UNIVERSITÉ DE STRASBOURG

Thèse présentée pour obtenir le grade de
Docteur de l'Université de Strasbourg

Discipline : Sciences du Vivant
Spécialité : Bioinformatique

par Yannick-Noël ANNO

Etablissement d'une architecture bioinformatique et biostatistique d'intégration et d'analyse des données génomiques, épigénétiques et phylogénétiques du génome humain.

Application aux sites de fixation du facteur de transcription hStaf/ZNF143

Soutenue publiquement le 24 septembre 2010

Membres du jury

Co-directeur de thèse :	M. Philippe CARBON, professeur, UdS
Co-directeur de thèse :	Mme Odile LECOMPTE, maitre de conférences , UdS
Examinateur Interne :	M. Jean-Marie WURTZ, professeur, UdS
Examinateur Externe :	M. Thierry LEVEILLARD, directeur de recherche, INSERM
Rapporteur Externe :	M. Marc ROBINSON-RECHAVI, professeur, UNIL
Rapporteur Externe :	M. Matthieu GERARD, directeur de recherche, CEA
Membre invité :	M. Olivier POCH, directeur de recherche, CNRS

Remerciements

Un grand merci général à tout ceux qui ont animé ces trois ans, j'en oublierai certainement dans la liste qui suit. Qu'ils soient remercié dans ces deux premières lignes.

Merci à …

- Mes encadrants (voir ci-dessous) avec qui nous avons formé une intéressante tétrade élémentaire d'eau (Odile), de terre (Philippe), de feu (Olivier) et d'air (votre serviteur). Une thèse est un bien étrange et long voyage qu'on sait d'avance être difficile et riche d'humain, du pire au meilleur et pourtant le bilan semble toujours le même. Des personnes épuisées mais satisfaites. C'est mon cas, merci pour ça.

  - Odile Lecompte : L'eau (dans tous ses états) du quatuor, que je remercie d'avoir su éteindre les brasiers pochiens et canalisé son énergie démesurée par ses kilomètres de digues impassibles. Merci pour son énervante et justifiée aptitude à corriger mes idées trop vite lancées et pour avoir enduré le rôle ingrat du recadrage. Merci aussi d'avoir su être là quand le besoin s'est fait sentir.

  - Philippe Carbon: Valeur sûre, îlot de terre esplanadienne où se poser dans la tempête. Merci avant tout pour la thématique passionnante, pour ses touches anonymes dans mon travail, pour nos discussions caféinées et détendues allant de la dynamique moléculaire à l'avenir de la recherche. Merci d'avoir compris mon pari osé sur l'avenir quand il aurait été plus simple et scientifiquement exaltant de poursuivre. J'espère ne jamais regretter.

  - Olivier Poch: Une thèse de remerciements n'y suffirait pas mais citons pêle-mêle: Pour sa flamme infinie et son incroyable talent visionnaire en sciences, souvent incompatible avec le sommeil de ses doctorants. Pour sa patience vis-à-vis dudit sommeil. Pour sa déroutante faculté à poursuivre toute conversation avec chaque membre du laboratoire sans plus de recadrage. Pour ses idiomes et son humour communément admis comme "spécial". Pour avoir passivement enrichi ma rhétorique par la redoutable sienne.. Pour avoir fait de gros efforts pour se recentrer et nous faire moins quitter les rails de nos thèses.

- Alain Krol: pour son abnégation et sa mansuétude vis à vis des membres de son équipe et pour m'avoir fait confiance en me laissant représenter notre équipe lors d'échanges internationaux.

- Dino Moras: Pour sa gestion pleine de sagesse et d'humilité d'un des meilleurs instituts de recherche au monde, qui sera à n'en pas douter regrettée. Pour m'avoir permis de lier mon master et ma thèse par une intéressante thématique qui s'avèrera fort utile par la suite.

- Eric Westhof: Pour voir, comme Dino Moras, permis de réaliser mon stage de master dans de décentes conditions. Pour tout ce que j'ai appris à le voir opérer à la vice-présidence Recherche de l'université lors des conseils et pour son aptitude déroutante à tant connaitre els ARN et pourtant toujours confondre les prénoms.

- Catherine Florentz: A tant de titres. Pour sa gestion exemplaire de l'Ecole Doctorale et son impartialité lors des conseils, réduisant les conflits passés entre mes mains de représentant à la portion congrue. Pour m'avoir permis d'être un consultant doctorale, inestimable expérience pour l'avenir, et plus globalement pour toutes les compétences qu'elle a su nous transmettre via les experts dont elle a su s'entourer pour parfaire nos formations. Enfin, pour m'avoir permis d'aller au Japon pour une ineffable expérience scientifique et humaine.

- Fanny Hummel: Pour sa gentillesse, sa dévotion à sa tache de secrétaire de l'Ecole doctorale et sa réactivité précieuse dans ce monde procédurier.

- Natacha Rochel : Pour notre collaboration sur le promoteur de RARb qui m'a aspiré dans le monde fascinant des récepteurs nucléaires

- Cecile Rochette-Egly : Pour notre collaboration sur les cibles de RAR et nous avoir laissé une précieuse carte-blanche.

- Ronald Carpio-Farro: Pour notre collaboration sur CNGA3 et pour m'avoir appris à dialoguer avec un biologiste

- Sebastien Lalevée : Pour nos discussions d'égal à égal lors de notre collaboration et pour ne pas m'avoir remercié dans sa thèse.

- Nicolas Gagnière : Pour ses précieux conseils, proches du matraquage, en matière de programmation et pour avoir activement contribué à mon embonpoint par fast-food interposés.

- Jean-Radwen Aniba : Pour avoir suscité en moi le goût de l'économie et du business et donc réorienté toute ma carrière. Un grand merci aussi pour sa mesure tempérée, son endurance face à notre humour des plus lourds et pour avoir été un si agréable compagnon d'infortune.

- David Kieffer : Pour son enthousiasme démesuré et ses conseils en propagation d'alignements pairwise.

- Laëtitia Poy'de Vine : Pour son moral systématiquement au beau fixe et son aptitude a endurer ma volubilité lors de pauses nicotiniques salutaires

- Raymond Ripp: pour sa systématique bonne humeur, son immense savoir informatique (cet homme a vu les dinosaures !) et sa contribution aux visualisations statistiques de GPS.

- Julie Thompson: Pour son humour "so british" et ses très précieuses révisions de mes écrits anglais.

- Wolfgang Raffelsberger : Pour ses conseils aiguisés en matière de statistique et pour avoir été systématiquement le "last man standing" du laboratoire, le soir, nous enjoignant à faire de même.

- Nicolas Wicker : Pour ses conseils en clustering, modèles de mélange et sectes étranges.

- N'goc Hoan N'Guyen : Pour son expertise en bases de données et son incroyable système. Merci aussi pour avoir permis de développer mes facultés auditives basées sur l'intuition.

- Viet Dao Luu : Pour avoir complété le peu de compétences qu'il me manquait en "franç'namien".

- Evelyne Harlé: Pour sa contribution au moteur phylogénétique de GPS et m'avoir enduré en maitre de stage.

- Laurent-Philippe Albou : Pour avoir inventé le thésard quantique, rendant ma fréquentation assidue du laboratoire quasi stakhanoviste.

- Laura Cammas, Dominique Kobi, Gioia Altobelli et Kinga Bujakovska : Pour avoir sollicité mon expertise en me montrant leurs problèmes respectifs.

- Patrick Llerena, Alice Coués, Clémence Dro, Sylvestre Gug et Stéphane -docteur -Heitz: pour avoir fait confiance à mes analyses de consultant.

- Laëtitia Gonzalez-Lerch : pour sa bonne humeur et son efficacité précieuse palliant mon organisation stratosphérique.

- Evelyne Acker : Pour avoir fait commencer ou se finir chaque journée par un franc salut.

- Nicolas Sarkozy et Valérie Pécresse : Pour m'avoir montré la voie de l'après-thèse.

Et last but not least :

- Joseph André, Audrey Argili, Sophie Schweitzer et Lydia Weingand, plus que des amis, mes piliers. Pour avoir été systématiquement là, chacun dans ce pan qui vous est propre et pour avoir aimablement fait semblant de comprendre en quoi consistait ma thèse.
- Alisson Boche qui a enduré sans broncher mon rythme de travail et mon quasi-emménagement au labo des dernières semaines. Merci de m'avoir forcé sans rien me demander à recoller avec le monde réel en franchissant la porte. Merci pour tes relectures de mes travaux et tes livraisons de paniers-repas les (trop nombreuses) nuits où je ne suis pas rentré. M4HoneyB.

"Je ne crois pas en l'évolution, Darwin et toutes ces choses, non. Je pense plutôt que nous venons de l'Atlantide."

Maryvonne Anno, ma mère.

Table des abréviations

- ADN : acide désoxyribonulcéique
- AR *:* acide rétinoïque
- API : application programming interface
- ARN : acide ribonucléique
- ARNl : ARN long
- ARNmi : ARN microscopique interférence
- ARNnc : ARN non-codant
- ARNr : ARN ribosomal
- ARNs : ARN court
- ARNsc : petit ARN cytoplasmique
- ARNsn : petit ARN nucléaire
- ARNsno : petit ARN nucléolaire
- ARNs-TSS : petit ARN attaché au site d'initiation de la transcription
- ARNt : ARN de transfert
- ARNtSec : ARNt sélénocystéine
- BIRD : *biological integration and retrieval data*
- CAGE : *cap analysis of gene expression*
- CDS : séquence codante
- ChIP : immunoprécipitation de la chromatine
- ChIP-on-chip : ChIP sur une puce à ADN
- CLON : *count-length-occupancy-nearest*
- CMS : *content manager system*
- CUT : transcrit cryptique inconnu
- DAVID : *database for annotation, visualization and integrated discovery*
- DBD : domaine de liaison a l'ADN
- DSE : d*istal sequence element*
- EncODE : encyclopedia of DNA elements
- FAIRE : f*ormaldehyde assisted isolation of regulatory elements*
- GeCo : gene Context

- HMM : modèle de Markov Caché
- IC : intervalle de confiance
- IUPAC :*international union of pure and applied chemistry*
- Med : médiane
- MIT : Massachussetts institute of technology
- MPSS : massively parallel sequencing signature
- NDR : régionn sans nucléosome
- PALR : ARN long associé au promoteur
- PASR : ARN court associé au promoteur
- PDO : *application programming interface*
- PET : *paired-end di-tag*
- PolI : polymérase I
- PolII : polymérase II
- PolIII : polymérase III
- PROMPT : transcrit en amont du promoteur
- PSE : *proximal sequence element*
- PSSM : matrice de score position-spécifique
- Q1 : premier quartile / vingt-cinquième percentile
- Q3 : troisième quartile / soixante-quinzième percentile
- QA : cinquième percentile
- QZ : quatre-vingt-quinzième percentile
- RACE : *rapid amplification of cDNA ends*
- RAR : Récepteur à l'AR
- RAR-RE : élément de réponse à RAR
- RBAM : radar de boites-à-moustaches
- SELEX : *systematic svolution ligand by exponential enrichment*
- SGBD : Système de gestion de bases de données
- SNP : polymorphisme mononucléotidique
- Staf : *selenocysteine tRNA gene activation factor*
- SUT : transcrit stable inconnu
- TASR : petit ARN associé au terminateur
- TSS : site d'initiation de la transcription

- TUF : transcrit à fonction inconnue
- UCSC : université de Santa-Cruz Californie
- (5'/3')-UTR : région non-traduite (en 5'/3')
- ZNF : doigt-à-zinc

Table des matières

Index des figures

Index des tableaux

Avant-propos

Que ce soit la bactérie dans son écosystème ou la planète dans son système solaire, tout système complexe ne peut-être appréhendé que par le contexte au sein duquel il s'insert, avec lequel il interagit et finalement qu'il modifie intrinsèquement. A cette règle, les éléments liés au génome ne font pas exception. Ainsi face au tsunami de nouvelles données, informations et connaissances submergeant la biologie moderne, il devient essentiel de pouvoir replacer, en temps réel, toutes ces nouveautés dans leurs contextes informationnels. Dans le cadre d'analyses impliquant les informations liées aux génomes, de tels contextes peuvent concerner aussi bien la taille d'un exon ou la densité en éléments répétés que le nombre de nucléosomes présents dans une région ou les distances séparant diverses entités génomiques. A notre sens, les outils bioinformatiques associés aux données biologiques modernes ne peuvent plus se contenter de fournir un traitement de données et les résultats en découlant, mais se doivent d'informer sur leur normalité (ou anormalité). Par delà le fait d'offrir une dimension de « contexte informationnel » à l'utilisateur, cette approche doit être à même de lui indiquer ce qui sort de l'ordinaire et peut ainsi l'aider à identifier ce qui doit être analysé en priorité. C'est à cette fin que nous avons développé la base de données GeCo et son portail web associé. Ce système intègre non seulement des annotations issues des grands centres mondiaux qui incluent des données de génomique, d'épigénétique et d'évolution, mais aussi génère sa propre connaissance. Pour ce faire, GeCo calcule nombre de valeurs statistiques qui constitueront la gamme d'étalonnage où seront replacées les données des utilisateurs, via des représentations innovantes.

Tout élément étant fondamentalement contexte d'un autre élément, les possibilités qu'offrent l'étude par le contexte sont infinies. Aussi, dans notre architecture, nous avons choisi de nous focaliser uniquement sur les éléments contextuels de « premier niveau », constituant le contexte informationnel des gènes. Nous présenterons dans l'introduction les éléments majeurs de ce premier niveau et verrons en quoi, au cœur de l'ère du séquençage à haut-débit, il remet en question nombre de dogmes liés à la transcription. Différentes analyses préliminaires réalisées sur des systèmes impliquant quelques centaines de gènes ou régions chromosomiques ont permis de montrer l'intérêt d'une analyse contextuelle à l'ère post-génomique.

Cependant c’est dans le cadre de l’étude des éléments régulateurs de gènes que nous avons pu exploiter pleinement notre architecture.

En tant qu'entités fortement dépendantes du contexte et pour lesquelles les données massives commencent à être exploitables, les éléments régulateurs des gènes constituent un exemple de choix. C’est pourquoi nos développements se sont inscrits dans le cadre d’une collaboration, avec le laboratoire du Pr Carbon de l’IBMC de Strasbourg où a été découvert un facteur de transcription très particulier : le facteur Staf. Ce facteur, qui sera présenté en détail au début de notre introduction, permet d'activer plusieurs types de gènes (protéiques et correspondants à divers ARN non-codants) et fait intervenir au moins deux machineries de transcription. Au regard des études précédentes, il semblait posséder un grand nombre de sites de fixation au sein du génome humain et représentait donc un cas idéal pour implémenter et tester notre architecture dans un contexte de haut-débit.

Dans cette analyse, nous avons pu nous appuyer sur les importantes connaissances et données biologiques qui ont été produites par l’équipe du Pr Carbon et confronter nos modèles et concepts avec les données expérimentales. Nous verrons plus précisément au sein des résultats combien la synergie biologie-informatique a été importante et notamment, comment les prédictions bioinformatiques ont permis, en première approximation, de faire émerger des pistes d'exploration, de même que la connaissance du contexte informationnel s’est avérée indispensable pour améliorer des modèles mathématiques se perdant dans le flot de la variabilité biologique. Par ces allers-retours permanents entre les deux univers, nous avons ainsi pu passer outre certaines limites et souligner le rôle majeur du facteur Staf dans la cellule humaine. L'ensemble de ces travaux a donné lieu à quatre articles. Deux d'entre eux ont été soumis et seront joints à ce manuscrit de thèse. Les deux autres sont actuellement en préparation.

INTRODUCTION

Chapitre 1 : le facteur Staf

La transcription est le phénomène universel par lequel un vecteur ARN est produit à partir d'une matrice ADN au sein de complexes protéiques. Ces complexes protéiques, les machineries de transcription, sont constitués de nombreux cofacteurs agrégés autour d'une ARN polymérase ADN dépendante. Cette dernière est chargée d'exécuter la synthèse de la molécule d'ARN en accord avec la complémentarité des bases nucléiques du brin transcrit de l'ADN. Chez les eucaryotes, le type de polymérase impliqué (Pol I, II ou III) est associé à 3 machineries de transcription distinctes ayant chacune son répertoire de gènes à transcrire (Figure 1):

1) L'ARN polymérase I (Pol I) transcrit la quasi-totalité des gènes d'ARN ribosomaux (ARNr)

2) L'ARN polymérase II (Pol II) est en charge de la transcription des gènes de protéines, des précurseurs de micro-ARN et d'une partie des gènes de petits ARN nucléaires (ARNsn) et de la totalité des petits ARN nucléolaires (ARNsno)

3) l'ARN polymerase III (PolIII) est en charge des autres gènes d'ARNsn et d'autres ARN non-codants (ARNnc) comme les ARN de transfert (ARNt) et l'ARNr.

Ces machineries de transcription sont recrutées au niveau de l'ADN via des facteurs de transcription reconnaissant spécifiquement une courte séquence nucléique et permettant ainsi la constitution du complexe d'initiation de la transcription à proximité du site d'initiation de la transcription (TSS). Parmi ces facteurs, le facteur Staf (Selenocysteine tRNA Transcription Activating Factor) constitue un facteur atypique, à même de recruter les machineries de transcription Pol II et III. Il s'avère essentiel à la survie cellulaire par son rôle indispensable dans l'expression du gène de l'ARNt sélénocystéine ($ARNt^{Sec}$) qui lui a donné son nom. Nous allons donc passer en revue les caractéristiques connues de la protéine Staf et de ses sites de fixation dans l'ADN afin de mettre en lumière son rôle particulier dans la transcription.

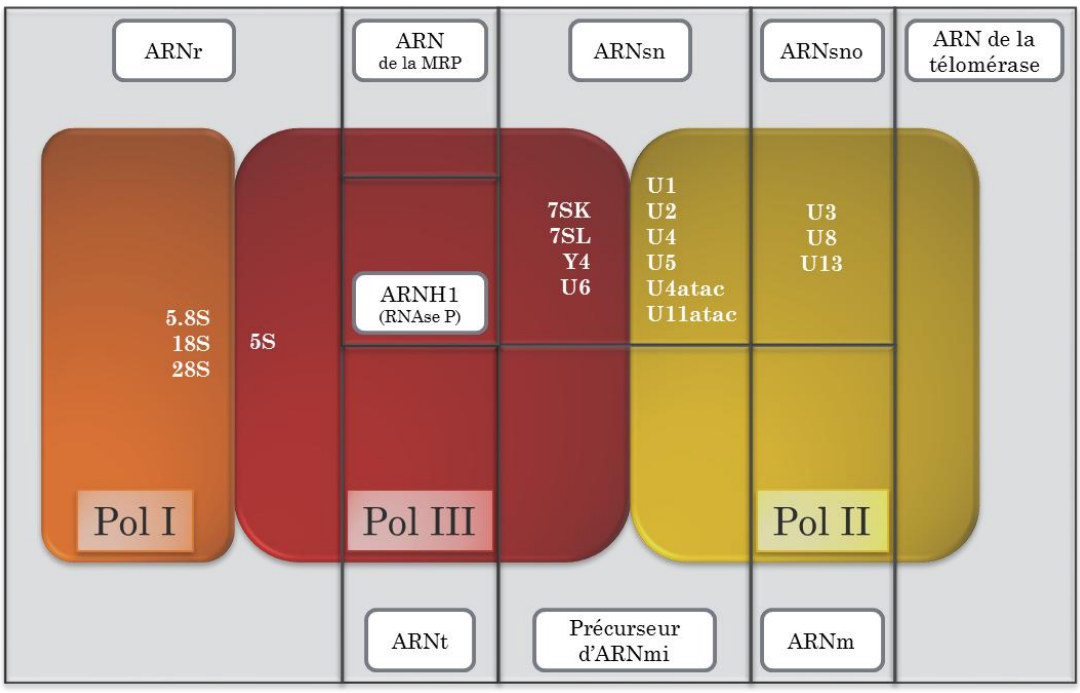


Figure 1 : **Représentation des types de gènes transcrits par les diverses ARN polymerases eucaryotes**

Certains types de gènes sont spécifiques d'une polymérase comme les ARNsno de la PolII alors que beaucoup sont répartis entre deux polymérases comme les ARNr activés selon l'ARN, par PolI ou PolIII.

