

2µg pL253
1µl Enzyme
2µl Tampon 10X
qsp H2O dd 20µl

Toutes les digestions sont déposées au bain marie à 37°C pendant 2h30

4- Purification sur gel des digestions (Kit Nucléospin extract II Macherey-Nagel)

Dépôt sur gel d'agarose 1% des digestions.
Utiliser des peignes grand format.
Laisser un puit vide entre chaque échantillon.
Migration à 80V (préconisé par le kit)
Placer le gel sur le transluminateur protégé avec du saran. Le gel est morcelé. Exposer un morceau de gel et découper le plus rapidement possible les bandes avec un scalpel.
Extraire chaque bande puis la découper en petits bouts pour faciliter l'extraction. Nettoyer la lame du scalpel entre chaque bande !!!!
Suivre le protocole d'extraction nucleospin extract II.

Vérifier sur gel de la bonne fonctionnalité des enzymes (cf contrôles de digestion).

- Traitement du plamide digéré à la Phosphatase alcaline
16µl plasmide digéré
2µl tampon NEB3
1µl CIP
Incuber 30 minutes à 37°C.
Inactiver l'enzyme 30 min à 65°C.
Extraction Phénol/Chloroforme. Précipitation EtOH, le culot est repris dans 8µl H2O.

5- Quantification des produits digérés et purifiés sur gel

- Préparation du marqueur

Concentration initiale du marqueur II (λ HindIII) = 0,5µg/µl
Concentration initiale du marqueur IX (φ Hae III) = 1µg/µl
Concentration finale : 20µg/ml.
Mélanger 40µl de marqueur II et 20µl de marqueur IX avec 940 µl de TAE 1X
TAE 1X = 94µl TAE bleu 10X + 846µl H2O dd
Diluer au demi un aliquot de cette solution et déposer 20µl pour avoir le marqueur 200ng, 10µl pour le 100 ng et 5µl pour le 50ng.

- Préparation des échantillons.

Charger sur gel l'échantillon non dilué et dilué au 1/5 pour les fragments PCR et non dilué, au 1/5 ou 1/10 pour les plasmides.
Faire migrer les échantillons sur un gel d'agarose 1%.
Estimer la quantité des échantillons sur le gel
La qualité de la photo est très importante pour l'estimation. Ne pas saturer le signal.

Formule : ng/µl * 10^3 / (635 * nb bases (de l'échantillon)) = pmoles/µl.

6- Ligation des miniarms dans pL253

Deux options : cf. IV.3.

- Double ligation

Mix réactionnel :

4µl Tampon 5X
0,1 pmole pL253 (Not I / Bam HI)
0,5 pmole miniarm 5' (Not I / Spe I)
0,5 pmole miniarm 3' (Spe I / Bam HI)
1µl Taq DNA Ligase
Qsp 20µl H2O dd

Incuber ON à 16°C ou 1hr à Température Ambiante. (PS : j'ai toujours ligué ON).

- Simple ligation

Mix réactionnel :

4µl Tampon 5X
0,1 pmole pL253 (Not I / Spe I)
0,5 pmole miniarm 5' (Not I / Spe I)
1µl Taq DNA Ligase
Qsp 20µl H2O dd

NB : Dans les 2 cas faire un contrôle sans insert.

7- Transformation des produits de ligation

Préparer des boîtes LB agar /Ampicilline 100µg/ml final (au laboratoire, ils utilisent 50µg/ml final).
100µl Topo 10 ou XL1-blue ou TBS chimiocompétentes avec 5µl de la ligation (j'ai travaillé en grande partie avec les TBS).
Incuber 30' sur glace
Choc thermique à 42°C pendant 45 sec.
Laisser 1 à 2' sur glace.
Ajouter 1 ml de LB puis incuber 1h à 37°C.
Etaler 100µl et le reste.

Inclure en contrôle des cellules sans ADN et faire le même traitement puis étaler 100µL sur LB et 100µl sur LB/Ampiciline.

8- Purification de l'ADN des clones qui ont poussé

Piquer 8 clones par construction et les ensemencer dans 3 ml de LB/Ampicilline. Incuber ON à 37°C.
Purification ADN plasmidique
Miniprep cf. protocole précédent.

9- Vérification du clonage du retrieval plasmid

Digestion pour sortir l'insert

4µl ADN
0,3µl Enz I
0,3µl Enz II
2µl Tampon
13,4µl H2O dd

Digérer 2-3hrs à 37°C dans un bain marie.
Migrer sur un gel agarose 1%.

Faire un stock glycérol de 2 clones positifs pour chaque construction.

NB : on peut également digérer pL253 et le retrieval plasmid par Spe I et les faire migrer sur gel.

Si la première option a été privilégiée, passer directement au 10.
Pour la 2ème option :
Après avoir vérifié le clonage du miniarm 5' dans pL253, il faut cloner le miniarm 3'.

Faire une midiprep du plasmide pL253-miniarm5' :
Digérer 20µg de pL253-miniarm5' :

Mix réactionnel :

20µg pL253-miniarm5'
2µl Spe I
2µl Bam HI
3µl React 4
H2O dd qsp 30µl

Digérer 2-3hrs à 37°C dans un bain marie.
Faire la ligation entre pL253-miniarm5' (Spe I / Bam HI) et PCR miniarm 3' (Spe I / Bam HI).
Transformer la ligation.
Faire une préculture de 6 clones qui ont poussé.
Analyser l'incorporation du bras d'homologie par digestion.

10- Séquençage des produits de ligation

On séquence deux clones de chaque construction. On séquence le brin sens et le brin anti-sens. L'amorce FW est l'amorce T7 car pL253 est un dérivé de pBSK. L'amorce Rev est celle qui a servi à amplifier le miniarm 3'.

Mix réactionnel :

1µl Big Dye
3,5 µl de diluant
0,5µl T7 FW 10µM ou Rev du miniarm 3'
14µl H2O dd

Programme PCR :

95°C 20 sec
55°C 30 sec } 30 cycles
60°C 4 min

Purification des produits PCR :
Déposer dans chaque tube 50µl d'éthanol absolu et 6µl d'acétate d'ammonium 10M.
Ajouter 5µg de glycogène (pour voir le culot)
Vortexer puis refroidir 10' sur glace.
Centrifuger à 13000 rpm pendant 20' à 4°C.
Laver avec 100µl d'éthanol glacé puis sécher le culot à 37°C.
Resuspendre le culot dans 10µl de HiDi Formamide.

11- Analyse des séquences

Utiliser le programme Sequencher.

V- Construction du Gap repaired Plasmid (GRP)

1- Digestion du retrieval plasmid

Faire soit une midiprep des retrieval plasmid ou purifier sur nucléospin les minipreps et les reprendre dans 30µl H2O dd (PS : j'ai purifié les miniprep, par contre, il y a peu de matériel à transformer entre 10 et 50 ng. Cependant je n'ai pas eu de problèmes de transformation).

Mix réactionnel :

2µg ADN miniprep
3µl Tampon React 4
1µl Spe I
H2O dd qsp 30µl

Digestion 3h30 à 37°C dans le bain marie.
Vérifier sur gel la digestion en chargeant 3µl sur un gel d'agarose 1%.
Purifier sur colonne Nucleospin les plasmides digérés par Spe I (il faut que la digestion du plasmide soit totale).
Quantifier sur gel en chargeant 5µl.

2- Transformation des SW102 renfermant les BmQ des constructions avec les retrieval plasmid correspondants.

1. Pré-culture de 5 ml LB/ Chloramphénicol ON à 30-32 °C.
2. Mettre de l'H2O dd, des falcons 50 mL des eppendorfs et des cuvettes à 4°C.
3. Diluer au 50ème (500µL dans 25 mL) et incuber 3-5h jusqu'à atteindre une DO_{600} =0,6.
4. Incuber les cellules dans la glace pendant 2 min puis les transférer (12,5 mL) dans 2 falcons 50 mL préalablement à 4°C.
5. Traiter pendant 15' un échantillon à 30°C et un autre à 42°C pour induire la production des protéines de recombinaison. Agiter manuellement.
6. Refroidir les cellules dans de l'H2O dd glacée pendant 5 min en agitant.
7. Centrifuger à 4°C, 5000 rpm pendant 5 min.
8. Enlever le surnageant et égoutter sur papier, puis ajouter 5 ml d'H2O dd à 4°C en gardant le tube dans la glace. Resuspendre le culot en agitant doucement (en faisant un cercle) le falcon placé dans la glace pendant 5 min. Quand les cellules sont resuspendues, compléter à 15 mL avec de l'H2O dd glacée puis inverser quelques fois.
9. Centrifuger à 4°C, 5000 rpm pendant 5 min.
10. Refaire les étapes 6 et 7 (la resuspension est plus facile cette fois).
11. Enlever le surnageant en inversant le tube sur du papier (Ne pas perdre le culot).
12. Resuspendre le culot dans le volume résiduel du tube et stocker les cellules compétentes à 4°C (environ 100µL par falcon).
13. Transférer les cellules électrocompétentes dans un eppendorf préalablement mis à 4°C et mélanger idéalement 50 ng de Retrieval plasmid digéré par Spe I.
14. Transformer par électroporation à 1750 V. Transférer les cellules dans 1 mL de LB puis incuber à 30-32°C pendant 1h.
15. Etaler les bactéries (1 µL 100µL et le reste) sur le milieu de sélection (LB/Ampicilline pour sélectionner le retrieval plasmid).