1. Caractéristiques structurales du facteur Staf

Le facteur Staf est une protéine constituée de 507 acides aminés (chez *Xenopus laevis*) qui renferme trois domaines majeurs. De son extrémité amino-terminale à son extrémité carboxy-terminale, on trouve successivement un domaine d'activation (AD), un domaine de liaison à l'ADN (DBD) et un troisième domaine à la fonction encore inconnue.

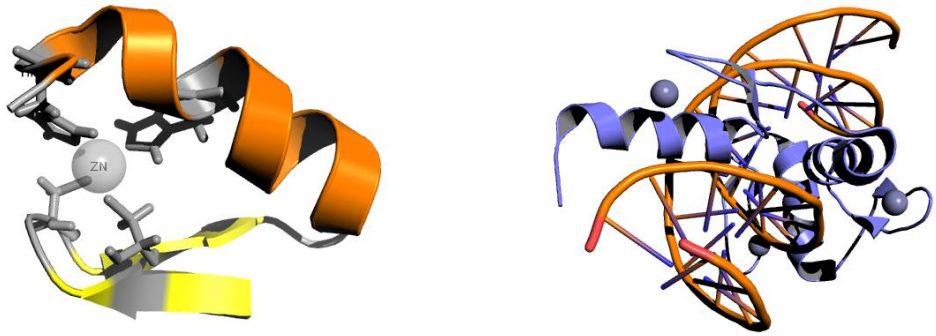
1.1. Le domaine d'activation de Staf

L'AD de Staf (268 acides aminés) présente deux sous-domaines d'activation fonctionnellement distincts. La transcription des gènes d'ARNm, liée à la PolII est activée via une première région s'étendant de la glutamine[84] à l'histidine[176] et renfermant un même motif actif répété quatre fois (Schuster *et al.*, 1998). Une deuxième région d'activation s'étendant de la proline[207] à la glycine[224] est impliquée dans la transcription des gènes d'ARNnc par la Pol II ou la Pol III. Ces deux sous domaines ont été finement caractérisés et la leucine[213] de la deuxième région s'avère indispensable à cet élément unique en charge de la transcription des gènes d'ARNnc. Un lien a été établi entre le nombre de répétitions du motif dans la première région et l'efficacité de la

transcription. Une à trois copies de ce motif répété peuvent être éliminées sans que l'activation transcriptionelle ne soit abolie. Enfin, Schuster et collaborateurs (Schuster *et al.*, 1998) ont également démontré que la leucine[166], l'aspartate[168] et la thréonine[170] du quatrième élément répété jouent un rôle majeur sur son aptitude à activer la transcription.

1.2. Le domaine de liaison à l'ADN de Staf

Ce domaine couvre 40% de la protéine et s'étend des résidus 268 à 468. Il renferme 7 doigts à zinc de type C2-H2 (Miller *et al.*, 1985) mais seuls les six premiers établissent des contacts avec l'ADN (Schaub *et al.*, 1999a) sur un site de liaison appelé Staf Binding Site (SBS). Les doigts à zinc de type C2-H2 sont largement répandus dans les protéines qui se lient à l'ADN ou il permettent l'établissement d'une reconnaissance spécifique entre une protéine et une séquence d'ADN. Classiquement, l'atome de zinc établit quatre liaisons de coordination avec les deux cystéines et les deux histidines invariantes. Un motif de liaison de 6 acides aminés est présent entre chaque doigt et sa séquence est très conservée dans l'ensemble des protéines qui renferment des doigts à zinc de ce type. L'interaction avec l'ADN des doigts à zinc de type C2-H2 se fait par pénétration du doigt dans le grand sillon. Ce faisant, des liaisons spécifiques avec trois paires de bases de l'ADN s'établissent par l'intermédiaire de quatre résidus d'une hélice α précédant une structure de type β–α–β, (Figure 2A). L'interaction de plusieurs doigts se fait par enroulement autour de l'ADN et cette interaction est réalisée de manière anti-parallèle, de sorte que le premier doigt à zinc reconnaît les 3 bases en 3' du site de fixation du facteur, le dernier doigt à zinc reconnaissant quant à lui les bases les plus en 5' du site (Figure 2B). Les six premiers doigts à zinc de Staf reconnaissant chacun 3 pb un SBS canonique est constitué de 18 pb (Schaub *et al.*, 2000). Le septième doigt à zinc de Staf n'établit pas de contact avec l'ADN mais stabilise le complexe en augmentant l'affinité du facteur pour sa séquence cible (Schaub *et al.*, 1999a). De plus, une participation différentielle à la reconnaissance de l'ADN a été mise en évidence pour les différents doigts à zinc. Elle dépend du type de séquence cible et selon les cas, la fixation du premier doigt à zinc peut s'avérer indispensable ou pénalisante pour la transcription du gène (Schaub *et al.*, 1999b). Enfin, les doigts à zinc 3 à 6 semblent être les seuls réellement indispensables à l'établissement d'une association fonctionnelle, du facteur sur son site (Schaub *et al.*, 2000).



A B

Figure 2: Représentation tridimensionnelle et interaction avec l'ADN d'un doigt à zinc de type C2H2

A/ Les cytosines et histidines impliquées dans la liaison à l'atome de zinc sont figurées en bâtonnets gris, l'hélice α en orange et les deux feuillets β en jaune. (PDB/1ZNF). B. Interaction de trois doigts à zinc de Zif268 (bleu) avec la double hélice d'ADN (orange)(PDB 1G2F).

1.3. Domaine carboxy-terminal de Staf

Ce domaine n'a pour l'heure pas de fonction connue. Une analyse au sein de la banque de données Pfam ne met en évidence ni domaine bien documenté, ni simplement identifié.

2. Homologues de Staf au sein du vivant

Le facteur Staf fut initialement découvert chez *Xenopus laevis* (Schuster *et al.*, 1995) mais une interrogation des banques de données protéiques avec la séquence batracienne permit rapidement d'identifier deux homologues humains, alors à fonction inconnue (Myslinski *et al.*, 1998). L'orthologue humain de Staf (hStaf) est ZNF143 avec qui le facteur Staf présente 84% de similarité. La seconde protéine, ZNF76, est vraisemblablement son paralogue avec qui Staf ne possède que 64% de similarité. Toutefois, ces similarités augmentent à 95% entre Staf et ZNF143 et 86% entre Staf et ZNF76 si l'on examine le seul DBD. Staf et ses orthologues semblaient jusque récemment être uniquement présents chez les vertébrés. Mais la haute conservation de séquence du DBD a permis aux programmes de recherche de similarité de mieux ancrer les alignements des séquences protéiques et de mettre en évidence des orthologues potentiels dans d'autres organismes dont l'anémone de mer et l'abeille (Putnam *et al.*, 2007). Une étude est en cours au laboratoire pour vérifier si ces protéines représentent réellement des orthologues de Staf car, en dehors du DBD, les séquences sont très divergentes.

3. Promoteurs reconnus par le facteur Staf

3.1. Promoteur du gène de l'ARNt^{Sec}

historiquement, le gène d' ARNt^{Sec} fut le premier identifié comme régulé par Staf. Au sein des gènes d'ARNt transcrits par la Pol III, le gène de l'ARNt^{Sec} est un gène particulier. En effet, des deux boites A et B typiques d'un ARNt situées dans la partie transcrite, l'ARNt^{Sec} ne conserve que la seconde. De plus, le promoteur renferme trois éléments, situés en amont de la partie transcrite, et caractéristiques des promoteurs de gènes d'ARNsn transcrits par la Pol III comme le gène de l'ARNsn U6. Deux de ces éléments sont la boîte TATA et la boîte PSE (Proximal Sequence Element) (Carbon *et al.*, 1987; Murphy *et al.*, 1987). Le troisième élément, activateur de la transcription, est constitué par la DSE (Distal Sequence Element) qui renferme souvent deux motifs, le motif octamer et le motif SBS, sites de liaison respectifs des protéines Oct1 et Staf. La DSE, la PSE et la boîte TATA de l'ARNt^{Sec} sont de plus fonctionnellemnt équivalents à ceux présents aux mêmes positions dans le promoteur du gène de l'ARNsn U6 transcrit par la Pol III (Carbon and Krol, 1991; Myslinski *et al.*, 1993) (Figure 3).

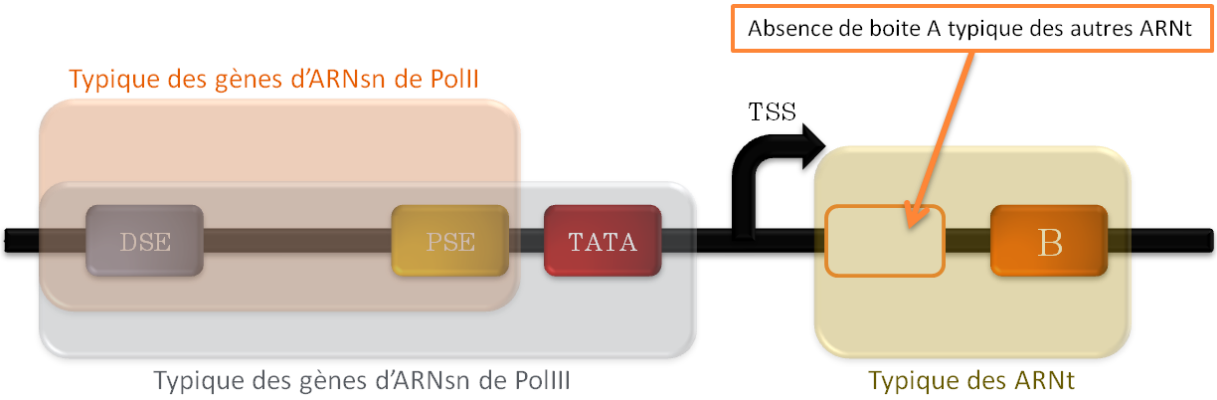


Figure 3 : Agencement du promoteur de l'ARNtSec

Le promoteur de l'ARNtSec présente i) en amont du site d'initiation de la transcription, des éléments communs avec les gènes d'ARNsn : la PSE recélant le site Staf, la DSE et une boite TATA typique des ARNsn transcrits par PolII (et de certains gènes de protéines) ; ii) en aval du site d'initiation, une seule des deux boites typiques des ARNt: la boite B..

3.2. Promoteurs des gènes d'ARNsn et d'autres ARNnc transcrits par la Pol II et la Pol III

Il a également été démontré que Staf activait la transcription des gènes des ARNsn U1, U2, U4 et U5 transcrit par la Pol II ainsi que celle des gènes de l'ARN de la MRP, de l'ARN 7SK , de

l'ARN Y4 et de l'ARN de la RNase P transcrit par la Pol III. La protéine humaine, ZNF143, orthologue de Staf de Xénope possède les mêmes propriétés d'activation pour les gènes transcrit soit par la Pol II soit par la Pol III (Schaub *et al.*, 1997)(Myslinski et la 1998). Tous ces éléments reflètent la complexité du facteur Staf et de son rôle dans l'activation de la transcription.

3.3. Promoteurs de gènes de protéines

C'est en plaçant le gène rapporteur de la chloramphénicol-acétyl-transférase en aval d'un promoteur chimère artificiel constitué du promoteur de base du gène de la thymidine kinase fusionné avec un SBS qu'il fut démontré que Staf et ses homologues humains portaient également la capacité d'activer la transcription de gènes de protéines (Schuster *et al.*, 1995). Par la suite, sept gènes de protéines régulés par Staf ont été identifiés. Chez la souris, mStaf active la transription du gène codant pour une sous-unité du complexe de chaperons cytosolique (Kubota *et al.*, 2000) et du gène codant pour l'aldéhyde réductase (AKR1A1) (Saur *et al.*, 2002). L'orthologue humain de ce dernier gène est lui aussi activé par ZNF143 (Barski *et al.*, 2004), de même que les gènes codant pour l'oxyde nitrique synthase neuronale (NOS1) (Saur *et al.*, 2002), le facteur de régulation de l'interféron 3 (IFR3) (Mach *et al.*, 2002), la transaldolase (TALDO1) (Grossman *et al.*, 2004), la synaptobrévine-like 1 (SYBL1) (Di Leva *et al.*, 2004) et la protéine ribosomique mitochondriale S11 (MRPS11) (Ishiguchi *et al.*, 2004). L'expression de cette dernière protéine, comme celle de ZNF143, augmente dans les cellules cancéreuses traitées par le cis-platine. Dans des cellules devenues résistantes au cis-platine, l'élimination de ZNF143 par des techniques d'ARN interférence se traduit par un rétablissement de la sensibilité des cellules au cis-platine (Ishiguchi *et al.*, 2004) (Wakasugi *et al.*, 2007).

L'identification de ces sept gènes de protéines laissait supposer que la transcription d'un nombre plus important de gènes de protéines pouvait être dépendant de Staf. Récemment, une étude fut entreprise à l'échelle du génome humain afin de trouver de nouveaux sites de fixation dans les promoteurs de gènes de protéines. En interrogeant la banque de promoteur DBTSS v4 (Yamashita *et al.*, 2006) avec une séquence consensus de Staf et dix variants de cette séquence, Myslinski et collaborateurs en 2006 identifièrent 938 promoteurs humains de gènes protéiques renfermant de tels sites. L'occupation in vivo par hStaf de 295 de ces promoteurs a été testée en recherchant par PCR, l'ADN de ces promoteurs, dans l'ADN isolé par la technique

d'immunoprécipitation de chromatine (ChIP) à l'aide d'un anticorps antiStaf. Il s'avère que Staf est bien associé à ces promoteurs dans 90% des cas (Myslinski *et al.*, 2006) et que les gènes correspondants présentent un enrichissement fonctionnel en protéines se liant à l'ADN (23%) ou intervenant dans la synthèse et le cycle de vie des protéines (21%)(Figure 4). Cette étude sera détaillée sous l'angle des sites de fixation au paragraphe 4.2.

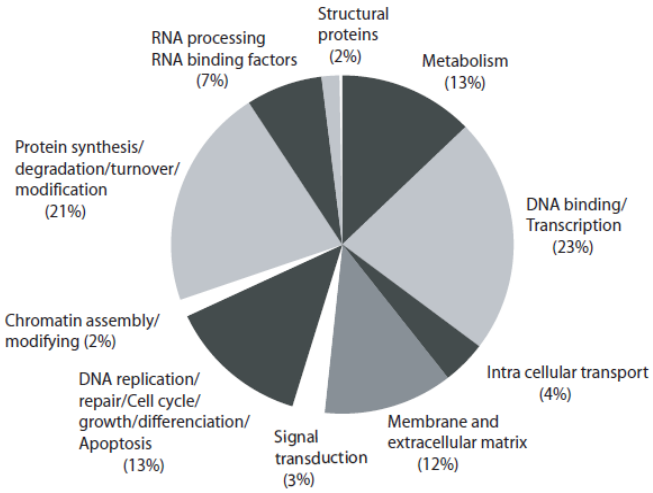


Figure 4 : Répartition des ontologies GO associées aux gènes potentiellement régulés par Staf

On note une prévalence des gènes liés à la transcription et à la vie des protéines (Myslinski et al., 2006)

Plus récemment, la dissection des promoteurs du gène TFAM, codant pour une protéine impliquée dans la réplication et la transcription de l'ADN mitochondrial et du gène BUB1B dont le produit est impliqué dans le contrôle du cycle cellulaire, a démontré que ces promoteurs renfermaient des éléments de réponse à Staf fonctionnels (Gerard *et al.*, 2007) (Myslinski *et al.*, 2007). Enfin, tout dernièrement, il a été démontré que ZNF143 est impliqué aussi dans la production de l'ARNnc SCARNA2 (Gerard *et al.*, 2010), ARN qui participe au sein d'une particule ribonucléoprotéique à l'introduction d'une base modifiée dans l'ARNsn U2 (Tycowski *et al.*, 2004). Ces différents sites sont présentés à la Figure 5

4. Interaction de Staf avec l'ADN

4.1. Séquence consensus

De par la reconnaissance de 3 pb d'ADN successives par chacun des 6 doigts de zinc fonctionnels du DBD de Staf il en découle que le site de fixation canonique de ce facteur est constitué de 18 pb. Ce site est capable à lui seul de recruter la machinerie de transcription Pol II, même en l'absence de tout autre élément classique du promoteur de base des gènes de protéine et d'initier la transcription (Myslinski *et al.*, 2006). Les séquences des sites de fixation identifiés dans les promoteurs de gènes d'ARNnc et dans les promoteurs des sept premiers gènes de protéines identifiés comme régulés par Staf sont données dans la Figure 5.

Figure 5 : Séquence et répartition des séquences des premiers sites de fixation de Staf identifiés

Ces sites sont ordonnés selon le type de gène impliqué et la machinerie de transcription recrutée, (1), (2) réfère à l'un des 2 sites présents dans le promoteur des gènes concernés .

La comparaison des séquences révèle l'extrême flexibilité de ce site de fixation dont 13 positions sont relativement dégénérées et dont seules les cytosines en position 3, 4, 5 et 12,

ainsi que l'adénine en position 6 semblent faire l'objet d'une relative forte pression de conservation. Des expériences de sélection de sites (Schaub *et al.*, 1999a; Schuster *et al.*, 1995) in vitro par la technique de (*Systematic Evolution Ligand by EXponential enrichment*) ont permis de dériver la séquence consensus de reconnaissance de Staf de 18 :
Y(A/t)CCC(A/g)N(A/C)AT(G/c)C(A/c)YYRCR ou YWCCCRNMATSCMYYRCR

4.2. Séquence couplée au SBS

L'étude réalisée par Myslinski et collaborateurs en 2006 permit de récupérer 1488 nouveaux sites potentiels de liaison de ZNF143 dont 648 (43.5%) étaient conservés chez la souris. Parmi eux, près de la moitié présentait une extension très conservée de 7 pb de séquence ACTACAN située immédiatement en amont du SBS. Cette séquence, couplée aux 12 premières bases du SBS permit alors d'extraire 204 nouveaux SBS conservés chez la souris complétés par 41 SBS supplémentaires suite à examen manuel. Parallèlement, une autre étude systématique visant à identifier les motifs de 12 bases maximum surreprésentés et conservés dans les séquences des promoteurs de mammifères - 4000 paires de bases centrées autour du TSS - , a mis en évidence un certain nombre de motifs répondant à ces critères (Xie *et al.*, 2005). Le quatrième motif le plus représenté (motif M4) de séquence ACTACANNTCCC fut classé comme la séquence orpheline (ne constituant pas la cible connue d'un facteur de transcription) la plus présente dans ces promoteurs. Dans le même temps, une autre étude vint confirmer ces résultats et souligna quant à elle la surreprésentation d'un motif de type ACTACAANTCCCA dans les promoteurs (Prakash and Tompa, 2005). Ces séquences contiennent d'une part le motif ACTACAN précédemment identifié en amont du SBS et d'autre part les 6 premières pb de ce site. Les séquences de type M4 de Xie (Xie *et al.*, 2005) ont été rallongées de 12 paires de bases en 3' par Myslinski et collaborateurs. Il s'est avéré que parmi les séquences obtenues 282 étaient en fait des motifs ACTACAN avec un SBS associé. Ainsi, grandement aidée par cette séquence ACTACAN, une collection de 1175 SBS potentiels fut mise à jour, permettant de proposer un logo affiné pour le site de fixation de Staf, site qui apparaît comme extrêmement dévoyé (Figure 6).

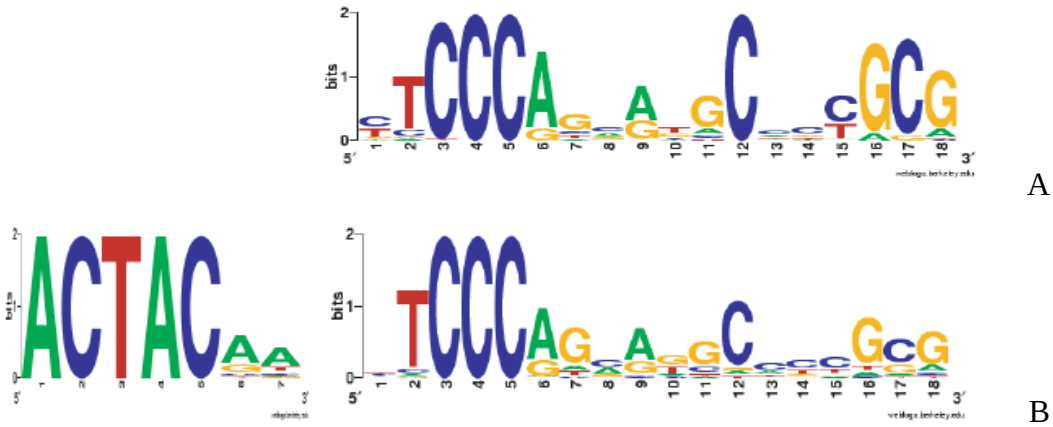


Figure 6 : Visualisation graphique de la répartition des quatre bases de l'ADN à chaque position du SBS moyen d'après Mylsinki *et al.*, 2006

Cette répartition fut obtenue suite à la recherche utilisant le consensus du SBS (A) et du SBS moyen étendu de 7 bases en 5' obtenu en partant du motif M4 décrit par Xie et al. en 2005 (B).

Plus récemment ces mêmes auteurs, en améliorant leur méthode de détection, sont parvenus à mettre en évidence des séquences associant le motif ACTACAN et les 12 premiers nucléotides du SBS, sans pour autant faire le lien avec le site de liaison de Staf et proposèrent un motif plus dévoyé de l'ACTACAN : l'ACTAYRN (Xie *et al.*, 2007).

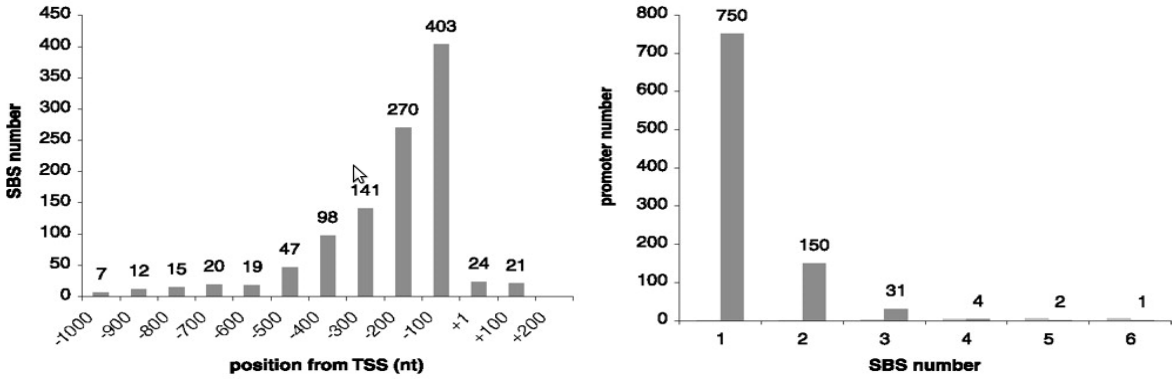
Cette séquence ACTACAN/ACTAYR apparaissant dans près de la moitié des cas à côté d'un SBS semble donc spécifique d'un sous-ensemble de ces derniers et plus fortement conservée que ceux-ci. Ce motif n'est peut-être pas le seul à être lié au SBS mais semble le plus fréquent au vu de l'étude systématique de Xie. Cette étude ne permettant pas de discerner le SBS entier lié à l'ACTACAN du bruit environnant, il se peut donc aisément que d'autres signaux liés aux SBS soient de fait trop faibles pour être détectés de manière systématique.

On ne peut pour l'heure qu'imaginer le rôle de ce motif ACTACAN, pouvant soit favoriser la fixation du facteur Staf, soit être reconnu par un autre facteur. Cette dernière option laisse entrevoir un possible jeu combinatoire de facteurs, voire une compétition pour la partie 5' de l'ACTACAN-SBS.

4.3. Caractéristiques propres aux SBS de promoteurs

L'étude de Myslinski et coll (2006) a mis en évidence que la majorité des promoteurs à SBS (80%) contient un seul site et qu'une proportion importante des SBS (15%) est en fait impliquée dans un promoteur partagé par deux gènes en orientation inverse, laissant supposer un rôle de hStaf dans l'expression des gènes à partir de promoteurs bidirectionnels. Enfin, une étude de la distribution des 1175 sites par rapport au TSS des gènes impliqués révèle une

apparente préférence pour la zone de 0 à 200 pb en amont du TSS où sont localisés 57% des sites (Figure 7).



A B

Figure 7 : Nombre et distribution des SBS attachés aux gènes

A/ Distribution des distances observées entre les sites de fixation de hStaf (SBS) et le TSS du gène le plus proche. B/ Répartition du nombre de promoteurs en fonction du nombre de SBS dans ces promoteurs (figure extraite de Myslinski et al., 2006).

4.4. Validation expérimentale et enjeux soulevés par l'abondance des SBS.

Un nombre significatif des 938 promoteurs identifiés *in silico* comme renfermant des sites SBS fut recherché par PCR dans l'ADN obtenu par immunoprécipitation de chromatine (ChIP) avec des anticorps antiStaf. Les résultats ont montré que plus de 90 % des sites identifiés *in silico* sont *in vivo* associés à hStaf. Ce nombre très important de sites validés expérimentalement, conforté par l'étude *in silico* ciblée sur une banque de promoteurs laisse à penser qu'un nombre conséquent de site demeurent encore inconnus à l'échelle du génome humain. Parmi eux, une partie sera certainement attachée à la séquence ACTACAN dont le rôle reste à élucider. Aussi, pour poursuivre l'analyse des sites facteur hStaf, il est désormais indispensable de disposer d'une architecture à même d'absorber un nombre important de données récentes issues des '`omics et de pouvoir replacer ces sites dans leur contexte informationnel.

Chapitre 2 : Le contexte informationnel des gènes

L'ère post-génomique et le tsunami de données qui l'accompagne a profondément modifié notre conception du génome et du gène (Gerstein *et al.*, 2007) en mettant en relief l'importance du contexte informationnel pour étudier et comprendre le rôle de toute région génomique. Celui-ci est multi-dimentionnel et hiérarchisé du nucléotide à l'organisme. Le premier niveau de ce contexte informationnel peut toutefois se réduire à trois dimentions majeures : la séquence elle-même, l'épigénétique et ses variations au cours de l'évolution. Ces dimensions interviennent à des degrés divers et de façon fortement intriquée dans la fonction et le devenir de toute région génomique, depuis les régions *à priori* non-transcrites comme les intergènes ou les éléments répétés jusqu'aux acteurs majeurs de la vie cellulaire, à savoir les gènes.

1. Types de gènes

Avant même de parler de contexte des gènes, il convient de définir la collection la plus riche possible de ceux-ci, en considérant à la fois les gènes d'ARN codants et non-codants (Figure 8).

Type de gène	Abbréviation
ARN de transfert	ARNt
ARN ribosomal	ARNr
Petit ARN nucléolaire	ARNsno
Petit ARN nucléaire	ARNsn
Petit ARN cytoplasmique	ARNsc
ARN microscopique interférence	ARNmi
ARN codant	ARNm

Figure 8 : Principaux types de gènes classiques d'ARN codant et non-codant

Dans l'état de nos connaissances , on peut penser que cette liste soit loin d'être exhaustive. Ainsi, récemment, une nouvelle catégorie de gènes correspondant à des précurseurs d'ARN mesurant *in fine* entre 26 et 33 nucléotides (Figure 9A), a été découverte (Aravin *et al.*, 2006). Il s'agit des ARNpi, interagissant avec les protéines PIWI (Figure 9B) et potentiellement régulés

par les protéines Tudor (Siomi *et al.*, 2010). Ils sont exprimés lors du développement ou ils interviennent dans le maintien de l'immortalité des cellules (Aravin *et al.*, 2007; Tang, 2010). Ils constitueraient le plus grand effectif d'ARN avec 35.000 entités localisés de manière unique chez la souris et le rat et atteignant 1,3 millions si l'on inclue les localisations multiples (http://pirnabank.ibab.ac.in/).

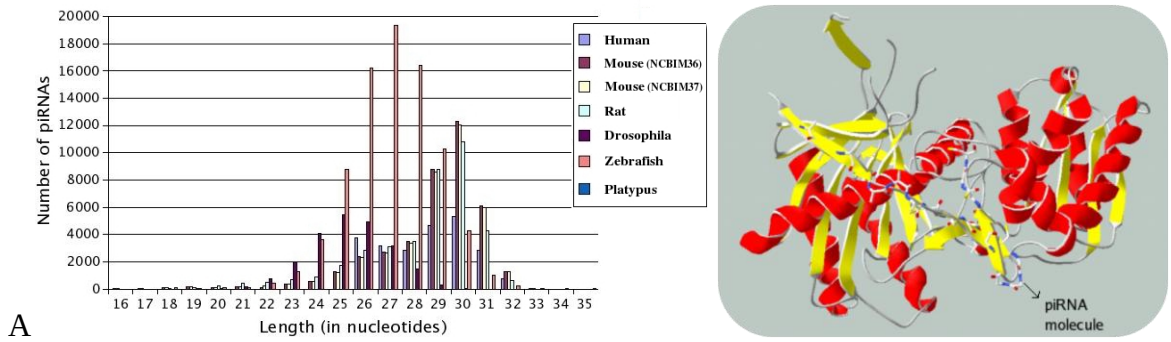


Figure 9 : Distribution en taille de ARNpi et interaction entre un ARNpi et sa protéine PIWI.

A/ Distribution en taille de ARNpi dans 7 génomes. B/Complexe ARNpi – PIWI par replacement de la séquence d'un ARN pi dans la structure de la protéine PIWI d'A. fulgidus (PDB/ 1YTU) (source : piRNAbank)

Dans le détail, ces ARN permettent d'une part de neutraliser l'expression des éléments répétés, dont ils dérivent pour la plupart, et d'autre part de bloquer l'expression des gènes localisés sur les chromosomes sexuels lors des phases de développement où cette expression est proscrite (Aravin *et al.*, 2007). Cette deuxième fonction s'opère dans le cytoplasme où ils séquestrent et engagent les ARNm correspondants dans des voies de dégradation.

Les réseaux d'interactions présidant l'expression des gènes s'avèrent donc de plus en plus complexes avec, en plus d'une régulation en cis, des acteurs géniques agissant en trans de plus en plus nombreux comme les ARNmi ou les ARNpi. Il est donc crucial de disposer de la collection de gènes la plus exhaustive pour appréhender pleinement le contexte informationnel d'un gène donné.

2. Cis-contexte : la séquence

2.1. Séquences régulatrices

Comme nous l'avons vu pour le facteur Staf, les séquences de fixation des éléments de réponses des protéines activant ou réprimant la transcription constituent une part informationnelle cruciale d'une région génomique. Leur identification et localisation permettent d'estampiller la région les contenant, comme « régulatrices ». Ces séquences sont fréquemment de petite taille, fortement dévoyées, localisées dans la région proche ou distale d'un gène et reconnues plus ou moins spécifiquement par différentes protéines. Dans le génome humain, il existe environ 2000 gènes codant pour des facteurs de transcription (Ravasi *et al.*, 2010), ceux-ci, pouvant reconnaître de nombreuses séquences différentes. De plus, du fait de la taille relativement courte de ces séquences, le promoteur d'un gène sera sous la régulation potentielle d'une combinatoire de plusieurs facteurs (Fedorova and Zink, 2008). Il en résulte qu'un réseau très dense d'interactions entres les protéines régulatrices et leur cibles mais également entre ces protéines elle-mêmes ,est mis en place et l'étude de ce réseau ne pourra se faire qu'à l'aune de la biologie des systèmes nécessitera. A cette fin, un atlas des interactions entre les facteurs de transcription a récemment été publié. Il contient pour l'instant 762 interactions chez l'humain, 877 interactions chez la souris et environ la moitié de ces interactions serait conservée entre les deux organismes (Ravasi *et al.*, 2010). Les données disponibles vont donc croissant mais ne sont pour l'heure pas suffisantes pour répondre de manière systématique aux problèmes biologiques adressés au génome entier. En conséquence, de telles analyses ne peuvent pour l'heure être menées que sur la base des prédictions que peuvent réaliser les modèles mathématiques et plus généralement la bioinformatique dont le rôle s'annonce de plus en plus central à mesure que les données s'accumulent sans pouvoir être toutes validées.

2.2. Eléments répétés

Véritables parasites du génome, issus de diverses lignées (Figure 10), constituant de 30 à 50% du génome des mammifères, ils sont une source d'informations sur des évènements géniques passés et ont également un impact sur le contexte présent des gènes (Levy *et al.*, 2007; Margulies *et al.*, 2005). Ils jouent un rôle dans l'évolution génomique, dans la structure du génome, dans la constitution des promoteurs, exons et terminateurs alternatifs, et peuvent générer de nouvelles jonctions d'épissage (Babushok *et al.*, 2007; Hasler *et al.*, 2007; Jurka, 2004; Kazazian, 2004; Peaston *et al.*, 2004; Speek, 2001).

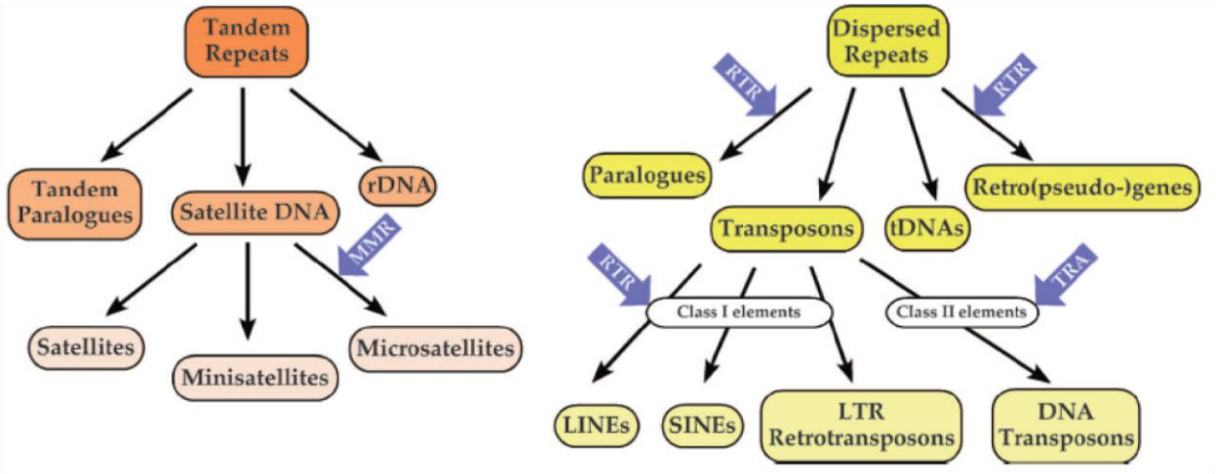


Figure 10: Classification des éléments répétés selon qu'ils dérivent d'éléments en tandem ou dispersés.

Certaines études les incriminent même dans la survenue de cancers (Lerat and Semon, 2007). De plus, il demeure possible qu'un ensemble de gènes présentant une distribution similaire d'éléments répétés, ou parasités par la même famille de ceux-ci puisse présenter certains traits contextuels communs. En effet, certains de ces éléments renferment de manière ultra-conservée de puissants promoteurs de la polymérase III. De fait, certains retrotransposons sont transcrits et jouent ainsi un rôle clé dans le transcriptome humain (Faulkner *et al.*, 2009). Ainsi, ces éléments répétés, supposément inutiles, présentent donc un contenu informationnel riche, notamment dans la régulation de la transcription.

2.3. Single Nucleotide Polymorphisms (SNP's)

Il s'agit de positions génomiques uniques connues pour présenter des mutations que l'on pourra retrouver à différentes fréquences dans différentes populations d'une même espèce. A ces données s'ajoutent les polymorphismes multinucléotidiques (MNP), les délétions et les insertions qui sont bien documentées. Dans le cadre des études associatives à l'échelle du génome, ces changements dans la séquence sont regroupés sous le terme de variants. Un amalgame est souvent fait entre SNP's et « variants fréquents » alors qu'il existe des SNP's rares. Pour les différencier, Cirulli et Goldstein proposèrent la classification donnée dans le Tableau 1 (Goldstein *et al.*, 2010).

Variant class	Minor allele frequency	Implications for analysis
Very common	Between 5 and 50%	Amenable to association analysis using current genome-wide association methods
Less common	Between 1 and 5%	Amenable to association analysis using variants catalogued in the 1000 Genomes Project
Rare (but not private)	Less than 1% but still polymorphic in one or more major human populations	Amenable to framework of extreme phenotype resequencing, as well as co-segregation in families
Private	Restricted to probands and immediate relatives	Difficult to analyse except through co-segregation in families. As linkage evidence will (by definition) be modest, discovery would be limited to the most recognizable of variants

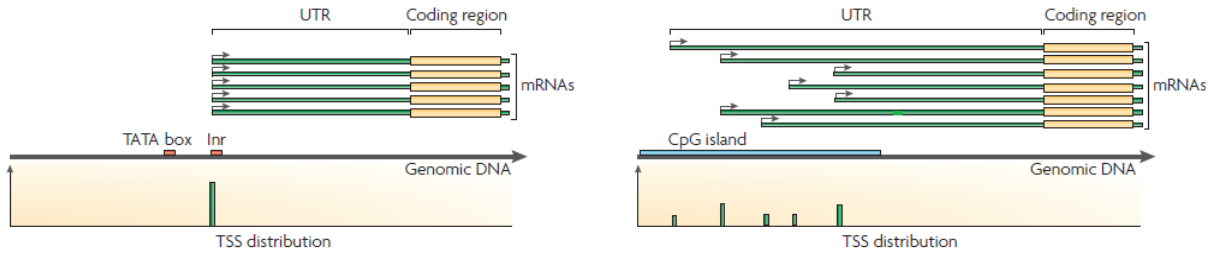
Tableau 1: Classification des SNP's liés à un phénotype donné selon leur fréquence d'apparition

La classification selon la fréquence d'apparition des variants est couplée aux impacts d'analyse que chaque classe de fréquence apporte (d'après Cirulli & Goldstein, 2010)

Le génome humain disponible et annoté étant un génome de référence, il ne rend pas compte du polymorphisme interindividuel où une mutation ponctuelle pourra avoir un effet drastique sur l'expression des gènes. C'est à cette fin que le projet « 1000 génomes» (www.1000genomes.org) - visant à séquencer au moins 1000 génomes humains et en déceler les variants – a été lancé en 2008. En fait à ce jour, près de 2000 génomes ont été séquencés dans le cadre de ce projet et le but est d'atteindre les 2500 d'ici fin 2011. Qu'elles soient liées à un phénotype, une pathologie ou simplement une perturbation de l'expression génique, ces mutations pourront changer drastiquement le contenu informationnel et fonctionnel d'une région génomique. Aussi, disposer de leur nombre et leur type dans cette région revêt donc un intérêt crucial pour en appréhender le contexte.

2.4. Ilots CpG

Ce sont des régions génomiques présentant un enrichissement supérieur à 10% en dinucléotides CpG sur une distance comprise entre 500 et 2000 paires de bases. Ils sont très abondants au sein des promoteurs de gènes de protéines. Chez l'homme, on dénombre une présence d'îlots CpG dans 72% des promoteurs où ils sont en général associés à des gènes dont l'expression est ubiquitaire (Saxonov *et al.*, 2006). Ils sont caractéristiques des promoteurs dits « larges » qui voient la machinerie de transcription se positionner dans une fenêtre statistique au niveau du promoteur de base et entrainer la création de plusieurs ARNm présentant différents TSS. Ce type d'initiation s'oppose à celle observée pour la plupart des gènes à expression tissu-spécifique qui aboutit à TSS mono-nucléotidique (comme c'est le cas pour la plupart des gènes dont l'expression est tissu-spécifique) (Figure 11).