Incuber à 32°C pour 16-24h.

NB : j'ai transformé de 5ng à 30ng faute d'en avoir plus. Dans tous les cas les transformations ont fonctionné.

Le lendemain :
Préculture de 6 clones dans 3mL de LB/ Ampicilline ON à 30-32°C.

Purification des minipreps
Cf protocole habituel

3- Vérification du Gap repaired Plasmid : GRP : vérification de l'insertion des 10 kb

Définir une stratégie de digestion entre le GRP et le Retrieval Plasmid.
Essayer d'avoir au maximum 3 bandes. Utiliser soit une simple ou une double digestion. Idéalement on linéarise le RP et on obtient 2 bandes pour le GRP.

Digérer 5µl d'ADN dans 20µl final.
NB : 3µl devrait être suffisants.

Faire un stock glycérol de la souche contenant le Gap repaired plasmid validé.

VI- Construction du mini-targeting vector (MTV)

Si vous utilisez les plasmides déjà construits par Fayçal ou Sébastien passez directement au 2

1- Sous clonage du Tag His-Flag-HA dans pL452 (Sal I / Eco RI) (figure)

Il faut utiliser le plasmide pL452 qui contient les séquences LoxP-Neo-LoxP (marqueur de sélection).
On doit insérer dans le plasmide :

- le TAG que l'on veut insérer en phase avec le C-ter du gêne d'intérêt.
- les sites de restriction utilisés pour le génotypage. Ces sites doivent se trouver en aval du TAG.

Le clonage des bras d'homologie se fera en Sal I / Asc I pour le bras 5' et Bam HI / Not I pour le 3'.
Remarque : le bras 3' est à l'extérieur du TAG.
Il faut pour cela ajouter dans la construction le site Asc I qui n'est pas présent dans pL452.

- Amplification par PCR du TAG :

Le plasmide contenant le Tag 6*His-Flag-HA est le pOZFHH (fournit par Fayçal)
On introduit à l'amorce FW le site Sal I et Asc I. Entre ces 2 sites est ajouté Xho I pour optimiser le clonage ultérieur du bras d'homologie 5' en Sal I / Asc I.
On introduit à l'amorce Rev les sites des enzymes utilisées pour le génotypage et pour finir le site Eco RI servant au clonage.

IMPORTANT : Comme Asc I coupe 8 nucléotides il faut en ajouter un pour conserver la phase de lecture. Ici un C est ajouté et par conséquent une proline sera crée.

NB : il n'est pas nécessaire d'ajouter les sites de restriction pour BamHI, NcoI, Eco RI, Xho I, et Xba I car ces sites sont respectivement présents dans la construction finale ainsi ou dans la séquence codant le TAG.

- Cloner le TAG en Sal I / Eco RI dans pL452.

Réaliser la PCR :

Programme PCR :

2' à 94°C
15'' à 94°C
30'' à 60°C
45'' à 72°C
} 30 cycles
4' à 72°C

Faire un gel 2% pour vérifier la taille du fragment PCR (151 pb).
Purifier sur nucléospin le produit de PCR.
Digérer 20µg de pl452 et les 30µl de PCR avec Sal I et Eco RI dans du NEB3.
Digérer en contrôle 2µg de pL452 avec Eco RI et Sal I.
Le produit PCR est digéré sur un gel 1%.
Repiquer ici 12 à 24 voire 48 clones.

- Séquencer le TAG pour voir s'il est bien en phase avec le C-ter.

2- Clonage des bras d'homologie 5' et 3' dans pL452-FHH

IMPORTANT : En fonction des sites de restriction présents dans les bras d'homologie on doit choisir dans quel ordre on clonera les bras d'homologie dans le pL452-FHH.
Ex : pour une construction, le bras d'homologie 5' (cloné en Sal I / Asc I) avait un site Bam HI. J'ai du cloner en premier le bras d'homologie 3' (cloné en Bam HI / Not I) puis en second le bras d'homologie 5'.

a- Clonage du miniarm 5' dans pL452-FHH

- Digestion du miniarm 5' par Sal I / Asc I

Les tampons de Sal I et Asc I ne sont pas compatibles. Il faut faire une digestion séquentielle.
Mix réactionnel :

30µL PCR homologie 5'
1µl Asc I
3,5µl NEB4

- Digestion de pL452-FHH

Mix réactionnel :

20µg pL452-TAG
2µl Asc I
3µl NEB4
qsp H2O dd 30µl

Vérifier sur gel en contrôle la digestion d'Asc I sur pL452-FHH.

- Purification Nucléospin des produits digérés par Asc I

Les échantillons sont repris dans 30µl d'H2O dd.

- Digestion de pL452-FHH (Asc I) et de PCR homologie 5' (Asc I) par Sal I

Mix réactionnel :

30µL PCR homologie 5' ou pL452-FHH
1µl Asc I (PCR homologie 5') ou 2µl (pL452-FHH)
3,5µl NEB4

- Purifier sur gel les produits de digestion.

Suivre le protocole habituel de clonage.

- Vérification de l'insertion du bras d'homologie 5' :

Digestion en Sal I / Eco RI.
Faire un stock glycérol du plasmide pL452-homologie5'-FHH vérifié.

- Midiprep de pL452-homologie5'-FHH :

Préculture de 100-200 ml LB/ Ampicilline ON à 37°C.

b- Clonage du miniarm 3' dans pL452-homologie5'-FHH

- Digestion du miniarm 3' par Not I / Bam HI

Mix réactionnel :

30µL PCR homologie 3'
1µl Not I
1µl Bam HI
3,5µl React 3

- Digestion du pL452-homologie5'-FHH par Not I / Bam HI

Mix réactionnel :

20µg pL452-homologie5'-TAG
2µl Not I
2µl Bam HI
3µl React 3
H2O dd qsp 30µl

Inclure en contrôle une digestion de pL452-homologie5'-FHH par Bam HI et Not I.
Digestion 2h30-3hr au bain marie à 37°C.
Suivre le protocole habituel de clonage.

- Vérification de l'insertion du bras d'homologie 5' :

Digestion en Not I / Bam HI.
Faire un stock glycérol du plasmide pL452-homologie5'-FHH vérifié.

- Vérification de la construction de la construction finale : pL452-homologie5'-FHH-homologie3'

- Séquençage des bras d'homologie 5' et 3'.
Le 5' est séquencé en FW avec l'amorce FB13 et en Rev avec l'amorce F14 (se fixe dans la cassette Néo). Avec F14 on pourra vérifier si le TAG est inséré en phase.
Pour le bras d'homologie 3' aucune amorce n'a été désignée. On peut utiliser les amorces utilisées pour la PCR homologie 3' ou alors synthétiser des amorces en amont et en aval du clonage.
- Digestion par Not I / Sal I (facultatif).

Faire un stock glycérol de la souche contenant le mini-targeting vector (MTV) vérifié.

IMPORTANT : vérifier dans le séquençage que le TAG est en phase et que les sites de restriction pour le génotypage sont bien présents.

VII- Construction du neo-targeted plasmid (NTP)

1- Midiprep du MTV

2- Extraction de la mini-targetting cassette

- Digestion MTV par Not I / Sal I

Mix réactionnel :

20µg MTV
2µl Not I
2µl Sal I
3µl React 3

Digestion 2h30-3h ou ON à 37°C

- Purification sur gel

Il n'est souvent pas possible de purifier les 2 bandes issues de la digestion. Cette purification sur gel est faite pour éliminer le MTV qui n'a pas été digéré.

- Quantification des produits.

3- Transformation de la mini-targetting cassette dans la souche SW102 contenant le Gap Repaired Plasmid.

1. Pré-culture de 5 ml LB ON à 30-32 °C.
2. Mettre de l'H2O dd, des falcons 50 mL des eppendorfs et des cuvettes à 4°C.
3. Diluer au 50ème (500µL dans 25 mL) et incuber 3-5h jusqu'à atteindre une DO_{600} =0,6.
4. Incuber les cellules dans la glace pendant 2 min puis les transférer (12,5 mL)dans 2 falcons 50 mL préalablement à 4°C.
5. Traiter pendant 15' un échantillon à 30°C et un autre à 42°C pour induire la production des protéines de recombinaison. Agiter manuellement.
6. Refroidir les cellules dans de l'H2O dd glacée pendant 5 min.
7. Centrifuger à 4°C, 5000 rpm pendant 5 min.
8. Enlever le surnageant et égoutter sur papier, puis ajouter 5 ml d'H2O dd à 4°C en gardant le tube dans la glace. Resuspendre le culot en agitant doucement (en faisant un cercle) le falcon placé dans la glace pendant 5 min. Quand les cellules sont resuspendues, compléter à 15 mL avec de l'H2O dd glacée puis inverser quelques fois.
9. Centrifuger à 4°C, 5000 rpm pendant 5 min.
10. Refaire les étapes 6 et 7 (la resuspension est plus facile cette fois).
11. Enlever le surnageant en inversant le tube sur du papier (Ne pas perdre le culot). Resuspendre le culot dans le volume résiduel du tube et stocker les cellules compétentes à 4°C (environ 100µL par falcon).
12. Transférer les cellules électrocompétentes dans un eppendorf préalablement mis à 4°C et mélanger 50ng de mini-targetting cassette.
13. Transformer par électroporation. Transférer les cellules dans 1 mL de LB puis incuber à 30-32°C pendant 1h.
14. Etaler les bactéries (1 µL 100µL et le reste) sur le milieu de sélection LB/Kanamycine de l'électroporation. En se mettant sur Kanamycine on sélectionne les clones qui auront recombiné. La sélection Kanamycine est apportée par la mini-targetting cassette.
15. Incuber à 32°C pour 16-24h.