A B

Figure 11 : Distinction entre promoteurs étroits (A), possédant généralement une boite TATA et promoteurs larges à îlots CpG (B).

Les promoteurs étroits disposent d'un site d'initiation unique débutant à l'élément Inr et sont en général couplés à une boite TATA alors que les promoteurs larges peuvent avoir de multiples TSS, ne présentent pas de TATA mais possèdent des îlots CpG. (Sandelin et al., 2007)

Ainsi la présence des îlots CpG, dans une région donnée, a peu de chance d'être le fruit du hasard et sera signe d'un contexte propice à la transcription et un bon indicateur contextuel d'une transcription ubiquitaire active. Une localisation atypique, au sein d'un gène par exemple et non autour de la région 5' des gènes pourra constituer une information importante, notamment dans l'optique de variants de transcrits.

3. Trans-contexte : Epigénétique

Au cis-contexte déjà très riche de valeur informationnelle vient s'ajouter une nouvelle dimension dès lors qu'on se focalise sur la structuration de la molécule d'ADN et toutes les protéines agissant en trans qui pourront l'affecter. Les règles présidant à l'interaction conditionnelle de ces protéines avec l'ADN et les modifications qui les affecteront seront des vecteurs informationnels capitaux car ce sont eux qui permettront ou non au éléments de cis-contexte de prendre sens. Dans le réseau informationnel de la transcription, les modifications des histones, l'état de condensation de la chromatine qui leur est directement lié et à la présence de la polymérase dépendant de cette accessibilité structurale chargeront du sens "activement transcrite" la région génomique correspondante.

3.1. Structure de la chromatine

La compaction tridimensionnelle du génome rend la séquence par elle même peu informative quant à un quelconque potentiel de la compaction dans la régulation ou le contrôle de

l'expression des gènes. L'accessibilité d'une région, dépendant directement de l'état de condensation de la chromatine est à considérer en premier lieu. Les molécules d'ADN longues de quelques centimètres sont compactées dans un noyau d'environ 5µm de diamètre en subissant plusieurs niveaux d'enroulements (Figure 12).

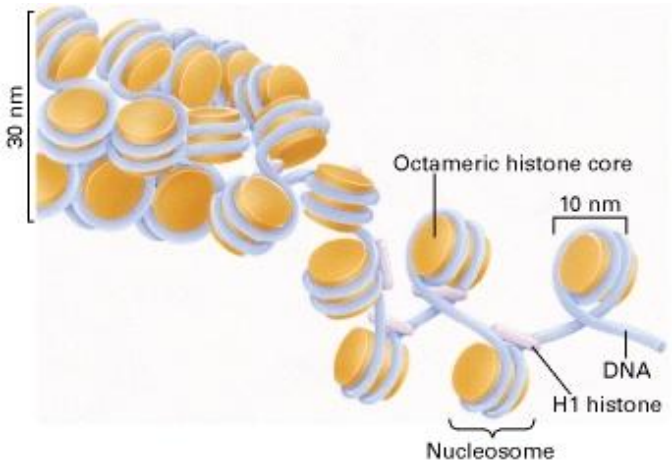


Figure 12 : Compaction de la chromatine selon le modèle solénoïde

L'ADN est enroulé autour de nucléosomes pour former la fibre de 10nm qui sera surenroulée pour former la fibre de 30nm. (adapté de Grunstein, 1992 dans Molecular and Cell biology 4th ed., 2000)

En premier lieu, l'ADN est enroulé autour d'un noyau d'histones associées en octamères. La fibre de 10 nm, renfermant les nucléosomes ainsi générés, subit des surenroulements successifs l'agençant en fibre de 30 nm pour, *in fine*, former les chromosomes métaphasiques que nous connaissons. La condensation de la chromatine n'est pas constante le long des chromosomes. On trouvera des zones d'euchromatine, correspondant à une transcription active et, à l'inverse, les zones d'hétérochromatine correspondront à des zones de transcription peu probable. L'état chromatinien variera en fonction de la phase du cycle cellulaire, de la lignée cellulaire et de la présence de facteurs de remodelage à même de la faire passer d'un état à l'autre. Expérimentalement, il est possible de mettre en évidence une sensibilité enzymatique différentielle à la DNase entre l'euchromatine et l'hétérochromatine. L'accessibilité limitée à l'ADN pour l'enzyme, générant de plus grands fragments lors de la digestion de l'hétérochromatine .

3.2. Occupation et modification des protéines-clé

3.2.1. Occupation génomique des nucléosomes

Les histones H2A, H2B, H3 et H4, assemblées en octamères sont les premières responsables de la compaction de la molécule d'ADN en formant les nucléosomes par association avec 146 pb d'ADN. L'histone H1 en s'associant à l'extérieur de la structure ainsi formée participe aussi à la compaction des nucléosomes. Les octamères d'histones sont constitués par l'association de deux exemplaires des histones H2A, H2B, H3 et H4. Ces quatre histones sont des protéines qui peuvent subir de nombreuses modifications post-traductionnelles, notamment au niveau des lysines et des arginines de leur domaine N-terminal. Les nucléosomes se distribuent de manière non uniforme sur le génome et sont retrouvés dans les régions codantes et non-codantes. La connaissance des règles présidant à leur positionnement n'est pas triviale puisque celles-ci est la clé de voute d'une discussion permanente du nucléosome avec son contexte informationnel, que ce soit au sein d'un organisme ou au cours de l'évolution. En effet, 5 mécanismes liant l'évolution de la régulation génique et le maintien, la déstabilisation ou la restauration du contexte chromatinien, au positionnement des nucléosomes ont récemment été identifiés chez la levure (Tsankov *et al.*, 2010). De plus, le partage de contexte entre un nucléosome et un élément répété de type Alu engendrerait une action réciproque qui conduirait entre autres à l'inactivation de l'élément répété, soulignant une fois encore les fortes connections entre les divers acteurs du contexte informationnel (Tanaka *et al.*, 2010).

La recherche d'un code signal au niveau de la séquence d'ADN permettant la mise en place du nucléosome a longtemps été noyée dans le bruit du positionnement indépendant de la séquence et par le glissement possible du nucléosome après la reconnaissance d'une séquence.
Les études structurales ont récemment levé le voile sur certains éléments-clés régissant le positionnement, la densité et la dynamique des nucléosomes dans une région génomique. Il semblerait que deux mécanismes soient impliqués. Tout d'abord un adressage fortement ancré du nucléosome sur une séquence répétée plusieurs fois de consensus YRRRRRYYYYYR – le plus souvent : CGGAAATTTCCG (M1)(Figure 13A) – assure une courbure optimale du polymère d'ADN (Gabdank *et al.*, 2009; Salih *et al.*, 2008b). Puis, un positionnement par « packing » se met en place où les nucléosomes se positionneraient par propagation, de manière aspécifique autour d'un nucléosome fortement ancré, faisant office de « barrière » (Figure 13B)(Mavrich *et al.*, 2008).

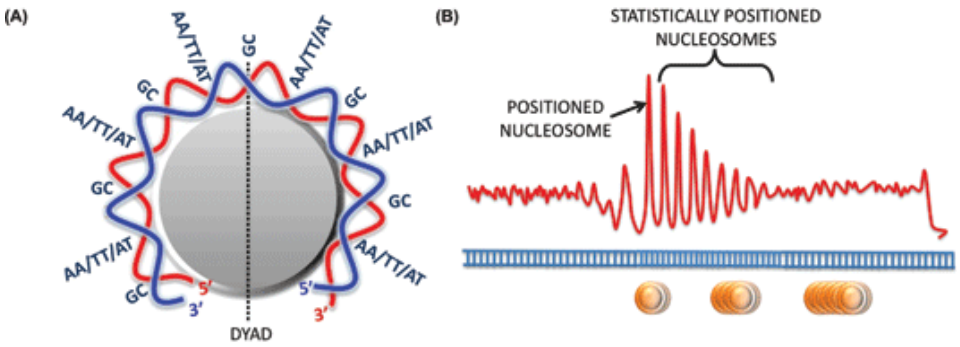


Figure 13 : Motif de séquence ADN enroulée autour des nucléosomes barrières et distribution des autres nucléosomes par aggrégation.

A/ Alternance des motifs de dinucléotides AA/TT/AT et CG dans l'ADN nucléosomal. B/ Effet « barrière » proposé par Mavrich et al. (2008) où un nucléosome sert de repère spécifique pour un positionnement aspécifique d'un gradient de nucléosomes. (Arya and Schlick, 2009)

De plus, le motif d'ancrage assurerait plusieurs positionnements possibles où le passage de l'un à l'autre serait ATP-dépendant (Kim *et al.*, 2006) et responsable du « flou nucléosomal » (Arya and Schlick, 2009)(Figure 14). La répétition du dinucléotide CG tous les 10 pb (10,4 statistiquement) revêt un caractère important dans le placement de la barrière et se retrouve également dans les éléments répétés Alu où l'on observe une périodicité de 31-32 pb, soit 10,4pb x 3 (Salih *et al.*, 2008a).

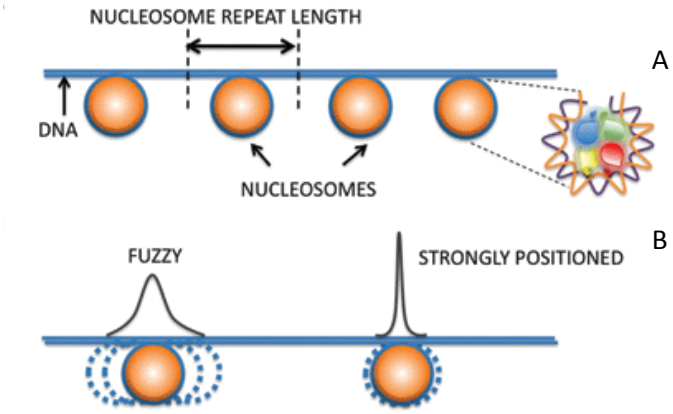


Figure 14 : Positionnement schématique de nucléosomes sur l'ADN

A/ Ce positionnement montre la distance inter-nucléosomes (A) définissant une longueur fixe et répétée le long de l'ADN. B/ Le « flou nucléosomal fruit du décalage possible de certains nucléotides autour de leur position d'ancrage théorique» (Arya and Schlick, 2009)

En outre, la distribution nucléosomale à proximité des gènes est particulière. En effet, on observe une absence caractéristique en nucléosome (NDR pour *Nucleosome Depleted Region*) directement en 5' et en 3' des gènes disposant d'un promoteur « ouvert » facilitant l'accès des facteurs de transcription à leurs sites de fixation. Sur ces promoteurs, les nucléosomes perdent rapidement leur périodicité d'agrégation (en 2-3 nucléosomes) et intègrent préférentiellement un variant de l'histone H2A. De plus la région en aval du TSS dans ces promoteurs « ouverts » est plus dense en nucléosomes (Figure 15).

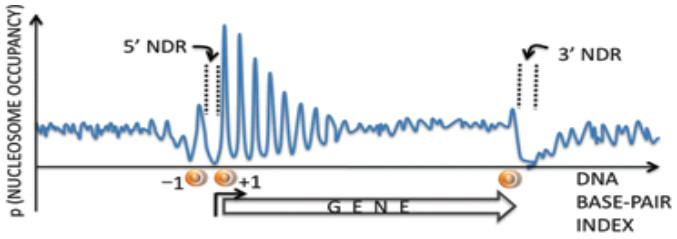


Figure 15 : Occupation relative en nucléosomes le long d'un gène entier.

Une région juste en amont du TSS est dépourvue de nucléosomes (NDR) et précède en aval une occupation par agrégation de nucléosomes de moins en moins marquée à mesure que l'on progresse dans le gène jusqu'à sa fin présentant une NDR caractéristique (Babbitt et al., 2010)

Toutefois, une présence dense de nucléosomes n'est pas toujours synonyme d'absence d'activation de la transcription car le site de liaison du facteur de transcription en question peut se situer sur la région liant deux nucléosomes successifs d'un promoteur dit « occupé ». En complément de ces comportements « promoteurs », le code génétique atténue la possibilité de rencontrer un signal fort de liaison d'un nucléosome dans les exons des gènes et l'on observera dans les gènes plutôt un gradient statistique en nucléosomes faiblement liés. Enfin, il a été montré qu'un nucléosome était moins spécifiquement associé à la partie 3' des gènes avant de laisser la place à la région en aval débutant par une autre NDR (Babbitt *et al.*, 2010) (Figure 15).

3.2.2. Modification post-traductionnelle des histones

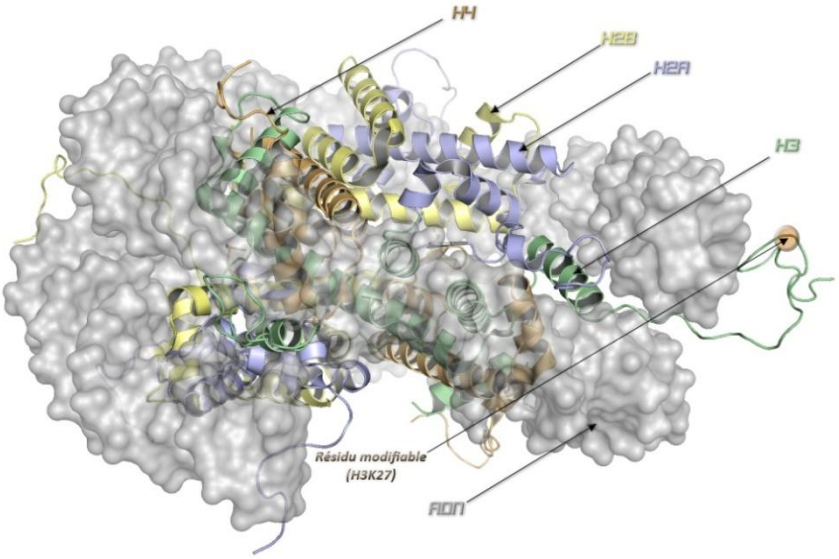


Figure 16 : Structure tridimensionnelle d'un octamère d'histones

L'octamère d'histones est le complexe protéique (en cartoon coloré) autour duquel est enroulé la double hélice d'ADN (en mesh gris transparent) avec mise en évidence de la lysine 27 de la queue de l'histone H3, résidu modifiable par N-acétylation occasionnant une décompaction de la chromatine et donc une activation de la transcription (PDB: 1EQZ)

Le nucléosome est une unité dynamique sujette à de nombreuses modifications de ses histones se répercutant sur la structure de la chromatine (Figure 16). Chaque résidu modifiable est donc porteur d'une charge informationnelle à même de changer profondément son contexte. On compte aujourd'hui plus de 300 couples différents de résidus d'histones modifiés et de types de modification, laissant entrevoir un code complexe dans la fonction de ces modifications (Figure 17). L'effet, selon le résidu et selon le type de modification, pourra être une répression ou une activation des gènes via respectivement une compaction ou une décompaction de la structure chromatinienne. L'acétylation de l'histone H3 sur ses lysines 9 et 14 et sa méthylation sur les lysines 4, 36 et 79 est synonyme de chromatine active alors que les méthylations des lysines 9 et 27 de H3 sont liées à une répression. Cependant, des études récentes ont montré un recouvrement des modifications actives et répressives dans différentes sous-populations chromatiniennes (Bernstein *et al.*, 2006; Roh *et al.*, 2006).

Outre ce rôle régulateur, les modifications des histones, en étroite collaboration avec les facteurs de réparation de l'ADN (pour revue : (Mendez-Acuna *et al.*, 2010)), interviennent également dans certaines fonctions cognitives (Koshibu *et al.*, 2009) et sont impliquées dans de nombreuses maladies. Les enzymes impliquées dans ces modifications des histones constituent donc des cibles thérapeutiques de choix (Spannhoff *et al.*, 2009) pour des pathologies allant de

l'infection par le virus de l'herpès (Liang *et al.*, 2009), au cancer (Bolden *et al.*, 2006) en passant par les maladies cardio-vasculaires ou rénales (Bush and McKinsey, 2010) et l'asthme (Barnes, 2009).

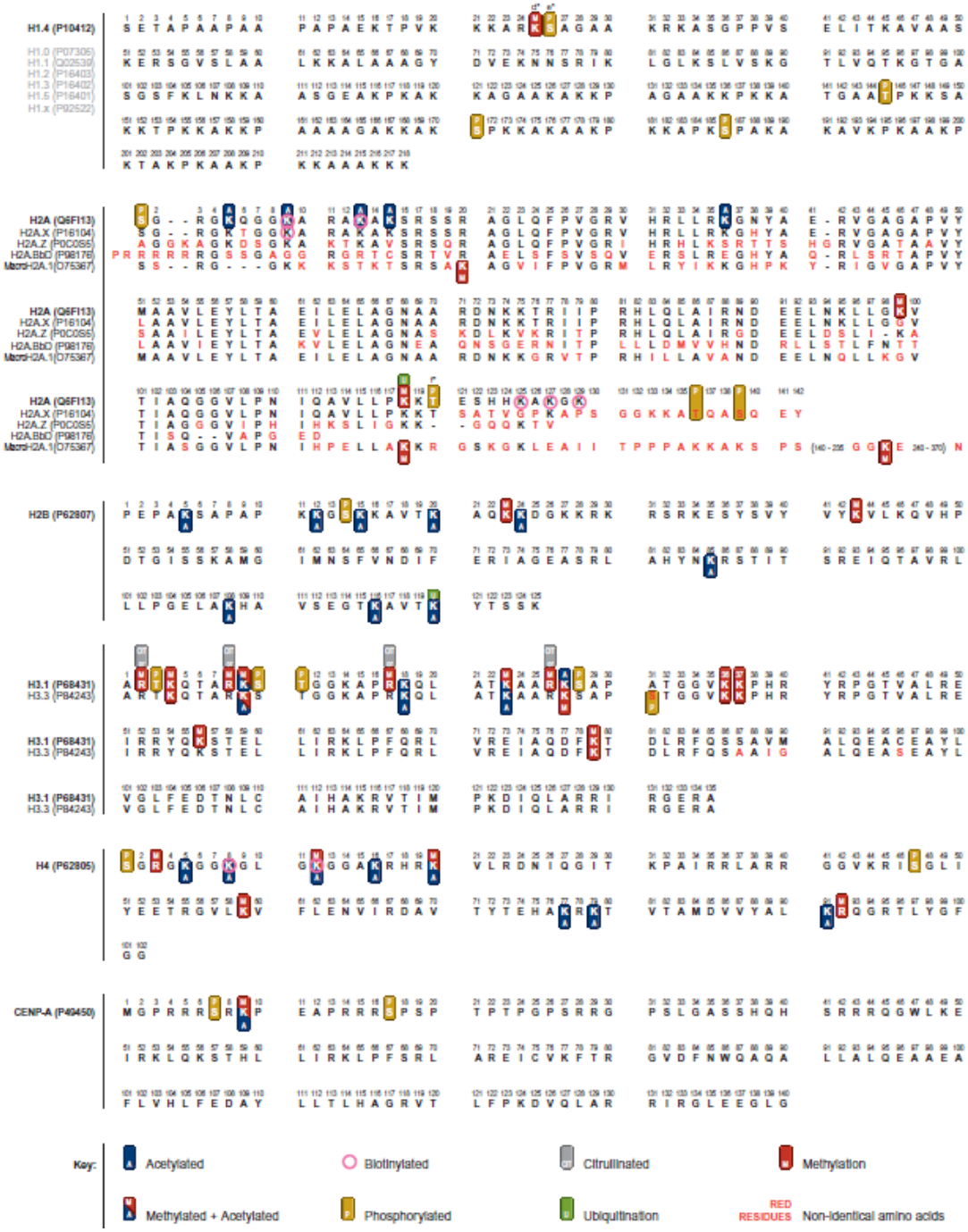


Figure 17 : Carte des résidus modifiables des diverses histones et variants d'histones ainsi que leur(s) modification(s) associée(s)

Les séquences des histones sont figurées horizontalement alors que les différentes histones sont disposés verticalement, se décantant éventuellement en divers variants . Six modifications sont figurées, couplées à une double modification. On voit ici que l'histone la plus modifiée est H3. (d'après : http://www.abcam.com)

3.3. Présence des machineries de transcription

La transcription est un phénomène disposant, entre autres, d'une phase d'initiation et d'une phase de terminaison qui sont les phases d'inertie chronophage du système et sont donc statistiquement les plus observées par les expériences de type « Chip-seq » (voir chapitre 2, paragraphe 1.2.1) . Il en résulte une vision "stroboscopique" de la machinerie de transcription que l'on verra "attendre" au début ou à la fin de la séquence à transcrire. Ainsi la présence de la machinerie de transcription, à l'intérieur ou à faible proximité de toute région génomique, participe évidemment du contexte et fournit une information sur l'aptitude de cette région à être transcrite. Au delà des cas attendus des promoteurs et des parties 3' de gènes annotés, la présence de la machinerie de transcription pourra s'avérer signe de promoteurs alternatifs, de variants d'épissage ou de gènes encore inconnus. Elle apporte donc une plus-value informationnelle venant en complément de l'état de condensation chromatinien et de sa régulation

A la lumière de ce qui précède, les éléments du trans-contexte que sont la structure chromatinienne, le positionnement et la modification des nucléosomes ou la présence d'une machinerie de transcription occupent une place stratégique dans le réseau informationnel. A la fois inter-imbriqués et ouverts sur les éléments du cis-contexte, ils sont de fait des vecteurs lourds de sens pour une région génomique

4. Contexte phylogénétique

La conservation au niveau de la séquence a été largement utilisée, que ce soit pour déterminer la présence d'un gène dans différentes espèces où permettre la caractérisation de l'importance de petites séquences potentiellement régulatrices. Dans l'idéal, les régions « utiles » sont plus conservées que les régions « inutiles », le tout étant de s'entendre sur la définition de conservation et d'utilité. En effet, une conservation stricte des bases n'est pas une absolue nécessité pour que la séquence homologue ait un rôle fonctionnel similaire, sinon identique à la séquence initiale. À l'échelle du résidu, une mutation pourra s'avérer plus ou moins neutre, selon sa position dans la séquence et la base par laquelle elle est substituée. À l'échelle du génome, une dynamique d'extinctions et d'apparitions de sites à potentiel régulateur reconnus chacun par des facteurs différents est un des moteurs de l'évolution. Ainsi, la non-conservation d'une séquence candidate à régulation, n'impliquera pas forcément que celle-ci ne soit pas

fonctionnelle mais simplement que tout ou partie du contexte informationnel de cette région aura été sollicité pour contrebalancer cette absence de conservation de séquence, que ce soit l'épigénétique (Tsankov *et al.*, 2010) ou le passage de relai à un autre facteur de transcription au cours de l'évolution par émergence d'un site reconnu par celui-ci. Ce phénomène de mutations est d'ailleurs très difficile à estimer dans les mécanismes de transcription où la pression de sélection dispose d'une marge de manœuvre assez large, du fait du faible impact d'une substitution sur l'affinité d'un facteur de transcription pour ses sites. Ces derniers évoluent en conséquence assez rapidement jusqu'à la substitution critique qui abolira la fixation du facteur à ces sites. Ces dernières années, le concept même de *phylogenetic footprinting* (voir chapitre 4, paragraphe 2.1) a montré ses limites face à de tels comportements et l'on préfère désormais considérer des régions plus larges que celles strictement conservées et parler dès lors de « région à potentiel régulateur » (Balmer and Blomhoff, 2009). Quelque soit la méthode pour l'établir la conservation entre organismes offre une dimension temporelle au contexte informationnel d'une région génomique. Attendu que les régions les plus conservées d'un génome sont majoritairement des exons et des séquences régulatrices, des éléments lourdement chargés en information, connaitre le degré de similarité phylogénétique permettra d'apprécier la conservation du contenu informationnel sur une échelle de temps.

Les trois axes majeurs du réseau d'information du génome, le cis-contexte, le trans-contexte et le contexte phylogénétique, contexte de premier niveau fortement ancré dans la transcription, sont, à la lumière de ce qui précède, des composantes majeurs concourant de l'être et du devenir d'une région génomique. Leurs multiples implications et imbrications en font un maillage dense, aux limites de l'écosystème génomique auquel les `omics ont donné accès et permis de jeter un nouveau regard, notamment sur la transcription, forçant à remettre en cause bon nombre de dogmes établis.

Chapitre 3 : `omics, nouveau regard sur la transcription

1. Le projet ENCODE

1.1. Buts et enjeux

Ces dix dernières années, quelques projets massifs de décryptage ou d'annotation plus ou moins fins de génomes ont vu le jour. Chez les organismes modèles de mammifères que sont l'homme et la souris, un réel effort mondial a été mis en place en vue d'accéder à une caractérisation idéalement exhaustive des données liées à ces génomes. Si chez la souris le projet FANTOM (Bono *et al.*, 2002) est à ce jour bien avancé, le projet EncODE (Encyclopedia of DNA éléments), pour l'heure uniquement dédié à l'homme, est le plus abouti. Il met en collaboration 16 laboratoires majeurs (MIT, Sanger…) qui échangent données et techniques. Du fait d'évidents problèmes de reproductibilité, le consortium s'est entendu sur les protocoles standardisés afin que les données soient les plus homogènes possibles et puissent être appréhendées ensemble lors d'une analyse commune. Cette nécessité est apparue évidente au vu du peu de reproductibilité des expériences de microarrays sur lesquelles reposaient le projet lors de la phase pîlote focalisée sur 1% (300 Mb) du génome humain (Birney *et al.*, 2007). Au passage à la phase de production considérant le génome entier, l'émergence de techniques issues du séquençage à haut-débit (*deep-sequencing*) détrôna les microarrays, celles ci se faisant plus anecdotiques au sein du projet. La firme Affymetrix en tant que partenaire du projet, pris en charge la tâche concernant les mesures d'expression génique via ses puces « exon arrays ».

En soi, la phase pîlote du projet EncODE permit à elle seule de décrire des mécanismes insoupçonnés liés à la transcription. Ainsi, des phénomènes aussi variés que le "trans-splicing" (fusion lors de l'épissage d'ARN chimères issus de deux gènes proches) ou l'abondance de sites de liaison de facteurs de transcription dans la région 3' des gènes, permettant d'envisager l'activation d'un autre gène proche ont été mis en évidence. Ceci a profondément altéré les dogmes et forcé à un changement des méthodologies de questionnement des gènes, obligeant notamment à considérer le génome dans son intégralité pour une recherche de sites de fixation de facteurs de transcription, et non plus aux alentours du TSS comme par le passé. Ce fait abonde également dans le sens de l'obligation de considérer l'environnement génique d'un gène car un autre gène peut receler les clés de son expression. Enfin, cette phase pîlote a aussi

entre autres phénomènes permis de reclasser la boite TATA, pensée ubiquitaire, à seulement 10% des promoteurs de gènes, essentiellement tissu-spécifiques.

1.2. Evolution de la technique

L'appréhension du génome humain complet via le *deep-sequencing* apporta donc une réponse adaptée aux nouveaux challenges lancés par la phase pîlote EncODE. Les techniques se diversifièrent et sont globalement classables en 3 catégories. Celles ciblant soit une protéine liée à l'ADN, soit l'ARN et celles ciblant directement l'ADN (Figure 18).

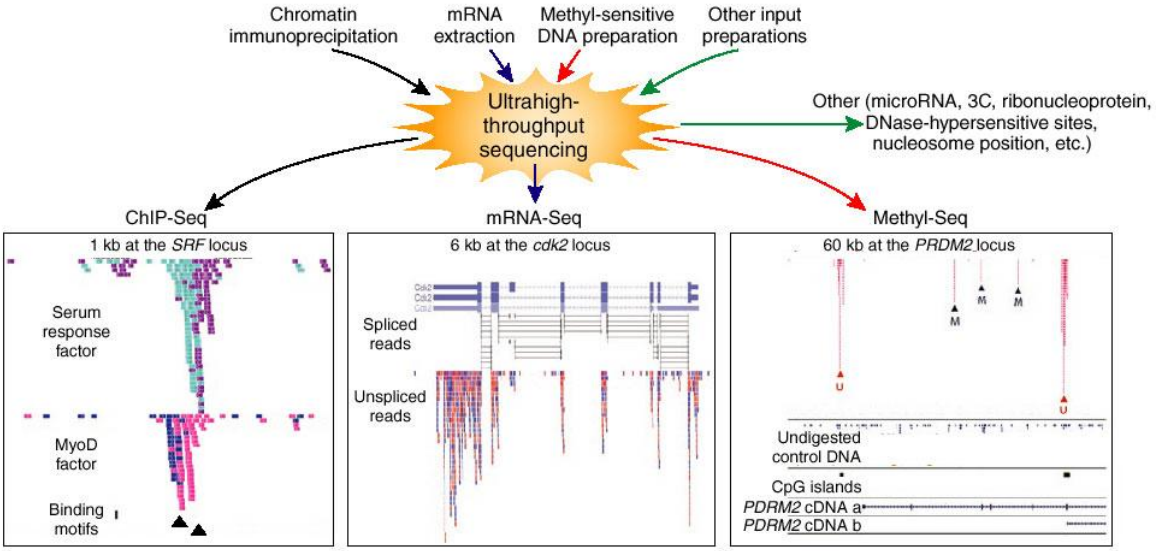


Figure 18 : Résumé des diverses techniques couplées au séquençage massivement parallèle.

Il est possible de bien noter ici les techniques ciblant une protéine (partie de gauche), l'ARN (partie centrale) et l'ADN (partie droite).(Wold and Myers, 2008)

1.2.1. Du Chip-on-chip, au ChIP-PET et au ChIP-Seq

L'application des nouvelles techniques d'analyse à l'ADN isolé par immunoprécipitation de la chromatine (ChIP) permet d'accéder à un nombre toujours plus important de sites de fixation d'une protéine dans l'ADN. La technique de ChIP consiste in vivo à fixer de manière covalente une protéine sur son site de fixation, à fragmenter la chromatine, puis à isoler, par l'emploi d'un anticorps spécifique, la protéine d'intérêt associée à l'ADN. Après réversion du pontage, l'ADN ainsi obtenu peut être analysé de différentes manières et une analyse par PCR est alors envisageable. Des expériences d'analyse à haut-débit, de l'ADN imunoprécipité, ont été

réalisées en hybridant cet ADN sur des puces à ADN (ChIP-on-chip). Toutefois, le choix de sondes étant très souvent défini, l'analyse exploratoire n'était pas permise et un biais *a priori* a donc souvent été introduit .

La technique de PET (Paired Ending diTags), couplée au ChIP, (ChIP-PET) permit d'assembler, dans un même vecteur, des étiquettes issues des extrémités de plus d'une dizaine de fragments différents d'ADN immunoprécipités (Wei *et al.*, 2006) . Après clonage et amplification, l'ADN était séquencé notamment sur un séquenceur Roche 454 (Figure 19). Il était ainsi possible d'obtenir des informations de séquence sur plusieurs centaines de milliers de fragments. Cette technique a désormais été supplantée par la technique du ChIP-Seq qui présente l'avantage de passer par moins d'étapes du fait d'un séquençage quasi-direct des produits immunoprécipités après amplification et génération de clusters (Figure 20).

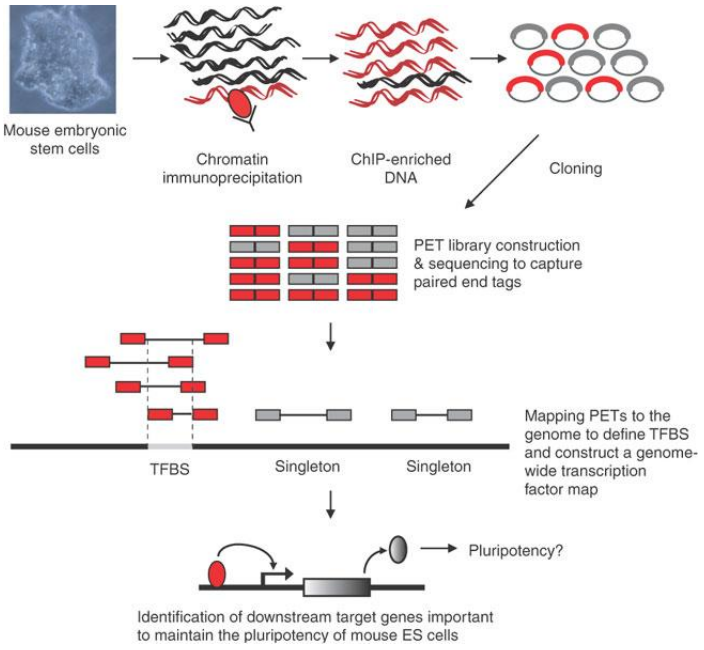


Figure 19 : Principe de la technique de ChIP-PET (Loh *et al.*, 2006)

Depuis 2007, la technique de ChIP fut couplée au *deep-sequencing* et cette approche appelée ChIP-Seq représente désormais l'approche de choix (Johnson *et al.*, 2007). L'ADN isolé par ChIP subit divers traitement (remplissage des extrémités, addition d'adaptateurs) avant d'être amplifié par PCR, puis séquencé. Cette technique présente peu de biais *a priori*, seul l'étape d'amplification par PCR pouvant éventuellement fausser l'analyse (Kozarewa *et al.*, 2009). Du fait des coûts encore élevés des techniques de *deep-sequencing* celles ci constituent parfois un complément au ChIP-on-chip qui est privilégié dans le cas de facteurs de transcription

spécifiques devant être testés de manière répétée dans de nombreuses conditions (Liu *et al.*, 2010).

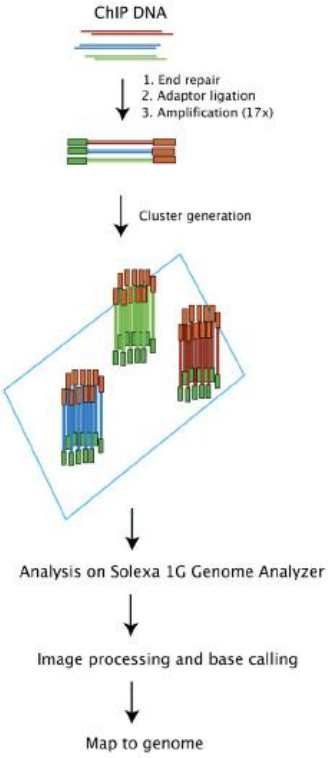


Figure 20 : Protocole schématique de la technique de ChIP-Seq

Un des très grands avantages de cette technique réside dans la production de la séquence de plusieurs millions d'étiquettes différentes (environ 20 millions d'étiquettes de 32 pb à l'heure actuelle sur un séquenceur Solexa Illumina) issues d'autant de fragments différents. Ces étiquettes sont ensuite replacées sur le génome par bioinformatique. L'apparition de pics constitués d'étiquettes se chevauchant est caractéristique de la présence d'un site de fixation de la protéine d'intérêt. De plus, dans la situation idéale, il y a co-localisation entre le sommet d'un pic et le site de fixation de la protéine étudiée dans le génome (Figure 21).

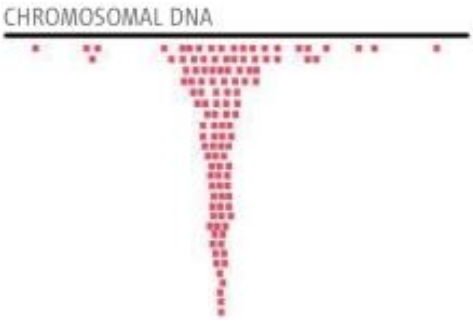


Figure 21: Replacement des tags de ChIP-Seq sur le génome. Idéalement le site de fixation se situe dans la partie la plus étroite du pic.

Un des avantages supplémentaires du ChIP-Seq est que l'ADN peut aussi être séquencé en utilisant le séquenceur de nouvelle génération Hélicos qui est capable d'analyser une molécule unique sans passer par l'intermédiaire de clusters (Goren *et al.*, 2010) et produire des séquences de 35pb en moyenne. Par ailleurs, il apparaît aujourd'hui indispensable pour certains investigateurs de systématiquement coupler une expérience de ChIP-Seq avec une expérience-contrôle de séquençage réalisée sur de l'ADN isolé de chromatine fragmentée pour détecter les biais de séquençage et ajuster les résultats.

1.2.2. De la transcriptomique au RNA-Seq

Les puces de transcriptomique permirent dans un premier temps de suivre les taux d'expressions de plusieurs gènes annotés d'une lignée cellulaire (Shalon *et al.*, 1996). Il fut possible avec leurs dernières versions, les "*tiling arrays*", de se détacher des annotations en couvrant le génome (Bertone *et al.*, 2005). Ceci alors même que le besoin se faisait sentir au vu du nombre insoupçonné de transcrits correspondants à des régions non-codantes, difficilement capturables. En dépit des informations obtenues par l'analyse de bibliothèques d'ADNc générées par les techniques de SAGE (Velculescu *et al.*, 1995) et par le séquençage de bibliothèques d'ADNc par l'approche de *Massively Parallel Sequencing Signature* (MPSS) (Brenner *et al.*, 2000), le *deep-sequencing* dédié à l'ARN (RNA-Seq) s'imposa comme un outil simple et abordable permettant d'accéder aux transcriptomes entiers, balayant les problèmes de traitement du signal des sondes de transcriptomique (Wang *et al.*, 2009).

Les transcriptomes déterminés par RNA-Seq couvrent déjà des lignées cellulaires issues d'une une douzaine d'organismes, incluant l'homme, la souris, la levure, le ver et la drosophile ainsi que quelques autres organismes non-modèles et quelques plantes et procaryotes. (Marguerat and Bahler, 2010). Le principe consiste dans un premier temps en une fragmentation des ARN, suivie de l'addition d'adaptateurs et la transformation en ADNc des différents fragments ainsi générés. Ensuite la séquence des extrémités des différents ADNc est établie et les différentes séquences sont replacées sur le génome (Figure 22) ce qui permet de détecter les jonctions exon/intron (Figure 23A). L'évolution de cette technique donne accès aux taux d'expression, permettant d'évaluer jusqu'au taux de chaque exon, de déterminer les jonctions d'épissage, de rechercher des transcrits non-annotés tels que de nouveaux ARNmi ou ARNpi ou encore des fusions de transcrits (Figure 23B). De plus la technique de RNA-Seq se décline désormais en

version courte (séquençage de 32 nucléotides à partir de l'extrémité) et en version longue (séquençage de 75 nucléotides) et peu identifier le brin d'ADN dont est issu l'ARN analysé.