Le lendemain :

Préculture de 6 clones dans 3mL de LB/Kanamycine ON à 30-32°C.

Purification des minipreps

Vérification du Minitargeting Vector MTV
Définir une stratégie de digestion entre le GRP et le MTV.
Essayer d'avoir au maximum 3 bandes. Utiliser soit une simple ou une double digestion.
Idéalement on linéarise le GRP et on obtient 2 bandes pour le NTP.

Digérer 5µl d'ADN dans 20µl final.

NB : 3µl devrait être suffisant.

Faire un stock glycérol de la souche contenant le NTP validé.

VIII- Eletroporation du neo-targeted plasmid (NTP) en cellules ES

1- Maxiprep du NTP.

Préculture de 100-200 ml LB/ Kanamycine ON à 30°C (bactéries SW102).

Quantifier l'ADN.
Vérifier l'intégrité du NTP par restriction.

2- Linéariser le NTP avec Not I

20µg de NTP linéarisé seront électroporés en cellules ES. L'ADN digéré est purifié par une extraction Phénol/Chloroforme. Seuls les 2/3 de la phase aqueuse seront prélevés afin que l'échantillon soit le plus pur possible. Ainsi, pour disposer d'une quantité suffisante, il faut digérer 35µg au départ.

Mix réactionnel :

35µg NTP
3µl Not I
3µl React 3
H2O dd qsp 30µl

Digestion 2h30-3h à 37°C

3- Purification par phénol/Chloroforme

1. Allonger le volume jusqu'à 200µl , et faire une extraction Phénol/Chloroforme.
2. Précipiter l'ADN en ajoutant 1/10^{e} d'Acéate de Sodium 3M, pH 5,2, et 2 volumes d'éthanol 100%.
3. Incuber 5 min dans la glace.
4. Centrifuger à 13000 rpm, 10 min.
5. Faire un lavage Ethanol 70%, reprendre dans 20µl H2O dd.
6. Quantifier le matériel obtenu.

Protocole d'Immunoprécipitation de Chromatine (ChIP)

Culture : la veille, ensemencer environ 8.10^6 cellules par boite de 10cm gélatinée.
Changer le milieu 2 heures avant.

Préparation de la chromatine

1. Crosslinker par ajout de 0,4% de formaldéhyde dans le milieu de culture (110µl de formaldéhyde 36,5% dans 10 ml de milieu de culture). Incuber 10 min à Température Ambiante (TA).
2. Arrêter la réaction par ajout de glycine à une concentration finale de 0,125M (stock 1,25M préparé le jour même, 1 ml par boite). Incuber 5 min à Température Ambiante (TA).
3. Rincer 2 fois les cellules avec du PBS 1X froid.
4. Scraper les cellules dans 2ml de PBS froid contenant des inhibiteurs (PMSF 1mM + Complete). Transférer dans un falcon. Conserver à 4°C.

A partir de cette étape, il faut travailler le plus possible dans la glace.

5. Centrifuger les cellules à 2500 rpm, 5 min, 4°C.
6. Aspirer le surnageant, et reprendre les cellules dans le tampon de lyse FA/SDS (plus PMSF 1mM, et Complete), 1 ml pour environ 20.10^6 cellules. Incuber 20 min dans la glace.
7. Purifier la chromatine par centrifugation à 13000 rpm, 20 min, 4°C. Aspirer le surnageant
8. Homogénéiser la chromatine avec du FA/SDS (plus PMSF 1mM, et Complete), 1 ml pour environ 20.10^6 cellules dans un eppendorf 2ml, avec une p1000. Incuber pendant une heure, sur une roue à 4°C.
9. Centrifuger la chromatine à 13000 rpm, 20 min, 4°C.
10. Regrouper tous les tubes, et reprendre la chromatine dans du FA/SDS (plus PMSF 1mM, et Complete), 1,6 ml pour environ 80.10^6 cellules (entre 4 et 5 boites 10 cm) dans un eppendorf 2 ml.
11. Procéder à la sonication. Les tubes sont dans un mélange eau/glace/NaCl pendant la sonication. Réglage de l'appareil : pulse 60%, puissance 4. 6 cycles 20 sec ON/40 sec OFF. Si la chromatine n'est pas suffisamment sonique, selon l'utilisation de la chromatine (ChIP-qPCR ou séquençage Solexa). Sur le Diagenode Bioruptor, rajouter plusieurs cycles de 20 sec ON/40 sec OFF (Attention, le volume maximale est 500 µl. Il faut de plus que le bain soit très froid).
12. Centrifuger la chromatine sonique à 13000 rpm, 10 min, 4°C. Garder le surnageant et le transférer dans un nouveau tube.
13. Prélever un aliquot de 20 µl afin de vérifier la sonication et quantifier l'ADN (Input). Conserver la chromatine à -80°C en aliquot de 500 µl.

Quantification de la chromatine et vérification de la sonication

Préparation de l'input de chromatine (ADN total, sonique, sans étape d'immunoprécipitation de chromatine)

1. Réverser le crosslink de l'aliquot de 20 µl : ajouter 480 µl d'H20 dd, 20 µl de NaCl 5M, et 15 µg de RNase A. Incuber 4 heures ou ON à 65°C.
2. Ajouter 10 µl d'EDTA 0,5M pH8, 50 µl de Tris pH8 1M et 4 µl de protéinase K 10 mg/ml. Incuber 1 heure à 50°C.
3. Extraire l'ADN par une purification Phénol/Chloroforme.

4. Précipiter l'ADN en ajoutant 1/10 vol d'Acétate de Sodium (NaAc) 3M, pH5,2, 2-2,5 vol d'éthanol 100% et 10 µg de Glycogène. Incuber 30 min à -20°C, puis centrifuger à 13000, 20 min, 4°C. Faire un lavage éthanol 70%. Sécher le culot et resuspendre dans 50 µl H20 dd.
5. Estimer la quantité et la pureté de l'ADN au nanodrop.
6. Vérifier la sonication en faisant migrer un échantillon de 5 µl sur un gel d'électrophorèse agarose 1,5%.

Si la sonication est incomplète, réaliser des cycles de sonication supplémentaires.

Immunopréciptation de la chromatine (simple ou tandem)

Jour 1 :
Préparer du PBS/BSA 2% + un tube frais à 0,1%, 4°C.

Préparation des billes magnétiques:

1. Prélever 50ul par point d'IP. Pooler dans 1 tube eppendorf 3 ou 4 points d'IP (150 à 200ul de billes.
2. Laver les billes 3 fois 1ml H2O
3. Laver les billes 2 fois 1ml PBS/BSA 0,1%
4. Resuspendre les billes dans 100ul PBS/BSA 0,1% par point d'IP (300ul ou 400ul)
5. Ajouter 3ug d'α-HA (ou α-FLAG) par point d'IP
6. Incuber 1300 rpm,1 heure, 30°C, dans l'agitateur thermostaté.

Blocage des billes magnétiques:

Bloquer les villes dans 1ml FA/SDS + 0,5 mg/ml de BSA + 0,2 mg/ml yeast tRNA. Saturer 1h à température ambiante sur la roue.

Immunoprécipitation 1:

Pour chaque ChIP, préparer 80 µg (RPC1 et RPC4) à 250 µg (BRF1 et BRF2) de chromatine soniquée, compléter si besoin le volume à 500 µl avec le tampon de lyse FA/SDS + inhibiteurs. Pour le ChIP tandem préparer l'équivalent de 3,5 IP.

1. Ajouter les billes et incuber 2 heures à 21°C.
2. Jeter le surnageant.
3. Laver les billes avec 1ml des tampons suivants:
   FA/SDS, 1 fois.
   Low Salt buffer, 1 fois, 5 min d'incubation.
   High Salt buffer, 3 fois, pour le dernier lavage, 10 min d'incubation.
   Low LiCl, 2 fois, 5 min d'incubation pour les deux lavages.
   TE, 2 fois, pour le dernier lavage, 5 min d'incubation
4. Au cours du dernier lavage, regrouper les billes dans un seul et unique tube pour chaque condition.