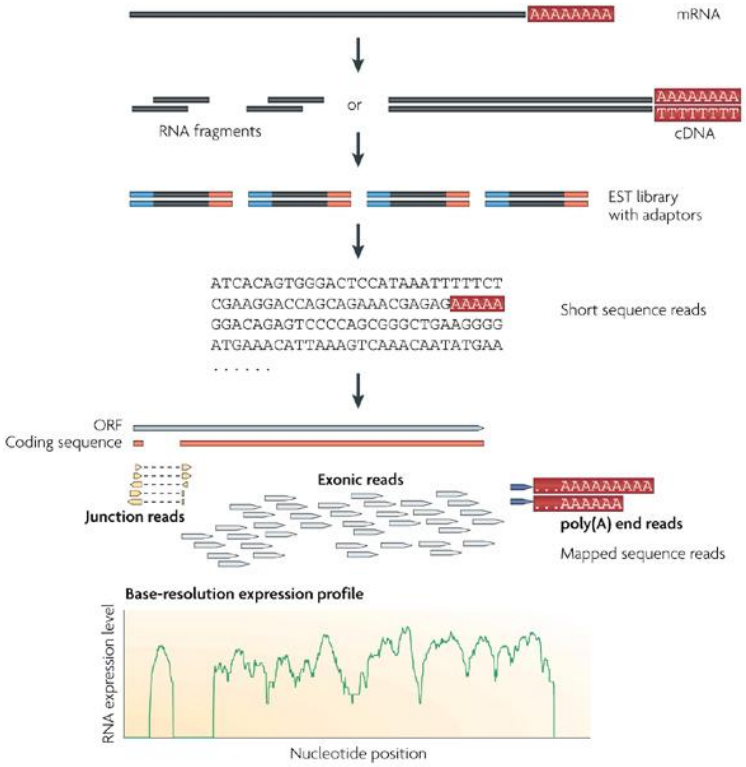


Figure 22: Principe de la technique de RNA-Seq (Wang *et al.*, 2009).

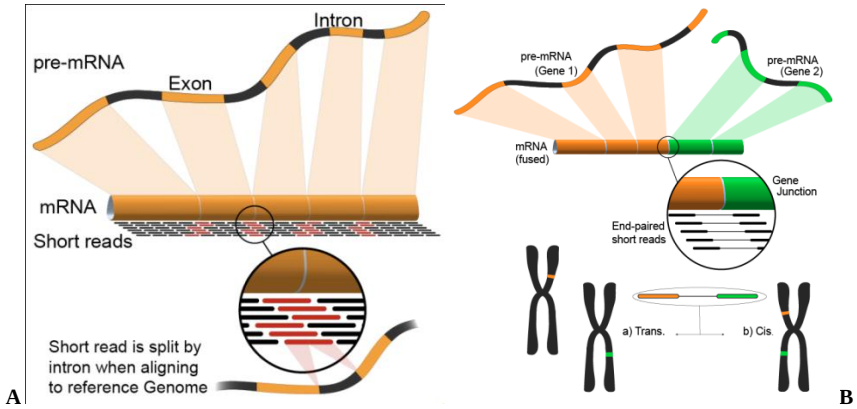


Figure 23 : Applications du RNA-Seq.

A/ Les tags localisés de manière non-continue sur le génome sont signes de la présence d'introns d'introns. B/ Les fusions de gènes s'opérant sur le même chromosome (cis-fusion) ou entre chromosomes (trans-fusion) sont également détectables par la technique. (source: en.wikipedia.org).

1.2.3. De la DNaseI au DNase-Seq

L'hypersensibilité à la nucléase de certaines régions de la chromatine est un fait bien établi et cette sensibilité préférentielle couplée en aval à la technique de *southern blot* a été largement utilisée pour identifier des régions hypersensibles et fort probablement actives du génome. (Keene *et al.*, 1981; McGhee *et al.*, 1981) Elle repose sur le fait qu'une région de chromatine active verra ses nucléosomes déplacés, de sorte que la compaction chromatinienne sera moindre et que l'accès à l'ADN (Gross and Garrard, 1988) d'une enzyme de clivage sera favorisé. Ces zones très accessibles sont particulièrement riches en éléments régulateurs de l'expression des gènes (Felsenfeld and Groudine, 2003; Gross and Garrard, 1988). La phase pîlote du projet EncODE permit de compléter cette richesse par la mise en évidence d'autres éléments proches des zones hypersensibles à la DNAse : histones modifiées, zones de réplication précoce et TSS. De plus, une partie de ces zones apparaît dans les gènes (45%) et plus particulièrement dans les introns (39%), ce qui en soit n'a rien d'étonnant vu la plus grande taille de ceux ci au regard de la taille des exons (Boyle *et al.*, 2008) (Figure 24).

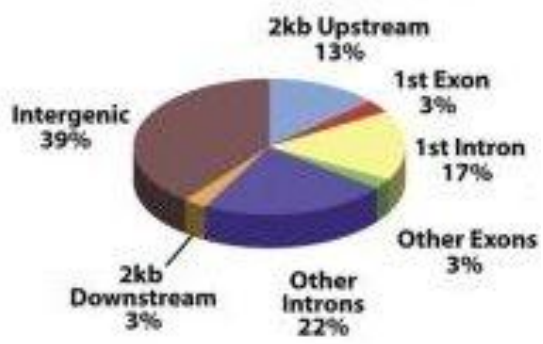


Figure 24 : Répartition des zones hypersensibles à la DNAse I du génome humain.

Le développement des nouvelles stratégies de séquençage permit, de remplacer les approches classiques d'analyse de l'ADN protégé contre l'action de la DNAse par le *deep-sequencing*. La méthode proposée par Boyle *et al.* fut optimisée pour donner le DNAse-Seq qui augmenta le rapport signal/bruit tout en limitant les séquences discontinues (Figure 25) (Song and Crawford, 2010).

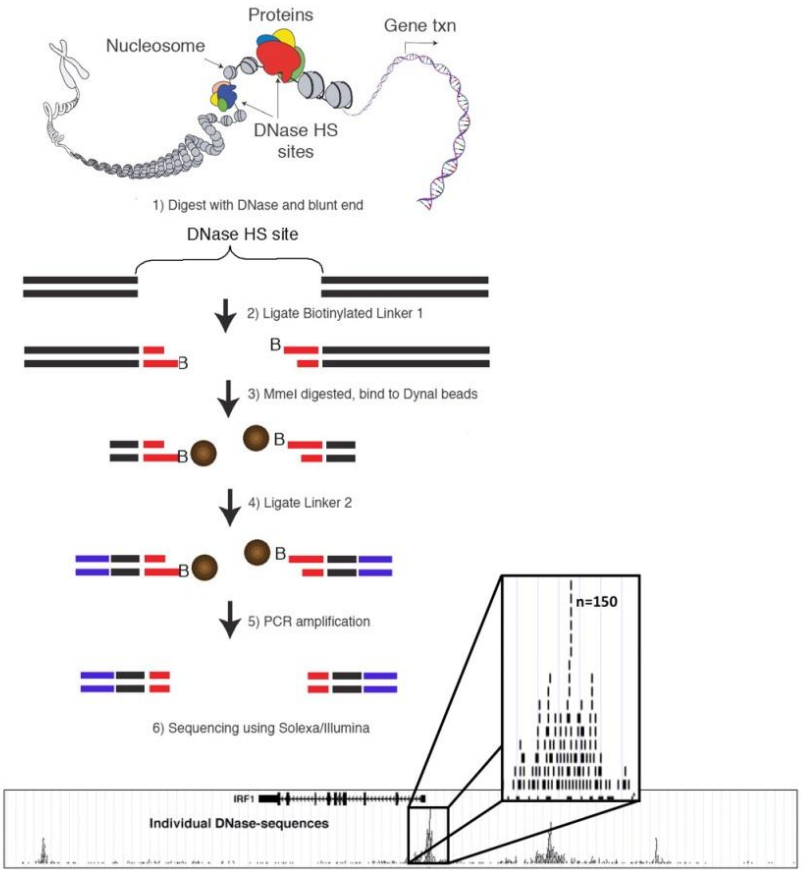


Figure 25 : Protocole de la technique de DNAse-Seq (Song et Crawfard, 2010)

Comme pour les autres techniques, il en résulte, après séquençage et placement des étiquettes sur le génome d'intérêt, que de nombreux tags sont agrégés en pics permettant de délimiter les zones hypersensibles. En plus de cette technique, le projet EncODE intègre également les expériences de FAIRE-Seq (*Formaldehyde Assisted Isolation of Regulatory Elements*) ou la digestion enzymatique est remplacée par un pontage au formaldéhyde suivi d'une fragmentation par sonication (Figure 26).

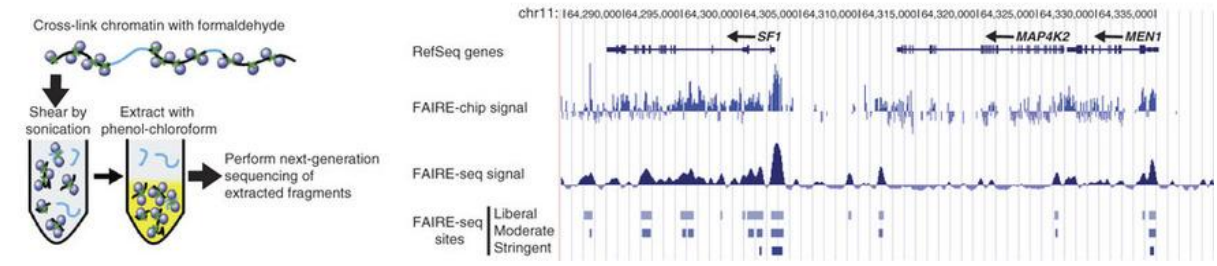


Figure 26 : Principe de la technique de FAIRE-Seq

A/ Des pontages sont réalisés entre les protéines de la chromatine et l'ADN à la manière du ChIP puis la chromatine est fragmentée par sonication, les fragments résistants sont alors massivement séquencés. B/ Après replacement sur le génome, les tags séquencés produisent un signal d'amplitude variable permettant de catégoriser les sites et les différents types de sensibilité.

1.2.4. Methyl-Seq

La méthylation de l'ADN, en position 5 de la cytosine, initialement observée comme associée avec l'hétérochromatine, (Holliday and Pugh, 1975) a de ce fait souvent été liée à la répression de l'expression des gènes (Jaenisch, 1997). Ceci étant, des études récentes font état d'activation différentielle de gènes tissu-spécifiques associée la présence de cette méthylation (Yagi *et al.*, 2008). Il en découle donc que la méthylation de l'ADN à un rôle complexe dans la régulation des gènes, interconnectée avec la structure chromatinienne. Il y a donc une importance cruciale à pouvoir accéder l'intégralité des sites méthylés au niveau du génome. Ceci fut rendu possible par le développement de diverses stratégies. Tout d'abord, la technique de Methyl-Seq (Brunner *et al.*, 2009) qui couple le séquençage haut-débit à la fragmentation de l'ADN obtenu par digestion par deux enzymes de restriction (Cedar *et al.*, 1979) MspI et HpaII qui reconnaissent tout deux la séquence CCGG. MspI digère l'ADN indépendamment de l'état de méthylation de la séquence cible. Par contre HpaII ne clive que les séquences non méthylées. De fait en soustrayant les produits issus de la première digestion, du total des produits obtenus par la seconde, il est possible d'accéder aux résidus méthylés. Le couplage au *deep-sequencing*, puis le replacement des fragments sur le génome, interrogé dans son intégralité, permet de positionner dans l'ADN les sites de méthylation. L'approche de choix pour accéder au methylome est désormais le séquençage à haut débit de l'ADN après traitement de cet ADN par le bisulfite (BS-Seq) ; cet agent modifie spécifiquement les cytosines non méthylées en uracile tout en laissant les m5C inchangés.

1.3. Expériences réalisées au cours du projet EncODE

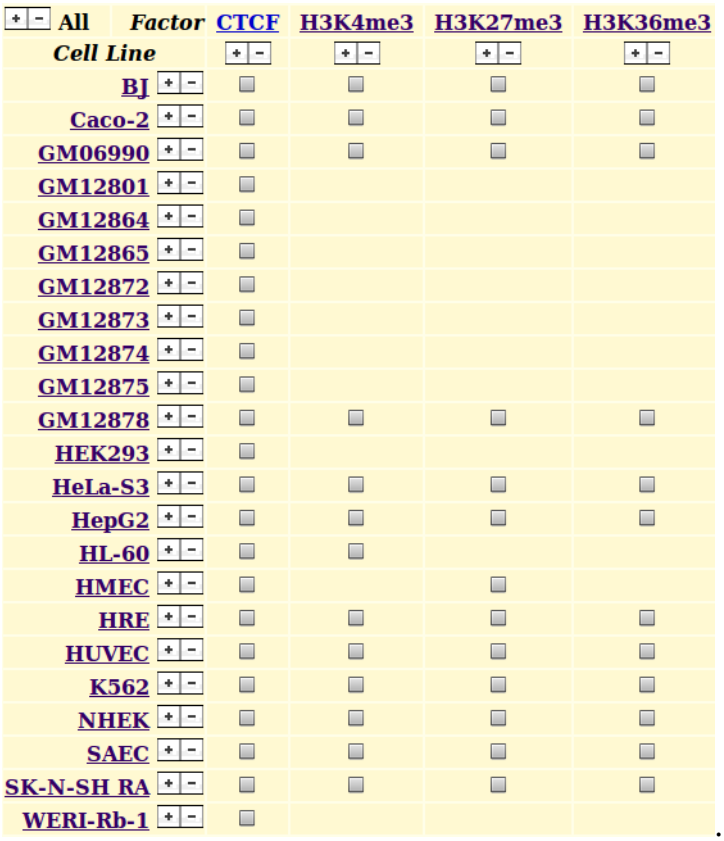
1.3.1. Cartographie des modifications d'histones

Le projet Encode, via le Broad Institute (http://www.broadinstitute.org) constitué du Massachussetts Institute of Technology (MIT) et d'Harvard, recense à ce jour la cartographie de 13 modifications d'histones. Après la formation d'un pontage covalent entre l'ADN et les protéines et fragmentation de la chromatine, l'histone présentant une modification sur un résidu donné est immunoprécipité avec l'ADN qui lui est associé par un anticorps dirigé spécifiquement contre cette histone modifiée. Les fragments d'ADN ainsi isolés sont ensuite séquencés et replacés sur le génome. L'application de cette stratégie de ChIP-Seq a permis de localiser les histones susceptibles d'être modifiées dans différentes lignées cellulaires (Figure 27 et Figure 28).

All *Cell Line* / *Antibody*	GM12878	H1-hESC	HepG2	HMEC	HSMM	HUVEC	K562	NHEK	NHLF
CTCF	☐	☐	☐	☐	☐	☐	☐	☐	☐
H3K4me1	☐	☐		☐	☐	☐	☐	☐	☐
H3K4me2	☐	☐	☐	☐	☐	☐	☐	☐	☐
H3K4me3	☐	☐	☐	☐	☐	☐	☐	☐	☐
H3K9ac	☐	☐	☐	☐	☐	☐	☐	☐	☐
H3K9me1						☐	☐	☐	
H3K27ac	☐		☐	☐	☐	☐	☐	☐	☐
H3K27me3	☐	☐		☐	☐	☐	☐	☐	☐
H3K36me3	☐	☐	☐	☐	☐	☐	☐	☐	☐
H4K20me1	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure 27: Modifications d'histones dans diverses lignées cellulaires mappées par le projet EncODE

Cette cartographie a été réalisé par le Broad institute (MIT et Harvard). Les expériences listées sont celles disponibles en juin 2010 (les carrés gris signifient que le couple modification-lignée fait l'objet d'une cartographie)



All	Factor	CTCF	H3K4me3	H3K27me3	H3K36me3
Cell Line					
BJ		☐	☐	☐	☐
Caco-2		☐	☐	☐	☐
GM06990		☐	☐	☐	☐
GM12801		☐			
GM12864		☐			
GM12865		☐			
GM12872		☐			
GM12873		☐			
GM12874		☐			
GM12875		☐			
GM12878		☐	☐	☐	☐
HEK293		☐			
HeLa-S3		☐	☐	☐	☐
HepG2		☐	☐	☐	☐
HL-60		☐	☐		
HMEC		☐		☐	
HRE		☐	☐	☐	☐
HUVEC		☐	☐	☐	☐
K562		☐	☐	☐	☐
NHEK		☐	☐	☐	☐
SAEC		☐	☐	☐	☐
SK-N-SH RA		☐	☐	☐	☐
WERI-Rb-1		☐			

Figure 28: Modifications d'histones dans diverses lignées cellulaires mappées par le projet EncODE

Ce mapping a été réalisé à l'université de Washington. Les expériences listées sont celles disponibles en date de juin 2010 (les carrés gris signifient que le couple modification-lignée fait l'objet d'une cartographie)

1.3.2. Mapping de la polymérase II et des facteurs de transcription

Pour obtenir cette information, dans le cadre du projet EncODE, les fragments d'ADN immunoprécipités par un anticorps dirigé contre la Pol II ont été analysés par *deep-sequencing*. La technique de ChIP-Seq permettant indirectement de quantifier l'abondance d'une protéine sur une région donnée, via le nombre de fragments constituant les pics, certaines régions, comme en particulier les TSS, se voient surreprésentées. Il semblerait donc que la phase d'initiation longtemps considérée comme sujette à régulation soit plus attentiste que pensée initialement. En effet, la polymérase se distribue en moyenne sur un grand nombre de promoteurs et semble « attendre » le signal de départ via d'autres signaux. Cette présence fréquente de la polymérase sur les promoteurs a été précédemment confirmée par la découverte de nombreux transcrits cryptiques correspondant à ces régions (Wyers *et al.*, 2005). D'autres

régions sont quant à elles étonnamment dépourvues de polymérase II. Aussi, la présence, l'absence de la polymérase II sur les promoteurs ou la distribution de cette polymérase sur les gènes constitue un critère important du contexte génomique et génique comme suggéré au chapitre 2, paragraphe 3.3. Ces données, croisées avec les données existantes sur la Pol III, ont permis de montrer que dans certains cas, ces polymérases travaillaient de concert, l'une favorisant probablement le recrutement de l'autre (Listerman *et al.*, 2007). Des travaux récents ont reporté la cartographie complète des sites de fixation de l'ARN Pol III au niveau du génome humain (Moqtaderi *et al.*, 2010). Au sein du projet EncODE, la cartographie de la polymérase II fait partie d'un lot plus général portant sur les "Facteurs de transcription".

Les centres d'HudsonAlpha (http://www.hudsonalpha.org) et de Yale (http://encode.gersteinlab.org/) sont en charge de la cartographie des sites de fixation de facteur de transcription (Figure 29). A l'heure de l'écriture, sur les 2000 facteurs potentiels, 91 ont fait l'objet d'une étude à l'échelle du génome par ChIP-Seq dans le cadre du projet EncODE. Certains facteurs restent encore à déterminer et, à l'opposé, certains sites prédits sont encore orphelins, comme par exemple les deux autres motifs co-régulateurs de l'expression du gène SCARNA2 mis en évidence par Gérard et collaborateurs en 2010 et évoqués au chapitre 1, paragraphe 3.3).