Elution 1:

5. Ajouter
   220 µl de FA/SDS + inhibiteurs
   30 µl de peptide HA ou Flag à 4ug/ul selon l'anticorps utilisé.
6. Eluer à 4°C sur la roue ON.

Jour 2 :
Récupérer l'équivalent d'1/2 IP, afin de contrôler la première étape d'immunoprécipitation

1. Préparer les billes magnétiques comme la veille, en ne comptant qu'un point d'IP par facteur. Utiliser bien évidemment un anticorps différent de celui utilisé à la première étape.
2. Ajouter l'éluat de la veille, en complétant le volume à 500 µl avec FA/SDS + inhibiteurs. Ajouter la chromatine aux billes.

3. Incuber 2 heures à 21°C sur la roue.
4. Jeter le surnageant et répéter les lavages comme lors de la première étape.
5. Eluer avec 250 µl de tampon d'élution à 65°C.
6. Collecter l'éluat et répéter l'étape d'élution. Combiner les deux éluats.

Réversion du crosslink :

Reverser le crossling du contrôle de la première IP, ainsi que des échantillons issus du tandem. Ajouter 20 µl de NaCl 5M, et 15 µg de RNase A. Incuber ON à 65°C.

Remarques :

- A l'issue de la deuxième immunoprécipitation, on peut également procéder à une élution peptide, comme lors de l'étape 5 du jour 1.
- Si l'on ne souhaite qu'une seule étape de ChIP, immunoprécipiter l'équivalent d'un seul point d'IP.
-

Jour 3 :

Purifier les échantillons comme lors de l'étape de purification de l'input (Quantification de la chromatine et vérification de la sonication).
Quantifier les échantillons obtenus au nanodrop, ou au Picogreen si la concentration est trop faible (seuil de détection du nandrop = 10 ng/µl).
Les échantillons d'Input, de ChIP simple, et de première étape de ChIP sont repris dans 50 µl H2O dd.
Les échantillons issus du ChIP tandem sont repris dans 20 µl d'H20 dd.

Solutions

FA/SDS
10 mM EDTA pH 8.0
50 mM Tris pH 7.5
150mM Nacl
1mM EDTA
1% Triton X100
0,1% Na-Deoxycholate
0,1%SDS

Low Salt Buffer (50 ml)
0.1 % SDS — 500 µl SDS 10 %
1 % Triton X-100 — 500 µl Triton X-100
2 mM EDTA pH 8 — 400 µl EDTA 250mM, pH 8
20 mM Tris, pH 8 — 1 ml Tris 1 M, pH 8
150 mM NaCl — 1.5 ml NaCl 5 M

High Salt Buffer (50 ml)
0.1 % SDS — 500 µl SDS 10 %
1 % Triton X-100 — 500 µl Triton X-100
2 mM EDTA pH 8 — 400 µl EDTA 250mM, pH 8
20 mM Tris, pH 8 — 1 ml Tris 1 M
500 mM NaCl — 5 ml NaCl 5 M

Low LiCl Buffer
0.25 M LiCl
1 % NP-40/Igepal

1 % deoxycholate
10 mM Tris, pH 8
1 mM EDTA, pH 8

(50 ml)

12.5 ml LiCl 1 M
0.5 ml NP-40/Igepal
5 ml deoxycholate 10 %
0.5 ml Tris 1M, pH 8
200 µl EDTA 250 mM, pH 8

High LiCl Buffer
0.5 M LiCl
1 % NP-40/Igepal
1 % deoxycholate
10 mM Tris, pH 8
1 mM EDTA, pH 8

(50 ml)
25 ml LiCl 1 M
0.5 ml NP-40/Igepal
5 ml deoxycholate 10 %
0.5 ml Tris 1M, pH 8
200 µl EDTA 250 mM, pH 8

IP elution Buffer
1 % SDS
0.1 M NaHCO3
(50 ml)
5 ml SDS 10 %
5 ml NaHCO3 1 M

Protocole d'Immunoprécipitation de Chromatine semi-natif

Culture : la veille, ensemencer environ 8.10^6 cellules par boite de 10cm gélatinée.
Prévoir environ 200.10^6 cellules (environ 20 boites 10cm).

Jour 1 :
Changer le milieu 2 heures avant.

Préparation de la chromatine

14. Crosslinker par ajout de 1% de formaldéhyde dans le milieu de culture (275 µl de formaldéhyde 36,5% dans 10 ml de milieu de culture). Incuber 10 min à Température Ambiante (TA).
15. Arrêter la réaction par ajout de glycine à une concentration finale de 0,125M (stock 1,25M préparé le jour même, 1 ml par boite). Incuber 5 min à Température Ambiante (TA).

A partir de cette étape, il faut travailler le plus possible dans la glace.

16. Rincer 2 fois les cellules avec du PBS 1X froid.
17. Scraper les cellules dans 2ml de PBS froid contenant des inhibiteurs (PMSF 1mM + Complete). Transférer dans un falcon. Conserver à 4°C.

Séparer en 4 tubes (environ 50.10^6 cellules par falcon).
Chaque falcon sera ensuite traité indépendamment. A l'issue de l'étape de lyse, les culots cellulaires seront poolés.

18. Centrifuger les cellules à 1000 g, 5 min, 4°C.
19. Aspirer le surnageant, et reprendre les cellules dans 10 ml de PBS froid. Faire 2 lavages PBS, centrifuger à 1000 g, 5 min, 4°C.

Perméabilisation des cellules

20. Reprendre les cellules dans 2 ml de solution I + inhibiteurs. Cette solution est hypotonique, et fait gonfler les cellules.
21. Ajouter 2 ml de solution II + inhibiteurs. Homogénéiser avec une p1000.
22. Incuber 10 min sur la glace. L'igépal permet de perméabiliser les cellules.
23. Centrifuger les cellules à 2000 g, 5 min, 4°C.

Digestion Mnase de la chromatine

24. Faire 2 lavages avec 10 ml de tampon KN, centrifuger à 2000 g, 5 min, 4°C. Pooler les cellules, afin d'avoir la même concentration cellulaire.
25. Reprendre chaque dans 2 ml de tampon KN (500 µl par IP) SANS inhibiteurs (inhibent l'action de la Mnase).
26. Bien resuspendre à la p1000, puis à la p200, sans faire de bulles. S'assurer qu'il n' y ait plus d'amas. Séparer de nouveau en 4 eppendorfs, transférer 500 µl dans chaque tube.
27. Digérer la chromatine à la Mnase, 100 unités pour 10.10^6 cellules. La solution stock Mnase est concentrée à 200u/µl, préparer une dilution à 20 u/µl dans du tampon KN (afin de prélever un volume précis). Bien homogénéiser à la pipette.
28. Incuber à 37°C, 10 min au bain-marie. Retourner une fois au cours de la digestion.
29. Stopper la digestion Mnase en transférant les tubes sur la glace. Ajouter 4 µl d'EDTA 0,5M pH8 (4mM final d'EDTA).
30. Ajouter les inhibiteurs de protéases (10 µl de Complete 50X, et 5 µl PMSF 0,1M). Laisser sur la glace 15 min environ.

Sonication de la chromatine digérée.

31. Sur le Diagenode Bioruptor, procéder à la sonication, 4 cycles de 20 sec ON/40 sec OFF (Il faut de plus que le bain soit très froid).
32. Centrifuger la chromatine sonique à 13000 rpm, 10 min, 4°C. Garder le surnageant, les pooler puis transférer dans un nouveau tube.
33. Prélever un aliquot de 20 µl afin de vérifier la sonication et quantifier l'ADN (input).
34. Réverser le crosslink de l'input : ajouter 480 µl d'H20 dd, 20 µl de NaCl 5M, et 15 µg de RNase A. Incuber 4 heures ou ON à 65°C.

Si l'on vérifie l'immunoprécipitation de la chromatine par WB, prélever 10 µl. Conserver sur la glace, jusqu'à la fin de l'expérience.

Première immunopécipitation Flag.

Préparation des billes Flag.

Par IP (4 IP pour chaque protéine, environ 50.10^6 cellules par IP), préparer 25 µl de billes sèches Flag.

1. Vortexer le flacon de billes Flag-M2-agarose (Sigma), et prélever 50 µl de slurry (50% billes sèches).
2. Ajouter 10 ml de tampon TEGN, centrifuger à 2000 rpm, 5 min, 4°C.
3. Laver une deuxième fois dans du tampon TEGN.
4. Laver 2 fois avec 3ml de tampon TKNSE.
5. Reprendre les billes dans 200 µl (4 volumes de billes) de tampon TKNSE, et ajouter à la chromatine centrifugée 50 µl de billes (1 volume).
6. Incuber ON à 4°C, sur la roue.

Jour 2 :

Lavages.

1. Centrifuger les IP à 2000 rpm, 5 min, 4°C. Prélever 1/400^e de la fraction non liée (5 µl). Resuspendre dans 1 ml de TEGN et transférer dans un falcon 14 ml.
2. Laver les billes avec 10 ml de TEGN, faire 8 lavages, centrifuger à 2000 rpm, 2 min, 4°C.
3. Au cours du dernier lavage, regrouper les billes dans un seul et unique tube pour chaque condition. Reprendre dans 500 µl de TEGN et transférer dans un eppendorf.
4. Centrifuger à 2000 rpm, 2 min, 4°C.

Elution.

5. Reprendre le culot dans 300 µl de TEGN + inhibiteurs. Ajouter 100 µl de peptide Flag (stock 4 mg/ml).Eluer 6 heures à TA.
6. Récupérer l'éluat en centrifugeant à 2000 rpm, 2 min 4°C, le garder dans la glace, durant la seconde élution.
7. Procéder à la deuxième élution comme précédemment. Eluer ON, à 4°C, sur la roue.

Jour 3 :

1. Transférer les éluats Flag (issus des 2 élutions, sur la journée, et sur la nuit) et les billes sur une colonne Pierce, avec filtre. Centrifuger 30 sec au maximum. Récupérer l'éluat Flag.
2. Prélever 1/80^e de l'éluat pour vérifier l'IP et l'élution Flag en WB. Pour le contrôle en qPCR, prélever ½ IP Flag.
3. Ajouter 200 µl de tampon Laemmli SDS 3X sur les billes, retenues sur la colonne. Incuber à 100°C, 10 min. Collecter l'éluat après centrifugation.

Deuxième immunopécipitation : HA.

Préparation des billes HA.

Préparer les billes HA-agarose comme pour les billes Flag. Prélever 25 µl seulement de billes, car on ne compte qu'une IP à cette étape.
Appliquer l'éluat issu de la première IP sur les billes, et incuber jusqu'au lendemain à 4°C sur la roue.

Jour 4 :

Prélever 1/80^{e} de la fraction non liée, jeter le reste du surnageant.

Lavages.

Procéder comme pour les billes Flag, 8 lavages TEGN, 10 ml.

Elution.

1. Reprendre le culot dans 300 µl de TEGN + inhibiteurs. Ajouter 100 µl de peptide HA (stock 4 mg/ml).Eluer 2heures à 4°C. Puis éluer à TA pendant 4 heures environ.
2. Collecter l'éluat après centrifugation, et prélever 1/40^{e}.
3. Ajouter le tampon d 'élution SDS/NaHCO3 directement sur les billes, 2 fois 250 µl, 15 min à chaque fois. Prélever 1/40^{e} pour contrôler l'élution SDS.
4. A la fin de la deuxième élution, collecter l'éluat, et ajouter 200 µl de tampon Laemmli SDS sur les billes. Chauffer à 100°C, 10 min.

Réversion du crosslink :

Reverser le crossling du contrôle de la première IP, ainsi que des échantillons issus du tandem. Ajouter 20 µl de NaCl 5M, et 15 µg de RNase A. Incuber ON à 65°C.

Jour 5 :

7. Ajouter 10 µl d'EDTA 0,5M pH8, 50 µl de Tris pH8 1M et 4 µl de protéinase K 10 mg/ml. Incuber 1 heure à 50°C.
8. Extraire l'ADN par une purification Phénol/Chloroforme.
9. Précipiter l'ADN en ajoutant 1/10 vol d'Acétate de Sodium (NaAc) 3M, pH5,2, 2-2,5 vol d'éthanol 100% et 10 µg de Glycogène. Incuber 30 min à -20°C, puis centrifuger à 13000, 20 min, 4°C. Faire un lavage éthanol 70%. Sécher le culot.

Les échantillons d'Input, de ChIP simple, et de première étape de ChIP sont repris dans 50 µl H2O dd.
Les échantillons issus du ChIP tandem sont repris dans 20 µl d'H20 dd. Les quantifier en utilisant le Picogreen.

Quantification de la chromatine et vérification de la digestion

Vérifier la digestion en faisant migrer un échantillon de 5 µl sur un gel d'électrophorèse agarose 1,5%.

Solutions

Solution I (500ml)

0,3M Sucrose	51,3g de Sucrose
60mM KCl	15ml KCl 2M
15mM NaCl	1,5ml NaCl 5M
5mM MgCl2	2,5ml MgCl2 1M
0,1mM EGTA	0,5ml EGTA 0,1M
15mM Tris-HCl	7,5ml Tris-HCl 1M pH 7,5

Solution II (10ml) l'Igepal est à rajouter au dernier moment.

9,2ml de Solution I	
0,8% Igepal	0,8ml d'Igepal 10%

Tampon TKNSE (500ml)

20mM Tris HCl pH7,5	10ml Tris HCl 1M pH7,5
15mM KCl	3,75ml KCl 2M
60mM NaCl	6ml NaCl 5M
0,34M Sucrose	58g Sucrose (M=342,3g/mole)
4mM EDTA	4 ml EDTA 0,5M PH8,0

Tampon MNase KN (500ml)

20mM Tris HCl	10ml Tris HCl 1M pH7,5
15mM KCl	3,75ml KCl 2M
60mM NaCl	6ml NaCl 5M
0,34M Sucrose	58g Sucrose (M=342,3g/mole)
1mM CaCl2	1ml de CaCl2 500mM

TEGN (500ml)

20mM Tris HCl	10ml Tris HCl 1M pH7,5
150mM NaCl	15ml NaCl 5M
3mM MgCl2	1,5ml MgCl2 1M
0,1mM EDTA	100ul EDTA 0,5M
10% Glycérol	50 ml Glycérol 100%
0,01% NP40	500 µl NP40 ou Igepal 10%

IP elution Buffer	(50 ml)
1 % SDS	5 ml SDS 10 %
0.1 M NaHCO3	5 ml NaHCO3 1 M

Protocole de RT-PCR

Extraction des ARNs.

L'ARN total est extrait des cellules non-taggées (46C) avec du Trizol (Invitrogen), suivant les instructions données par le fabriquant.

Les petits ARNs sont dans une certaine mesure, solubles dans l'éthanol 75%. On effectue alors un lavage à l'éthanol 80%, sans resuspendre le culot d'ARN.

Quantifier au nanodrop (1 unité DO 260=40 µg/µl), et vérifier l'intégrité de l'ARN en faisant migrer un échantillon sur un gel agarose 1%, TAE 1X (Attention à bien nettoyer les cuves avant la migration, le tampon de migration doit être changé). Deux bandes doivent être clairement visibles, l'ARN 28S et 18S.

Transcription inverse (RT).

Traitement DNase.

1. Eliminer l'ADN génomique contaminant, présent dans l'échantillon ARN.

ARN	5 µg
tampon 10X	5 µl
Dnase RQO1 1u/µgARN (Promega)	5 µl
H2O qsp 50 µl	37,5µl

2. Incuber 1 heure à 37°C.
3. Ajouter 1 µl de solution STOP, et inactiver la Dnase 10 min à 65°C.
4. Ajouter 5 µg de glycogène, 50 µl d'isopropanol.
5. Précipiter 30 min à -20°C.
6. Centrifuger à 14000 rpm, 20 min, 4°C.
7. Faire un lavage éthanol 80%, centrifuger à 14000 rpm, 10 min, 4°C.
8. Sécher le culot, resuspendre dans 10 µl (on peut vérifier à cette étape l'intégrité de l'ARN, sur gel agarose).

RT.

Protocole de la SuperScript II (SS II, Invitrogen).
Ne pas oublier le contrôle RT-, c'est-à-dire un échantillon ARN traité à la Dnase, sans ajout de SS II, afin de contrôler si il reste de l'ADN génomique.

1. Ajouter les réactifs suivants à l'ARN traité à la Dnase :

ARN 5 µg	10 µl
Random hexamers 10 µM (Invitrogen)	2 µl
dNTP 5mM chaque	1 µl

2. Chauffer à 65°C, 5min.

3. Mettre dans la glace, puis centrifuger brièvement.
4. Ajouter :

First strand buffer 5X	4 µl
DTT 0,1M	2 µl

5. Incuber 2 min à 25°C.
6. Ajouter 1 µl de SS II.
7. Incuber - 10 min à 25 °C
   - 1 heure à 42°C
8. Inactiver l'enzyme 15 min à 70°C.
9. Ajouter 1 µl de RNase H (Invitrogen) et incuber 20 min à 37°C.

On peut ensuite amplifier l'ADNc en PCR, ou en qPCR.

PCR.

ADNc		2 µl
Amorces	-Forward 10 µM	0,375 µl (150 mM final)
	-Reverse 10 µM	0,375 µl
Tampon PCR 10 X		1 µl
MgCl2 (25 mM)		0,5 µl
dNTP 10 mM		0,2 µl
Taq		
H2O qsp 25 µl		

Programme PCR

2' à 94°C
15'' à 94°C
30'' à 60°C } 30 cycles
45'' à 72°C
4' à 72°C

Déposer les produits obtenus sur un gel agarose TAE 1X, dont le pourcentage agarose est adapté à la taille des amplicons attendus.

qPCR.

Préparer différentes dilutions d'ADNc : 1/5^{e} à 1/50000^{e}.
L'ARN U6 est utilisé comme contrôle interne pour la normalisation.