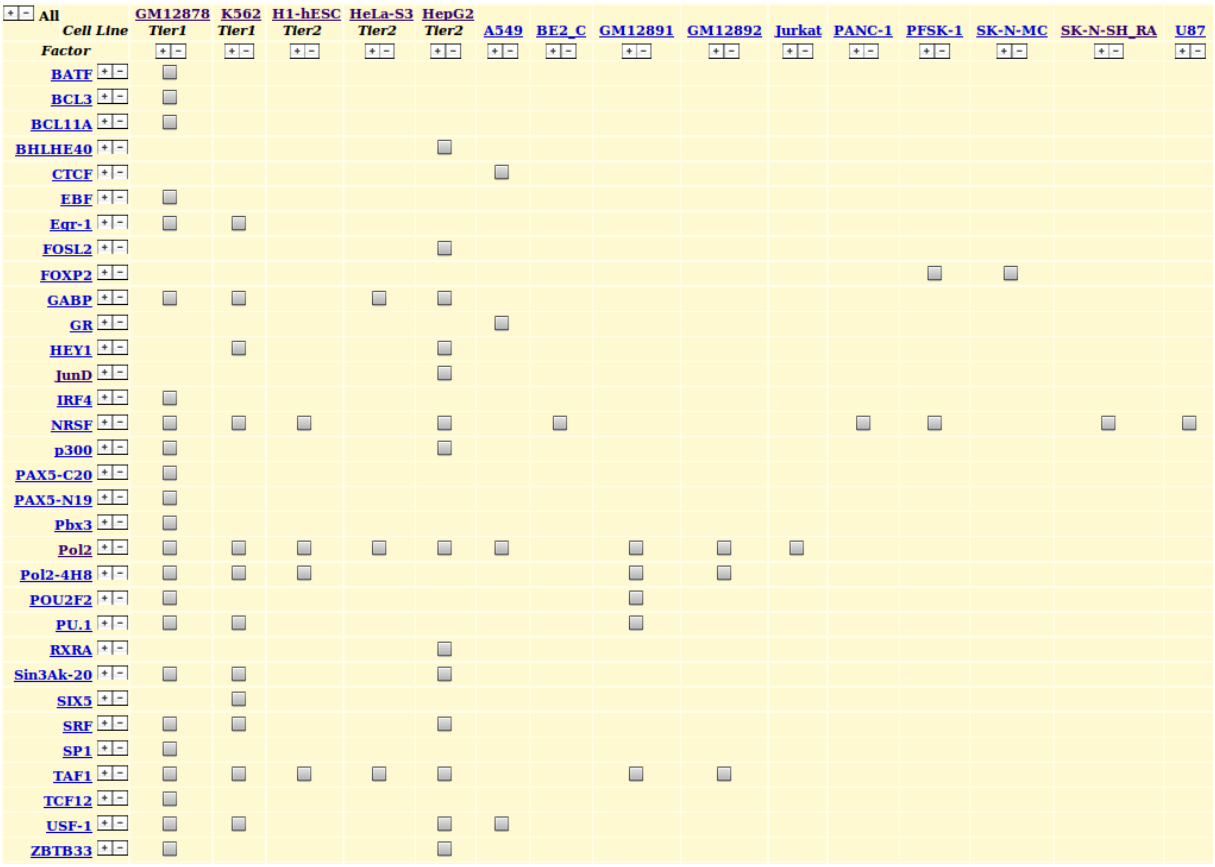


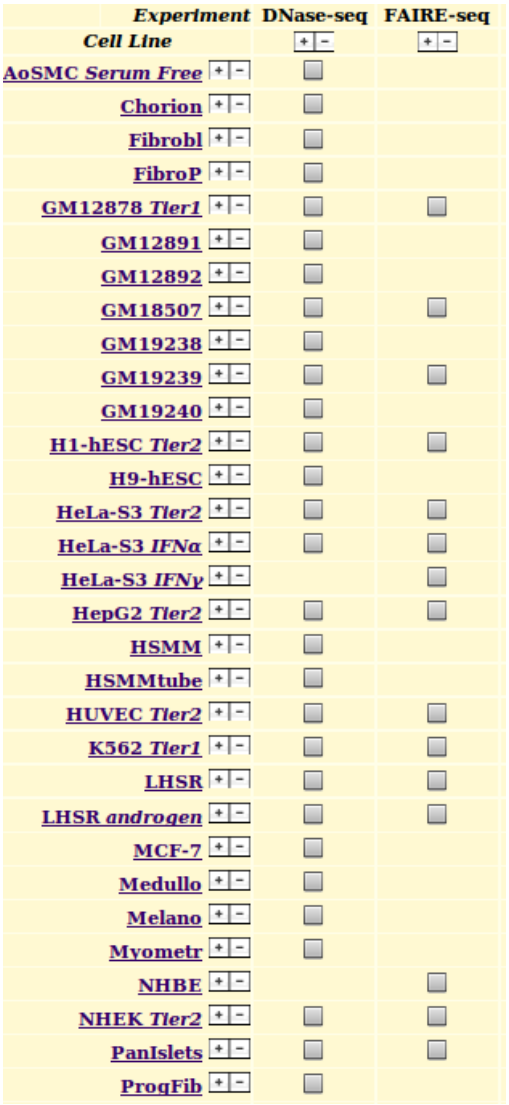
Figure 29 : Sites de fixation de « facteurs de transcription » dans diverses lignées cellulaires mappées par le projet EncODE

Ces cartographies ChIP-Seq a été réalisées à l'université d'Hudson. Les expériences listées sont celles disponibles en date de juin 2010 (les carrés gris signifient que le couple facteur-lignée fait l'objet d'une expérience).

De manière plus anecdotique, un dépôt supplémentaire a été réalisé par Elnitski au sein du projet EncODE. Celui-ci contient la cartographie des éléments régulant négativement l'expression des gènes dans la lignée K562.

1.3.3. Mapping de l'euchromatine

Se basant sur les techniques de DNAse-Seq et FAIRE-Seq, la cartographie de l'euchromatine n'est pas redondante avec le précédent car l'accessibilité du génome est indépendante du type de polymérase et certaines régions d'euchromatine sans ARN polymérase II pourront probablement faire apparaître de nouveaux ARN non-codants produits, par exemple, par l'ARN Pol III. (Figure 30). Cette cartographie présente donc un complément additionnel aux modifications d'histones (dont le catalogue est loin d'être complet) et aux cartographies des ARN polymérases II et III à l'échelle du génome (Moqtaderi *et al.*, 2010; Oler *et al.*, 2010).



Experiment / Cell Line	DNase-seq	FAIRE-seq
AoSMC *Serum Free*	☐	
Chorion	☐	
Fibrobl	☐	
FibroP	☐	
GM12878 *Tier1*	☐	☐
GM12891	☐	
GM12892	☐	
GM18507	☐	☐
GM19238	☐	
GM19239	☐	☐
GM19240	☐	
H1-hESC *Tier2*	☐	☐
H9-hESC	☐	
HeLa-S3 *Tier2*	☐	☐
HeLa-S3 *IFNα*	☐	☐
HeLa-S3 *IFNγ*		☐
HepG2 *Tier2*	☐	☐
HSMM	☐	
HSMMtube	☐	
HUVEC *Tier2*	☐	☐
K562 *Tier1*	☐	☐
LHSR	☐	☐
LHSR *androgen*	☐	☐
MCF-7	☐	
Medullo	☐	
Melano	☐	
Myometr	☐	
NHBE		☐
NHEK *Tier2*	☐	☐
PanIslets	☐	☐
ProgFib	☐	

Figure 30 : Cartographie des zones d'euchromatine dans diverses lignées cellulaires par le projet EncODE

Ces expériences de DNAse-Seq et FAIRE-Seq ont réalisées par l'université , l'université de Caroline du Nord, L'université du Texas, l'Institut Européen de Bioinformatique et l'université de Cambridge. Les expériences listées sont celles disponibles en date de juin 2010 (les carrés gris signifient que le couple facteur-lignée fait l'objet d'une expérience).

1.3.4. Transcriptomes

La technique de RNA-Seq fut utilisée par l'institut Caltech, Le Genome Institut de Singapour et celui de Cold Spring Harbor pour déterminer le transcriptome de diverses lignées cellulaires. Les études, par l'emploi des techniques de RNA-Seq « longs » (Figure 31A et C) ou « courts » (Figure 31B), mono ou double-brin (Figure 31C) se focalisèrent sur la cellule entière ou certains de ses compartiments et organelles comme le cytosol, les polysomes, le noyau, le

nucléoplasme, le nucléole ou la chromatine elle-même (Figure 31A et B). De plus, une contre-étude sur les compartiments cellulaires et organelles fut également menée à Singapour en utilisant cette fois la technique de *paired-end ditags* (PET) (Figure 31D).

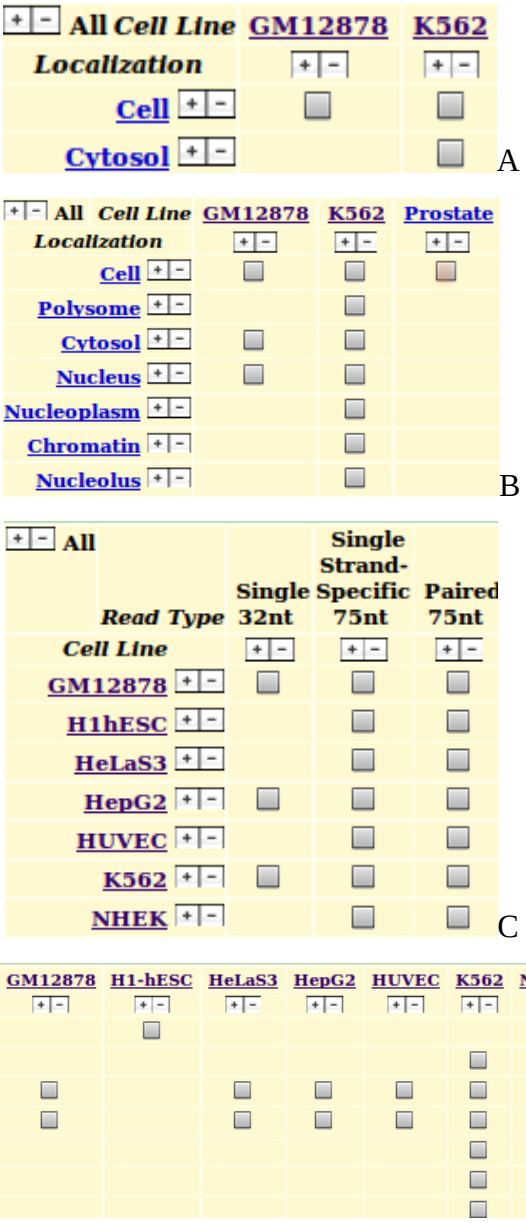


Figure 31: Expériences ciblant l'ARN menées au sein du projet EncODE

Les carrés gris signifient que le couple compartiment-lignée fait l'objet d'une expérience par les laboratoires de Caltech [A à C] et du Genome Institute of Singapour [D]. Les expériences listées sont celles disponibles en date de juin 2010.

1.4. Disponibilité et visualisation des données du projet EncODE

Du fait du choix du format des données, le Genome Browser de l'Université de Californie Santa-Cruz (UCSC) fut retenu pour proposer à la fois la visualisation et le téléchargement du

projet EncODE (Rosenbloom *et al.*, 2010), en remplacement de l'interface limitée EncODEdb développée en début de projet (Elnitski *et al.*, 2007) et abandonnée depuis. Ce faisant, le projet EncODE s'appuie sur le savoir-faire de l'UCSC en termes de bases de données relationnelles, permettant de visualiser une grande quantité d'informations en temps réel. Les fichiers issus du *deep-sequencing* qui y sont entreposés sont répartis par expérience et propoent le plus souvent deux réplicas. Une distinction est réalisée également entre les données brutes et celles traitées statistiquement. Une politique de sécurité des données est de rigueur, le téléchargeur s'engageant à ne pas publier de résultats issus de ces données dans les 9 mois qui suivent leur mise à disposition publique.

1.5. Autres données contextuelles disponibles à l'UCSC

L'UCSC fournit également ses propres données d'annotation contextuelle. Ainsi, on y trouve des données sur la localisation des îlots CpG, des éléments répétés de 3 types différents (totaux, simples ou « nichés ») et des polymorphismes mononucléotidiques. Par un jeu de choix de données à visualiser ou non, de manière succincte ou dense, l'utilisateur peut virtuellement accéder, d'une manière unitaire, à l'intégralité des données disponibles sur une région chromosomique donnée.

2. Dogmes et axiomes à l'aune du projet EncODE

2.1. Exceptions aux modes canoniques d'épissage et de transcription

Par des expérience de RACE (*Rapid Amplification of cDNA Ends*), il a été possible de mettre en évidence des transcrits non-cryptiques et non issus de la fusion de gènes mais générés par *trans-splicing* et donnant donc des ARNm chimères (Figure 32) (EncODE consortium, 2006). Il semblerait par ailleurs que dans bien des cas les deux brins du gène soient transcrits (EncODE consortium, 2006) générant une grande variété d'ARN codant ou non codant. Il en découle donc un rôle bien plus complexe de la transcription, la notion même de gène devenant de plus en plus difficile à définir.

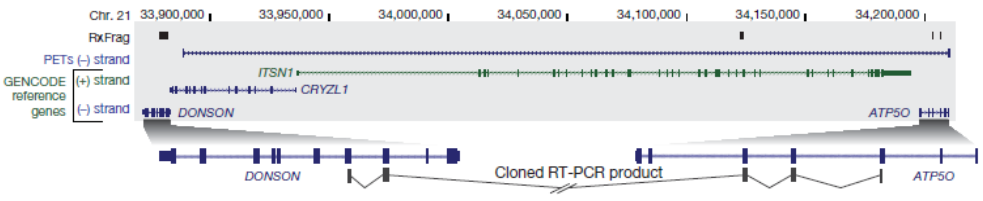


Figure 32 : Expérience de RACE montrant un épissage trans-gène (EncODE, 2006).

2.2. Corrélation entre modifications des histones et accessibilité chromatinienne

Le projet EncODE dans sa phase pîlote permit de valider l'axiome concernant la corrélation entre nucléosome modifié et remodelage chromatinien. En se focalisant sur 1 mégabase, il a été démontré sans équivoque qu'il existait un lien entre la bi-méthylation de la lysine 4 de la queue de l'histone H3 et la sensibilité à la DNase, montrant que cette modification participait à la de-structuration de la chromatine et la rendait plus accessible à l'attaque de la DNase (Figure 33).

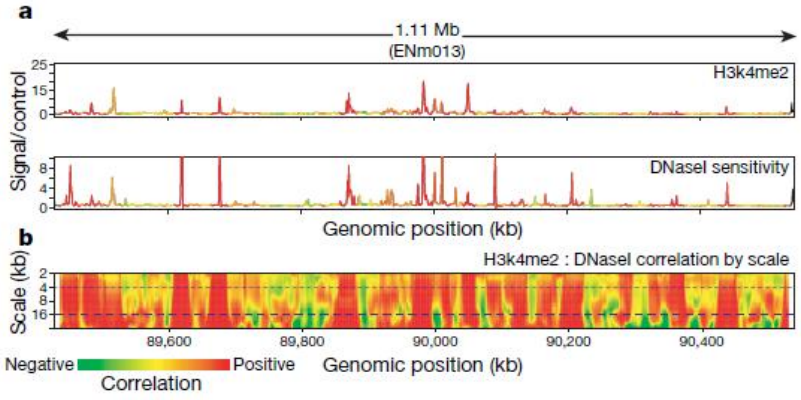


Figure 33 : Corrélation entre la bi-méthylation de la lysine 4 de H3 et la sensibilité à la DNAse

2.3. Abondance des transcrits

La technique du RNA-Seq a mis à jour une quantité insoupçonnée de nouveaux ARN non-codants forçant à une redéfinition des concepts, à partir de leur taille et des endroits du génome où ils co-localisent. Globalement, on parle de petits ARN (ARNs) quand leur taille est inférieure à 200bp et de longs ARN (ARNl) au-delà. Les deux sont associés à un promoteur et constituent respectivement les PASR (pour *Promotor-Associated Small RNA*) et PALR (pour *Promotor-Associated Long RNA*) alors que seuls les premiers peuvent être associés de manière plus précise au TSS (ARNs-TSS). Des équivalents de ces derniers existent au niveau de la fin du gène et sont alors appelés TASR (pour *Terminator Associated Small RNA*). En cas de transcrit situé en amont du promoteur, on parle de PROMPT (pour *PROMoteur uPstream Transcript*), transcrits présents sur l'un ou l'autre brin et dont la fonction concernerait le remodelage de la chromatine. Par ailleurs, chez la levure, on identifia également des transcrits cryptiques instables (CUT) et d'autres, stables mais non-annotés (SUT) et l'ensemble des ARN inconnus est désormais classé sous la bannière de « transcrits à fonction inconnue » (TUF). La

Figure 34 ci-après montre la distribution des nouveaux ARN identifiés par RNA-Seq et par *tiling arrays*.

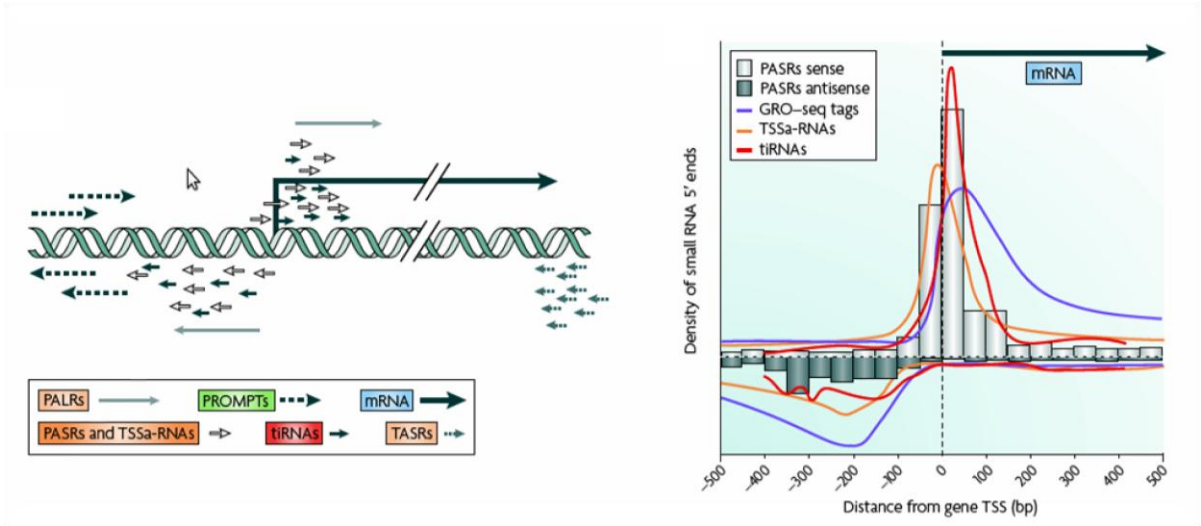


Figure 34 : Distribution des nouveaux petits ARN mis en évidence aux bornes des gènes

Ne nombreux petits transcrits sont associés, en sens ou en antisens au TSS des gènes, d'autres, plus longs sont spécifique su promoteur. Des transcrits sont également associés à la partie 3' des gènes, essentiellement en antisens.

A la manière des ARNmi et des ARNpi, les ARNti constituent également une nouvelle classe de très petits ARN (18 nucléotides) associés à l'initiation de la transcription et mis en évidence dans les cellules humaines, l'embryon de poulet et divers tissus de la drosophile (Taft *et al.*, 2009).

2.4. Abondance des variants de transcrits

Le nombre des variants de transcrits identifiés n'a cessé de croître au fur et à mesure que les techniques pour les détecter s'amélioraient. La multiplication des expériences de RNA-Seq fait état de nombres très hétérogènes de variants compris entre 1400 (Tang *et al.*, 2009) et 22000 (Wang *et al.*, 2008). Un message désormais indiscutable est que ces nombres ne font sens qu'à la lumière d'un type cellulaire, d'un tissu particulier ou d'un stade de développement. A cette fin, les expériences de RNA-Seq et de *microarrays* ont permis de constituer un catalogue des profils d'expression de diverses isoformes (Clark *et al.*, 2007)(Castle *et al.*, 2008), ces variations tissulaires ayant statistiquement plus de poids que les variations interindividuelles (Wang *et al.*, 2008) (Kwan *et al.*, 2008). Le *deep-sequencing* donne accès à des données d'expression complètes pour une variété de lignées cellulaires et le traitement de ces données

devrait permettre de faire émerger des réseaux de régulations. Aussi, les regards sont-ils déjà tournés vers les protéines régulant ces évènements d'épissage alternatifs afin de faire apparaître les sous-programmes impactant les divers sets de gènes (Blencowe, 2006).

2.5. Promoteurs alternatifs

Il apparaît au vu des études récentes (Cooper *et al.*, 2006)(Carninci *et al.*, 2006)(Kimura et la 2006) que la plupart des gènes humains et murins soit associée à plus d'un promoteur, offrant un autre contexte d'initiation de la transcription (passant parfois d'un promoteur large à étroit, dont les définitions opérationnelles ont été présentées à la Figure 11. Ce mécanisme qui peut être tissu-spécifique permet de produire des protéines différentes, selon que le codon initiateur soit touché ou non. Par exemple, l'UDP-glucuronosyltransférase disposerait d'au moins sept promoteurs ayant chacun leur propre profil d'expression tissulaire et l'expression à partir de ces promoteurs génèrent des protéines avec des extrémité amino-terminale différentes (Cooper *et al.*, 2006). Cette création de plusieurs protéines peut même drastiquement changer les propriétés physico-chimiques ou l'adressage de la protéine, à l'image de la gelsoline passant d'une forme cytosolique à une forme plasmatique selon le promoteur utilisé (Carninci *et al.*, 2006).