ADNc		
Amorces	-F 10 µM	
	-R 10 µM	
Mix qPCR (Eurogentec)		
10 µl		
0,375 µl		
0,375 µl		
12,5 µl		

H2O qsp 25 µl

Programme qPCR

5’ à 95°C
15’’ à 95°C
30’’ à 60°C } 40 cycles

Melting curve
3’ à 95°C
1’ à 55°C
10’’ à 55°C

Amorces utilisées pour le recombineering

Tableaux 12. Oligonucléotides utilisés pour la construction des vecteurs de recombinaisons, et pour la sonde utilisée lors du génotypage des lignées cellulaires étiquetées.

RPC4

RefSeq Gene Polr3d

RefSeq: NM_001164082.1
Description: Mus musculus polymerase (RNA) III (DNA directed) polypeptide D (Polr3d), transcript variant 2, mRNA.
DNA-directed RNA polymerase III subunit RPC4

chr14 : 70828555-70848555

Sonde Bam HI	5' G CAT GAA TTC ACC CTT CCA GTT GAA GAG CA 3'	5' G CAT GAA TTC CCT CCG ACT AGA GCA ACC AC 3'
5' Miniarm (Not I/Spe I)	5' G CAT GAA TTC GCG GCC GCG GTG ACG GAG TCC AAA GCT A 3'	5' G CAT GAA TTC ACT AGT AGG CAG CAC ATG AAT GAA CA 3'
3' Miniarm (Spe I/Bam HI)	5' G CAT GAA TTC ACT AGT TCA AGG CCA GTC TGG GAT AC 3'	5' G CAT GAA TTC GGA TCC TGC ATC CAG AGA CAA TGG AG 3'
5' homologie (Sal I/Asc I)	5' G CAT GAA TTC GTC GAC TTC CTG GGA AAC AGA AAT GG 3'	5' G CAT GAA TTC GGC GCG CCC CGG TGT TTG TGA TCC AAG AG 3'
3' homologie (Bam HI/Not I)	5' G CAT GGA TCC **TGA** GAT GGA CAG ATG GAG GAG 3'	5' G CAT GGA TCC GCG GCC GCA GGG AGC TGG GAC TGA GTT T 3'

RPC1

RefSeq Gene Polr3a

RefSeq: NM_001081247.1
Description: Mus musculus polymerase (RNA) III (DNA directed) polypeptide A (Polr3a), mRNA.
DNA-directed RNA polymerase III subunit RPC1

chr14 : 25267916-25306268

Sonde Kpn1	5' G CAT GAA TTC TGC AGC TAG GGA GGG AGA TA 3'	5' G CAT GAA TTC CTA CCC ATA CCC CAC ACA CC 3'
5' Miniarm (Not I/Spe I)	5' G CAT GAA TTC GCG GCC GCG CCA CTG TCA AGT CCC GTA T 3'	5' G CAT GAA TTC ACT AGT TTT GCC TCC TGA AGA GCA GT 3'
3' Miniarm (Spe I/Bam HI)	5' G CAT GAA TTC ACT AGT CTG CAG GGA TGA TCA GAA CC 3'	5' G CAT GAA TTC GGA TCC AAT TCT TCT GCA GCC CAC AG 3'
5' homologie (Sal I/Asc I)	5' G CAT GAA TTC GTC GAC GGC TGA GCT GAC TTG GTT TC 3'	5' G CAT GAA TTC GGC GCG CCT GTA ACA AGA GGG ATG TGG AA 3'
3' homologie (Bam HI/Not I)	5' G CAT GGA TCC **TAG** TTT GTG GCA AGA GGT CAC 3'	5' G CAT GGA TCC GCG GCC GCC ACA CTG GAC CTC ACA TTG G 3'

TFIIIC220

RefSeq Gene Gtf3c1

RefSeq: NM_207239.1
Description: Mus musculus general transcription factor III C 1 (Gtf3c1), mRNA.

chr7 : 132784468-132851202

Sonde Eco RI	5' G CAT GAA TTC CAT GCT GTG TTG AGC TTG GT 3'	5' G CAT GAA TTC CCG ACA GCT GAC TCT TGT TTC 3'
Sonde Interne	5' AAC CAG GTG CTG CTG AGA GT 3'	5' CTG GAG AAC TGG TGC AGT CA 3'
5' Miniarm (Not I/Spe I)	5' G CAT GAA TTC GCG GCC GCA ACT CCA TTA GCC CCC AGA T 3'	5' G CAT GAA TTC ACT AGT TGG TCA GCA CGC TCT TTC TT 3'
3' Miniarm (Spe I/Bam HI)	5' G CAT GAA TTC ACT AGT CAG AAC TTT CC3'	5' G CAT GAA TTC GGA TCC TCT TTT GAG CAC GAT GTT CG 3'
5' homologie (Sal I/Asc I)	5' G CAT GAA TTC GTC GAC GCA CCC ACA CCT TCC ATA CT 3'	5' G CAT GAA TTC GGC GCG CCT AAA TGG ATC CAC TTG TTC CA 3'
3' homologie (Bam HI/Not I)	5' G CAT GGA TCC **TAG** ACG CTG CCC CAG GCC AGC 3'	5' G CAT GGA TCC GCG GCC GCA ACC TAT AGG CCT GGC TTC TG 3'

TFIIIC110

RefSeq Gene Gtf3c2

RefSeq: NM_027901.2
Description: Mus musculus general transcription factor IIIC, polypeptide 2, beta (Gtf3c2), mRNA.

chr5 : 31458379-31482517

Sonde Eco RI	5' G CAT GAA TTC CTA CCT CCG GGA GAC TGA CA 3'	5' G CAT GAA TTC TTC CAT CCC CCA AGT AAA CA 3'
5' Miniarm (Not I/Spe I)	5' G CAT GAA TTC GCG GCC GCT GGC CCC TTG TAG ACT CAT C 3'	5' G CAT GAA TTC ACT AGT GAC CCC ACA GTT CCA ATC TC 3'
3' Miniarm (Spe I/Bam HI)	5' G CAT GAA TTC ACT AGT ATG ACT AGA AGC CGG GTG TG 3'	5' G CAT GAA TTC GGA TCC AAG CTC CCA AGT GAG CAG AA 3'
5' homologie (Sal I/Asc I)	5' G CAT GAA TTC GTC GAC TCG TTT TGG GTG GTT TTG AT 3'	5' G CAT GAA TTC GGC GCG CCG GGA TTG GGG AGA AGG CA 3'
3' homologie (Bam HI/Not I)	5' G CAT GGA TCC **TAG** TCT GGG CCA CAC AGA ACT 3'	5' G CAT GGA TCC GCG GCC GCA TCG GAA GGC AAT AAC ATC G 3'

TFIIIC90

RefSeq Gene Gtf3c4

RefSeq: NM_172977.3
Description: Mus musculus general transcription factor IIIC, polypeptide 4 (Gtf3c4), transcript variant 1, mRNA.

chr2 : 28677820-28695880

Sonde Eco RI	5' G CAT GAA TTC GGC TCG GGT TAT GAA ATT GT 3'	5' G CAT GAA TTC GGA AGG CAT GGA CAT CAG TT 3'
5' Miniarm (Not I/Spe I)	5' G CAT GAA TTC GCG GCC GCG AGA ACA CTG GAC AGC GTC A 3'	5' G CAT GAA TTC ACT AGT TGC CTT ATA GGC ATG GGG TA 3'
3' Miniarm (Spe I/Bam HI)	5' G CAT GAA TTC ACT AGT ATG ACT AGA AGC CGG GTG TG 3'	5' G CAT GAA TTC GGA TCC AAG CTC CCA AGT GAG CAG AA 3'
5' homologie (Sal I/Asc I)	5' G CAT GAA TTC GTC GAC GCT ACT TCC TGG CTG GTC TG 3'	5' G CAT GAA TTC GGC GCG CCG AAG ACA GGA GAA TCA CAG AAA GG 3'
3' homologie (Bam HI/Not I)	5' G CAT GGA TCC **TGA** GTT CAG TGA GGA GGA TGG 3'	5' G CAT GGA TCC GCG GCC GCT CTG GAG CCA GAG ACA GTC C 3'

TFIIIC63

RefSeq Gene Gtf3c5

RefSeq: NM_148928.2
Description: Mus musculus general transcription factor IIIC, polypeptide 5 (Gtf3c5), mRNA.

chr2 : 28421783-28438799

Sonde Eco RI	5' G CAT GAA TTC CAT GCT GTG TTG AGC TTG GT 3'	5' G CAT GAA TTC CCG ACA GCT GAC TCT TGT TTC 3'
5' Miniarm (Not I/Spe I)	5' G CAT GAA TTC GCG GCC GCA GCC AGG GCT ACA CAG AGA A 3'	5' G CAT GAA TTC ACT AGT CCT GAC CTC ACC TTG AGT TTG 3'
3' Miniarm (Spe I/Bam HI)	5' G CAT GAA TTC ACT AGT GCC TGA GTA CCA GCT CCA AC 3'	5' G CAT GAA TTC GGA TCC GTC CCCCAA CTC TGT GAA GA 3'
5' homologie (Sal I/Asc I)	5' G CAT GAA TTC GTC GAC TTC CTG GCT GTT TGA GGT GT 3'	5' G CAT GAA TTC GGC GCG CCC ACG TAA TCC AGA ATC TCT GT 3'
3' homologie (Bam HI/Not I)	5' G CAT GGA TCC **TGA** GGA AGC GCT GTT CCC AGG 3'	5' G CAT GGA TCC GCG GCC GCC CTG AAT GAG GGA AGG AAC A 3'

BRF1

RefSeq Gene Brf1

RefSeq: NM_028193.3
Description: Mus musculus BRF1 homolog, subunit of RNA polymerase III transcription initiation factor IIIB (S. cerevisiae) (Brf1), mRNA.
transcription factor IIIB 90 kDa subunit

chr12 : 114198073-114238832

Sonde Nco1	5' G CAT GAA TTC TGA TCC GAT CTG AAG CAT CC 3'	5' G CAT GAA TTC CAC TCC TCA CCG CAC TAA CA 3'
5' Miniarm (Not I/Spe I)	5' G CAT GAA TTC GCG GCC GCC TTC ATG CCC TTC ACC CTT A 3'	5' G CAT GAA TTC ACT AGT CTT GGC CCA CAG GTA AAG AA 3'
3' Miniarm (Spe I/Bam HI)	5' G CAT GAA TTC ACT AGT GGT GAG CTG AGG GCA GAT AG 3'	5' G CAT GAA TTC GGA TCC GCC TGA TTC CAC CTC ATC AT 3'
5' homologie (Sal I/Asc I)	5' G CAT GAA TTC GTC GAC TCT TAA CCG CTG AGC CAT CT 3'	5' G CAT GAA TTC GGC GCG CCG TAG CCA TCA TCT TCA TCA CC 3'
3' homologie (Bam HI/Not I)	5' G CAT GGA TCC **TGA** GTG GAA CTC AAG GCC AGG 3'	5' G CAT GGA TCC GCG GCC GCC AG ACC CAG CTA CTC TCT G 3'

BRF2

RefSeq Gene Brf2

RefSeq: NM_025686.2
Description: Mus musculus BRF2, subunit of RNA polymerase III transcription initiation factor, BRF1-like (Brf2), mRNA.
transcription factor IIIB 50 kDa subunit

chr8 : 28234304-28239104

Sonde Eco RI	5' G CAT GAA TTC CCA GAG GAA AAA CGT GGT GA 3'	5' G CAT GAA TTC TCT CAG ACG AAG GGC AGT TT 3'
5' Miniarm (Not I/Spe I)	5' G CAT GAA TTC GCG GCC GCG ATG GTG CAT CCG GAA TAA G 3'	5' G CAT GAA TTC ACT AGT TTT CGT TCA TTC ACC AAC CA 3'
3' Miniarm (Spe I/Bam HI)	5' G CAT GAA TTC ACT AGT CAC AGC TCA CAA CCA TT 3'	5' G CAT GAA TTC GGA TCC TCT GCT TGC AGT GCT CAG AT 3'
5' homologie (Sal I/Asc I)	5' G CAT GAA TTC GTC GAC GGC CTG GCT ACA GGT TCT AA 3'	5' G CAT GAA TTC GGC GCG CCG GGA GGG TTA GGG ACA CGC AT 3'
3' homologie (Bam HI/Not I)	5' G CAT GGA TCC **TGA** TGC ACG TTC TTT AAG GTG 3'	5' G CAT GGA TCC GCG GCC GCG CAA AAC TCT TTG AGC GAC A 3'

Stratégie de vérification par restriction de l'insertion correcte des étiquettes

Protéines	Restriction	Allèle sauvage	Allèle étiqueté Néo +	Allèle étiqueté Néo -
RPC1	Digestion Eco RI (sonde Eco RI)	7,7 kb	9,6 kb	-
	Digestion Bam HI (sonde NEO)	-	7,5 kb	Absence de bande
RPC4	Digestion Bam HI (sonde Bam HI)	14,1 kb	10,6 kb	8,6 kb
BRF1	Digestion Nco1 (sonde Nco1)	10,8 kb	7 kb	
	Digestion Bam HI (sonde NEO)	-	3 kb	Absence de bande
BRF2	Digestion Eco RI (sonde Eco RI)	17,6 kb	10,3 kb	8,3 kb
TFIIIC220	Digestion Eco RI (sonde Eco RI)	13,2 kb	8,6 kb	
	Digestion Eco RI (sonde Interne)	13,2 kb	6,6 kb	4,8 kb
TFIIIC110	Digestion HindIII (sonde HindIII)	8,2 kb	5,7 kb	3,7 kb
TFIIIC90	Digestion Bam HI (sonde Bam HI)	20,5 kb	7,5 kb	-
	Digestion Bam HI (sonde Bam HI)	-	14,9 kb	-
TFIIIC63	Digestion Bam HI (sonde Bam HI)	10,8 kb	6,8 kb	

Tableau 13. Tailles des bandes attendues lors du génotypage des lignées cellulaires étiquetées pour les différentes protéines.

Tableau 14. Amorces utilisées pour la qPCR et la RT-qPCR

Vérification de l'occupation aux nouveaux gènes

gènes	amorces	séquence Forward et Reverse	
ARBP	SG55	SG109 AGCAAGGAAGCTGGATCAGA	SG110 ACCTCTACCTCCCCATTGCT
2819	lc31	LC031f TGGCCCTTTCTCTTCTTGTCA 21	LC031r GAACCCTGGCACACTGTGTTAA 22
1283	lc40	LC040f GGAGGAGGCCCCTGGATT 18	LC040r CAGGTTTTCTTTCCCCCTTAGG 22
1263	lc43	LC043f TGCACTCGGCAGTCAGAAAG 20	LC043r AAGTGAACAGACCTGAGACGAAATT 25
2039	lc47	LC047f TTGAGGGAAAGAGAAGGAACACA 23	LC047r TTGTTAACATATGGAAGACTGCAAGA 26
1788	lc55	LC055f CAATCATCCTGAGTATGACTCCAGTT 26	LC055r CCAGTGTGTAAGACGCAGGTAGA 23
3752	lc59	LC059f TGTGAGTGTGTCAGACATTGGTACA 25	LC059r CCTCTCCCACTCACAAATTCG 21
2774	lc104	Lc104f GACCTGCACAGAACACCTCA 20	Lc104r TGAGCAGGCTGAGAAAAGGT 20
1252	lc124	Lc124f CCACTAGGAAATGGGCAGAA 20	Lc124r CATGGGCGTCTCTAGCATTT 20
1284	lc131	Lc131f CTGGCTTTCGGGCTTTACTA 20	Lc131r GCCCTGCATAACAAGAAAGG 20
1382	lc141	Lc141f CGGCCACAGCTCACTCTTAT 20	Lc141r GATGGCTGAAAATGGGAATG 20

Amorces utilisées pour vérifier le ChIP

gènes	amorces	séquence Forward et Reverse	
H1	PCR SG59	SG117 CGCTGTGCTTTGTGGGAAAT	SG118 GCTTCCTCCGCCCACTTT
tRNA-Valine	PCR SG65	SG129 CCCAGCGTCACCACAGTTCT	SG130 TGCGTGTAAAGCAGACGTGAT
tRNA-Leucine	PCR SG61	SG121 GTCCTCGGCTCTTTGTTTGG	SG122 ATATCCGCGTGGGTTCGAA
5S	PCR SG48	SG095 GCCATACCACCCTGAACG	SG096 AGCCTACAGCACCCGGTATT
ARBP	PCR SG55	SG109 AGCAAGGAAGCTGGATCAGA	SG110 ACCTCTACCTCCCCATTGCT
U6	PCR SG49	SG097 CGCTTCGGCAGCACATATAC	SG098 AAAATATGGAACGCTTCACGA
Untr	Untr1-Untr2	Untr1 TCAGGCATGAACCACCATAC	Untr2 AACATCCACACGTCCAGTGA

gènes	Localisation
H1	chr14:51,427,431-51,427,530
tRNA-Valine	chr19:12086362-12086461
tRNA-Leucine	chr19:12085509-12085608
5S	amorces localisées dans la région codante ciblant un grand nombre de gènes 5S
ARBP	chr5:116010573-116010673
U6	amorces localisées dans la région codante ciblant un grand nombre de gènes U6
Untr	Chr6: 120740575–120740790

Vérification de la mappabilité des gènes

gènes	amorces	séquence Forward et Reverse	
ARN 7SL	lc177	lc177f CGCCGCAGCTCTAGTATCTC 20	lc177r CCCGAGTAGCTGGGACTACA 20
	lc182	lc182f ACCCCTCCTTAGGCAACCT 19	lc182r CCGCACTAAGATCTGCATCA 20
ARN 7SK	lc184	lc184f ATCAACCCTGGCGATCAAT 19	lc184r AGCCTGCTTACTCTCGGATG 20
	lc186	lc186f GCGTTCACTGCTGCTTTCTACA 22	lc186r CGCCCTCACATCCTGGACTA 20
U6	lc190	lc190f AAATTCGTGAAGCGTTCCAT 20	lc190r GCCTCCAATAGGATGTTAGGG 21
	lc193	lc193f TCCATCCCATGTTCCTGAAT 20	lc193r AGGGGCCATGCTAATCTTCT 20
HY3	lc197	lc197f AGGCCTAACTTTCGGTTGGT 20	lc197r GAGCGGAGAAGGAACAAAGA 20
	lc200	lc200f GGCTGTTTCCAGCTAATTGTTCA 23	lc200r GGGAAGCTTCTTTCTCTCAACCT 23
tRNA-Valine (GTY)	lc204	lc204f ATTGCGAGAGATCCATTGCT 20	lc204r CGTGTTAGGCGAACGTGATA 20
tRNA-Leucine (TAA)	PCR SG61	SG121 GTCCTCGGCTCTTTGTTTGG	SG122 ATATCCGCGTGGGTTCGAA
tRNA-Lysine (AAA)	PCR SG63	SG125 CCGGAAATACAGGAGCCTAAAA	SG126 CCGGATAGCTCAGTCGGTAGA
tRNA-Proline (CCY)	lc209	lc209f GAGAGGTCCTGGGTTCAAATC 21	lc209r TGCAAAGAATCAACACTCTGAA 22
tRNA-Lysine (AAG)	lc210	lc210f AGAGAGGAGTTGCCCAGCTA 20	lc210r ACAACATGGGGCTCCAAC 18

gènes	amorces	Etat	Localisation
ARN 7SL	lc177	bound	chr12:70260282-70260581
	lc182	unbound	chr1:192306734-192307036
ARN 7SK	lc184	bound	chr9:78023109-78023440
	lc186	unbound	chr5:140320884-140321144
U6	lc190	bound	chr10:79371940-79372045
	lc193	unbound	chr15:96,501,279-96,503,279
HY3	lc197	bound	chr6:47731623-47731723
	lc200	unbound	chr4:129408260-129408350
tRNA-Valine (GTY)	lc204	bound	chr11:48670139-48670214
tRNA-Leucine (TAA)	PCR SG61	bound	chr19:12085509-12085608
tRNA-Lysine (AAA)	PCR SG63	bound	chr19:12084157-12084256
tRNA-Proline (CCY)	lc209	unbound	chr1:78294370-78294444
tRNA-Lysine (AAG)	lc210	unbound	chr3:3123739-3123816

RT-qPCR				
Unité de transcription	**amorces**	**séquence Forward et Reverse**		**Présence d'un transcrit ?**
2819	Lc31	LC031f TGGCCCTTTCTCTTCTTGTCA 21	LC031r GAACCCTGGCACACTGTGTTAA 22	Ok
1283	Lc40	LC040f GGAGGAGGCCCCTGGATT 18	LC040r CAGGTTTTCTTTCCCCCTTAGG 22	Ok
1263	Lc43	LC043f TGCACTCGGCAGTCAGAAAG 20	LC043r AAGTGAACAGACCTGAGACGAAATT 25	Ok
2039	Lc45	LC045f GGTGACCTGAGTTGCCAAAAG 21	LC045r AGGACCTGCCTCTCATATCCTAAG 24	Ok
1788	Lc55	LC055f CAATCATCCTGAGTATGACTCCAGTT 26	LC055r CCAGTGTGTAAGACGCAGGTAGA 23	Ok
3752	Lc59	LC059f TGTGAGTGTGTCAGACATTGGTACA 25	LC059r CCTCTCCCACTCACAAATTCG 21	Ok
2798	Lc93	Lc093f GGCCTAAAGGCTGAGACACA 20	Lc093r ATGTCGCCCTAGGGAAAGTC 20	Ok
2800	Lc94	Lc094f GCACAGAGCAGAAAAGTGTCC 21	Lc094r CCCAGCCAGAGTAAGGGATT 20	Non
2832	Lc99	Lc099f GCTGATGTCCTTGGGCTCTTT 21	Lc099r CAGGGAGATCTAGGAAGGCTTCT 23	Ok
2774	Lc104	Lc104f GACCTGCACAGAACACCTCA 20	Lc104r TGAGCAGGCTGAGAAAAGGT 20	Ok
2843	Lc106	Lc106f GCCTCAAAACTCAGGACAGG 20	Lc106r GGTTGAGGAATGGGGTTCTT 20	Ok
3686	Lc121	Lc121f GTGGGCTTTGCTAGAAGCTG 20	Lc121r TTGTGTTCCCGATGATGTGT 20	Ok
1247	Lc123	Lc123f CACACAGGTGCTCTGCAAAT 20	Lc123r CACTTGGCCTTTTTGGTAGC 20	Ok
1252	Lc124	Lc124f CCACTAGGAAATGGGCAGAA 20	Lc124r CATGGGCGTCTCTAGCATTT 20	Ok
1284	Lc131	Lc131f CTGGCTTTCGGGCTTTACTA 20	Lc131r GCCCTGCATAACAAGAAAGG 20	Ok
2119	Lc134	Lc134f AACAGTTACCTGGGATTGAACC 22	Lc134r GTATTTACCTTGCCTGTCTTAGGG 24	Ok
1382	Lc141	Lc141f CGGCCACAGCTCACTCTTAT 20	Lc141r GATGGCTGAAAATGGGAATG 20	Ok

Summary

In eukaryotes, RNA polymerase (RNAP) III transcribes the tRNAs, the 5S ribosomal RNA and a half a dozen known untranslated RNA. Mammalian genome contains several thousand of repeated elements, the Short interspersed repetitive elements (SINE). *In vitro,* they are transcribed by RNAP III. RNAP III transcription levels determine cell growth and proliferation and, importantly, its deregulation is associated with cancer. Looking at the genome-wide distribution of RNAP III and its transcription factors, TFIIIB and TFIIIC, we develop a highly specific tandem ChIP-sequencing method. We have determined the set of genes that are transcribed by RNAP III in mouse embryonic stem cells. We discovered that not all known class III genes were transcribed in ES cells. We also observed that RNAP III and its transcription factors were present at thirty unannotated sites on the mouse genome, only one of which was conserved in human. Only a couple of hundreds of SINEs out of more than half a million are associated with RNAP III in mouse ES cells. Our study reveals numerous 'TFIIIC-only' sites, called ETC for extra-TFIIIC loci in yeast. These sites are correlated with association of CTCF and the cohesin. Cohesin has been shown to occupy sites bound by CTCF and to contribute to DNA loop formation associated with gene repression or activation. This observation suggests that TFIIIC may play a role in chromosome organization in mouse. We also demonstrated that TCEA1, the ubiquitous isoform of TFIIS RNAP II elongation factor, is associated with active class III genes suggesting that TFIIS is a RNAP III transcription factor in mammals. Finally, the distribution of TFIIS on RNAP II-transcribed genes indicated that its recruitment does not control the transition of RNAP II paused at genes 5' end into elongation.

Résumé

Chez les eucaryotes, l'ARN polymérase (RNAP) III transcrit les ARN de transferts, l'ARN ribosomique 5S, et plusieurs douzaines d'autres ARNs non traduits. Le génome des mammifères contient plusieurs milliers d'éléments répétés, les SINEs. *In vitro*, leur transcription dépend de la RNAP III. Le taux de transcription de la RNAP III détermine la croissance et la prolifération cellulaires, sa dérégulation a été associée à de nombreux cancers. Afin de caractériser la distribution sur l'ensemble du génome de la RNAP III et de ses facteurs de transcription TFIIIB et TFIIIC, nous avons développé un protocole très spécifique de ChIP-seq en tandem. Nous avons déterminé l'ensemble des gènes liés par la RNAP III dans les cellules souches embryonnaires de souris. Cet ensemble est bien inférieur au nombre de gènes prédits dans le génome. Nous avons également observé la RNAP III et ses facteurs de transcription liés à 30 régions non annotées, seule une d'entre elles est conservée chez l'humain. Un très faible nombre de SINEs sur un demi-million prédits est associé à la RNAP III. Notre étude révèle de nombreux sites liés uniquement par TFIIIC, nommés « extra-TFIIIC loci », ETC chez la levure. Ces sites sont associés à la protéine CTCF, et à la cohésine. La cohésine occupe les sites liés par CTCF, et contribue à la formation de boucles ADN, associées à la répression ou à l'activation de l'expression des gènes. Ces données suggèrent que TFIIIC peut jouer un rôle dans l'organisation de l'architecture chromosomique chez les souris. Nous avons également démontré que TCEA1, l'isoforme ubiquitaire de TFIIS, le facteur d'élongation de la RNAP II, est associée aux gènes actifs de classe III. Ceci suggére que TFIIS est un facteur de transcription de classe III. Finalement, la distribution de TFIIS aux gènes de classe II indique que le recrutement de TFIIS n'est pas suffisant pour contrôler la transition de la RNAP II pausée en 5' des gènes en élongation.