2.6. Activation trans-génique de la transcription

Il s'agit de cas particuliers de promoteurs alternatifs, conservés dans la région 3' des gènes et plus particulièrement dans les 20 derniers pourcents du dernier exon, renfermant des sites d'initiation de la transcription dans le même sens que le transcrit primaire (Carninci *et al.*, 2005). Chez la souris plus de 1000 gènes présentant une telle architecture ont été identifiés et celle-ci serait suffisante pour initier la transcription d'un gène protéique avoisinant ou un petit ARN non-codant (Carninci *et al.*, 2006).

2.7. Le rôle altruiste des ADN égoïstes

Les études à l'échelle du génome permettent désormais d'appréhender l'impact des fonctions connues des éléments transposables, apportant avec eux parfois des promoteurs tissus-spécifiques et capables de cis-activer des gènes de protéines. De récentes études révèlent que ces fonctions « altruistes » ne constituent pas une exception, mais serait un mécanisme largement répandu du fait d'un fort enrichissement en élément transposables à proximité des

gènes codant pour des protéines et d'une transcription importante des éléments transposables (Faulkner *et al.*, 2009). Ceci conduit à une remise en cause du scénario de quiescence des éléments transposables attendant leur activation. Dès lors, on propose une approche par le contexte, où gène et élément transposable deviennent vecteur de leur contexte informationnels entrant en conflit lors de la transposition (Faulkner and Carninci, 2009).

3. Nouveaux challenges soulevés par les techniques de *deep-sequencing*

L'accès aux données de séquençage via les diverses techniques évoquées ci-dessus a soulevé de nouveaux challenges. Au-delà des challenges techniques inhérents à toute technique émergente et demandant à être perfectionnée, c'est avant tout la diversité biologique qui, comme par le passé devient un facteur limitant.

3.1. Forte dépendance du type cellulaire

Les expériences à l'échelle du génome ne se comprennent qu'à la lumière de la lignée cellulaire dans laquelle elles ont été réalisées. Les empreintes épigénétiques comme la méthylation de l'ADN ou les modifications post-traductionnelles d'histones ne constituent qu'une photo d'un état et d'un type cellulaire et nécessite une animation pour saisir le scénario du film. Sur les 900 lignées cellulaires recensées à l'European Collection of Cell Culture, seules 18 sont représentées dans les données contextuelles d'EncODE. De surcroit, du fait de leur facilité de croissance, il s'agit majoritairement de cultures de cellules cancéreuses présentant parfois de nombreux réarrangements chromosomiques et autres aberrations comme la cellule HeLa-s3 possédant 82 chromosomes. Il convient donc à la fois de manipuler les résultats disponibles avec un certain recul mais aussi de répéter les expériences sur de nombreuses autres lignées et types cellulaires, saines si possible.

3.2. Des données à compléter

Moins de 5% des facteurs de transcription présents dans les cellules ont été utilisés (ou a été utilisé) pour une analyse par ChIP-Seq à l'échelle du génome complet. De même, sur les 300 modifications d'histones seules 13 ont fait l'objet d'une cartographie. Si ceci permet bien évidemment de construire les architectures et plateformes expérimentales et d'analyses, l'exhaustivité thématique semble encore bien loin et, à l'heure de la biologie des systèmes, bien lacunaire pour prétendre assimiler une machine biologique dans son ensemble.

3.3. Nécessité de modifier les conditions d'induction pour couvrir les divers états cellulaires

Un autre point important, mis en évidence par les techniques de RNA-Seq, mais applicable aux autres, est la condition de traitement des cellules. En effet, la cellule étant un système inductible et adaptatif, un autre challenge sera d'étudier les perturbations des systèmes mesurés en réponse à divers stimulus biochimiques.

Les `omics et parmi elles, les expériences menées dans le cadre du projet EncODE ont ouvert les portes de grandes avancées et de remises en question dans l'appréhension du génome forçant les raisonnements sous l'angle du contexte informationnel à se développer. Toutefois, en dépit de la croissance exponentielle des volumes de données, le reste à faire est grand et, dans le cadre d'analyses exploratoires, c'est à dire la majorité, l'incomplétude des données requiert une part importante de prédiction et la construction de modèles pour tracer les grands axes d'investigation. Ceci est particulièrement vrai dans le cadre de l'étude des éléments régulateurs dont l'analyse typique débute avec peu de données sur le facteur correspondant et doit s'appuyer sur les méthodes *in silico* dédiées au questionnement des promoteurs. Au chapitre suivant, nous allons présenter quelques-unes de ces méthodes parmi lesquelles nous avons choisi les briques de notre stratégie dans l'analyse du facteur Staf.

Chapitre 4 : Méthodes bioinformatiques dédiées à l'analyse des séquences régulatrices

La branche de la bioinformatique traitant de l'analyse des séquences régulatrices est une discipline couvrant des méthodes et des algorithmes dédiés à la prédiction et à la validation de sites de fixations dans l'ADN de protéines régulatrices. Elle dispose de ressources spécifiques et d'outils répartis entre les grands centres généralistes et des portails plus spécialisés.

1. Méthodes de prédiction

1.1. Méthodes basées sur un modèle

Ces méthodes permettent d'utiliser l'information reposant sur un ensemble de séquences connues pour construire un modèle probabiliste avec lequel on va pouvoir explorer d'autres régions. Elles se basent globalement sur deux principes : l'utilisation de réelles probabilités ou d'un simple motif basé sur le code IUPAC représentant de manière binaire les bases pouvant apparaitre ou non à une position donnée. Ces méthodes basées sur un modèle ont grandement été et sont toujours utilisées. Permettant d'encoder une réalité biologique en un modèle mathématique. Elles ne pêchent pas plus que les autres dans leurs prédictions.

1.1.1. Méthode par profile

Comme séquence de biopolymère pouvant se décliner en 4 occurrences (A,T,C,G), un site de fixation est intrinsèquement défini par une matrice de longueur égale à celle du site et de largeur 4. Aussi, les méthodes par profile recensent dans un premier temps, à partir d'un ensemble de séquences connues, les probabilités d'occurrence de chaque base à chaque position du site. Ces positions sont alors converties en log-vraisemblance par rapport à un arrière-plan, typiquement le génome ou le promotome (ensemble des séquences promotrices du génome) . Les algorithmes sont en général couplés à une base de données de matrices et de sites de fixation de facteurs. En effet, si l'utilisateur ne sait pas ce qu'il cherche, un ensemble de matrices, souvent réalisées sur peu de sites lui est proposé.

Si cette méthode est séduisante du fait de sa simplicité et des ses fondements théoriques, elle part du point de vue qu'aucune corrélation n'existe entre les positions, ce qui peut parfois

s'avérer inexact (Bulyk *et al.,* 2002; Stormo and Tan, 2002). De plus, ce genre de méthodes prédit beaucoup de faux-positifs, quand bien même les algorithmes proposent un seuil optimal censé les limiter. Le principe du seuil optimal diffère d'un programme à l'autre. Dans le cas de MatInspector (Quandt *et al.,* 1995) ou Match (Kel *et al.*, 2003), il s'agit d'une transformation en percentile à partir de l'agrégation des scores unitaires alors que pour PATSER (ural.wustl.edu/software.html), il s 'agit d'une *p*-valeur basée sur les valeurs extrêmes. De plus, s'agissant d'un log-ratio, des correctifs à l'effet parfois drastiques sont apportés pour les petits échantillons (nombre de séquences disponibles pour construire le modèle). Ces correctifs sont indépendants du facteur. (Nishida *et al.,* 2009). Le taux de faux positif est non-seulement dû à l'extrême variabilité des séquences mais aussi, au fait que ces méthodes ne se basent que sur la séquence-cible donc n'intègrent aucune donnée biologique ou simplement de séquence flanquante. De fait, ces méthodes sont pauvres d'un point de vue intégratif. A titre d'exemple, en utilisant les matrices des seuls vertébrés lors d'une analyse de séquence de promoteur classique, les outils prédiront en moyenne 2 sites par base de la séquence analysée. Des méthodes de filtrage complémentaires s'avèrent donc nécessaire pour affiner ces prédictions.

1.1.2. Méthodes basées sur des modèles de Markov cachés.

Les modèles de Markov cachés apportent une dimension supérieure comparée au profil. De manière simple, le biopolymère d'ADN est assimilé à une séquence de variables aléatoires liées, entre lesquelles il existe des relations cachées. Ces relations peuvent intervenir au sein du site mais également en dehors du site. Ainsi, un modèle de Markov inclut dans l'analyse la région 5' ou 3' entourant le site et un modèle supplémentaire caractérisera quant à lui le passage cryptique du site à ces modèles flanquants. Ceci fut notamment utilisé pour détecter les gènes complets qui possèdent une composition et des séquences particulières au regard de leurs intergènes (Krogh and Mitchison, 1995). Le plus gros problème des logiciels basés sur ces modèles est leur temps d'exécution (augmentant exponentiellement en fonction de l'ordre n du modèle) c'est-à-dire en son aptitude à décrire des transitions cachées à n résidus d'intervalle. Il en découle que pour mener des analyses dans un temps acceptable, les modèles construits sont généralement d'ordre faible. Pour exemple, le logiciel BSSHMM3 (Xu *et al.,* 2005) dédié à l'analyse de séquences régulatrices innovait récemment en réussissant à proposer des modèles de Markov d'ordre 3, lui assurant un gain discutable de 12% en sensibilité et 13% en spécificité comparé à Match,

De plus, une certaine imprécision apparait si le modèle du site de fixation s'avère proche de celui de la séquence qui l'entoure (arrière-plan), comme pour les facteurs Sp1 ou Staf dont les sites sont tous deux riches en dinucléotides GC et retrouvés dans des zones qui le sont également. Des efforts ont été récemment faits pour augmenter la qualité des modèles de ces séquences d'arrière plan (Kyeong *et al.,* 2006) car en général, on utilise deux modèles définis il y a près de 10 ans. (Liu *et al.*, 2001; Thijs *et al.*, 2001).

1.2. Dépendance inter- et intra-sites

De nombreux facteurs nécessitent selon les cas, la liaison ou l'absence de liaison d'autres facteurs à leur propre site pour se fixer (Hochschild and Ptashne, 1986; Lomvardas and Thanos, 2001). Aussi, certaines approches récentes visant à modéliser de telles dépendances au contexte ont-elles été utilisées pour inférer la présence d'un site de fixation (Das *et al.*, 2004) (Wang *et al.*, 2005). Par ailleurs, les dépendances entre les positions des sites de fixation peuvent être examinées à l'aide de logiciels reposant sur des algorithmes de type chaine de Markov – MonteCarlo (MCMC) comme GMMPS (Zhou and Liu, 2004) qui typiquement vont chercher à modéliser les dépendances et offriront de s'en servir comme critère de recherche en couplant les inférences bayésiennes au Gibbs sampling (voir ci-dessous) comme OpenBUGS (http://www.openbugs.info/w/).

1.3. Méthodes de prédiction *ab initio*

1.3.1. Gibbs sampling

Le *Gibbs sampling* (German et German 1984) a été utilisé en biologie pour optimiser les alignements de séquence (Lawrence *et al.*, 1993). On peut donc voir dans la recherche de motifs communs à un groupe de séquences, un mini-alignement au niveau des motifs. Classiquement, l'algorithme démarre de manière aléatoire, aligne et calcule les paramètres de matrices de motif et les fréquences de l'arrière plan. A partir de cette génération aléatoire, il va, de manière itérative, retirer des séquences et, pour chaque séquence écartée, déterminer un score de positionnement en la déplaçant base-à-base au début de la partie alignée ce qui permet à la séquence d'être réalignée et aux paramètres d'être recalculés avant qu'une autre soit écartée. Et ainsi de suite jusqu'à convergence.

1.3.2. Algorithmes espérance-maximisation

Les méthodes de prédiction *ab initio* reposent sur des modèles mathématiques à même de détecter les sites de fixation sans les connaitre *a priori*. Parmi les logiciel reposant sur ces méthodes, le programme MEME (Bailey and Elkan, 1994) est sans doute un des plus utilisés. Il repose sur un modèle de mélange à deux composantes, l'une en charge des sites, l'autre de l'arrière-plan. L'algorithme lié est de type espérance maximisation (Dempster *et al.*, 1977) tentant de maximiser la vraisemblance des paramètres du modèle. Dans un premier temps, le programme MEME évalue la fréquence des motifs de longueur w et détermine au passage un seuil de classification bayésienne, définit les modèles et tente d'en minimiser les paramètres. Il s'agit d'un algorithme de descente de gradient visant par sauts proportionnels à minimiser la fonction à optimiser en se déplaçant dans le sens inverse de celui des variations de cette fonction. Cette méthode trouvera des minima locaux dont elle pourra avoir du mal à s'échapper et l'algorithme pourrait converger à tort ou très lentement. Pour palier ce problème, une part d'heuristique est introduite pour se défaire de ces minimas, ce qui est une des deux différences majeures avec les algorithmes de recherche d'alignements générés par *Gibbs sampling* pour qui ces "sauts échappatoires" se font de manière aléatoire pouvant faire perdre un nombre imprévisible d'itérations en cas de plateau. Le second est l'absence de besoin de classement préalable des séquences par un biologiste, ce qui le rend plus aisé. Au final, la seule réelle contrainte est l'obligation de fournir soi-même la longueur w, l'algorithme ne pouvant comparer deux motifs de longueurs différentes. Depuis sa première version, MEME a soigné sa connectivité et permet désormais de coupler ses recherches avec des accès aux bases de données pour relier les motifs trouvés à de réels facteurs via TOMTOM (Gupta *et al.*, 2007) ou d'utiliser les motifs qu'il a mis en évidence pour scanner d'autres séquences via MAST (Bailey and Gribskov, 1998).

1.4. Méthodes de prédiction basées sur la biophysique

Devant l'absence de réelles avancées théoriques et les résultats discutables des techniques existantes, un engouement a été observé récemment pour remettre au gout du jour les méthodes de prédictions basées sur l'affinité des protéines régulatrices pour leur site de fixation. Certaines de ces méthodes partent de la structure de la protéine et notamment, de son domaine de liaison à l'ADN du fait du lien direct entre la structure de ce domaine et l'affinité pour le site de fixation dans l'ADN. Ce domaine va par ailleurs pouvoir être spécifique d'une famille de facteurs qui aura, du fait de mêmes contraintes biophysiques, des comportements similaires.

Ainsi, par inférence, l'affinité du facteur, la dégénérescence non-délétère des sites de fixation et leur place dans un réseau de régulation pourront être propagées. aux autres facteurs de la famille. D'autres méthodes proposent un remplacement pur et simple du seuil optimal des matrices par un seuil d'affinité physique (Lusk and Eisen, 2008) ou bien tentent de définir des modèles intégrant de nombreuses sources, en plus de données physiques, comme l'expression génique et la séquence d'ADN (Pan *et al.,* 2008). Enfin, certaines méthodes biophysiques se basent sur la stabilité du duplex d'ADN, partant du postulat contre-intuitif que l'énergie nécessaire à la dissociation de la double-hélice serait plus grande aux endroits contenant des éléments régulateurs. Ceci s'expliquerait par la nécessité pour certains facteurs d'interagir avec les deux brins structurés en double hélice (Gordan and Hartemink, 2008).

1.5. Méthodes de prédictions indirectes

Celles-ci reposent en général sur une analyse de gènes co-régulés. Ainsi, une étude systématique alignant tous les promoteurs de l'homme et de la souris a permis de mettre en évidence des régions conservées qui servirent de base pour une interrogation des banques Jaspar (Portales-Casamar *et al.,* 2010) et Transfac (Wingender *et al.,* 1996) et permirent ainsi d'attribuer un ensemble de sites de fixation connus de facteurs de transcription à chaque gène. Stockés dans une base de données, ces gènes permettent d'examiner des ensemble d'autres gènes co-régulés fournis par l'utilisateur et de déceler d'éventuels régulateurs (Chang *et al.,* 2006). De manière similaire, le logiciel Asap (Marstrand *et al.,* 2008) permet d'interroger de multiples gènes co-régulés par de nombreuses matrices simultanément et d'en retirer des candidats enrichis en se basant sur des lois binomiales et des tests de Fisher exact. Les résultats de cette méthode ne présentent pas de gain en sensibilité mais en vitesse, elle est donc idéale pour une première analyse brute. Enfin, des méthodes basées sur les interactions de facteurs seules ont non-seulement permis d'assigner un rôle tissulaire aux facteurs non-tissu-spécifique (Yu *et al.,* 2006), mais ont également ouvert la voie des analyses d'étude combinatoire de plusieurs facteurs.

1.6. Méthodes de représentation

1.6.1. Consensus

Il illustre les combinaisons possibles des bases observées à une position, sans se soucier de leur fréquence d'apparition. Ainsi, posons les deux séquences suivantes : TGCTTTAGA et CGCTTTACA. On note que la position 1 peut prendre les modalités T et G et que la position 8

les modalités C ou G. Ainsi, le consensus sera (T/G) CTTTA(C/G)A. Ce consensus ne changera pas si la séquence 1 a été observée dix fois plus que la séquence 2. Du fait de la grande flexibilité des séquences, des tri-modalités comme C/A/G apparaissent souvent et rendent complexe la lecture de tels motifs. Aussi le code de l'International Union of Pure and Applied Chemistry (IUPAC) a été mis en place afin de ne garder qu'une lettre à chaque position (Cornish-Bowden, 1985). Très récemment, une évolution du code IUPAC à été proposée (Johnson, 2010) pour faire état simplement de la prévalence des possibles bases à une position donnée. Si cette évolution s'avère assez aisée d'utilisation avec deux bases possibles, elle devient hiéroglyphique en passant à trois (Figure 35).

	Translation					
	Code	1°		2°		3°
C–G–T/U combinations (Not A)	B	C	>	G	>	T/U
	B	C	>	G	=	T/U
	B	C	=	G	=	T/U
	b	T/U	>	G	>	C
	b	T/U	=	G	>	C
	L	G	>	T/U	>	C
	L	G	>	T/U	=	C
	l	C	>	T/U	>	G
	l	C	=	T/U	>	G
	O	T/U	>	C	>	G
	O	T/U	>	C	=	G
	o	G	>	C	>	T/U
	o	G	=	C	>	T/U
A–G–T/U combinations (Not C)	D	A	>	G	>	T/U
	D	A	>	G	=	T/U
	D	A	=	G	=	T/U
	d	T/U	>	G	>	A
	d	T/U	=	G	>	A
	E	G	>	T/U	>	A
	E	G	>	T/U	=	A
	e	A	>	T/U	>	G
	e	A	=	T/U	>	G
	F	T/U	>	A	>	G
	F	T/U	>	A	=	G
	f	G	>	A	>	T/U
	f	G	=	A	>	T/U
A–C–T/U combinations (Not G)	H	A	>	C	>	T/U
	H	A	>	C	=	T/U
	H	A	=	C	=	T/U
	h	T/U	>	C	>	A
	h	T/U	=	C	>	A
	I	C	>	T/U	>	A
	I	C	>	T/U	=	A
	i	A	>	T/U	>	C
	i	A	=	T/U	>	C
	J	T/U	>	A	>	C
	J	T/U	>	A	=	C
	j	C	>	A	>	T/U
	j	C	=	A	>	T/U
A–C–G combinations (Not T/U)	V	A	>	C	>	G
	V	A	>	C	=	G
	V	A	=	C	=	G
	v	G	>	C	>	A
	v	G	=	C	>	A
	X	C	>	G	>	A
	X	C	>	G	=	A
	x	A	>	G	>	C
	x	A	=	G	>	C
	Z	G	>	A	>	C
	Z	G	>	A	=	C
	z	C	>	A	>	G
	z	C	=	A	>	G

Figure 35 : Proposition d'encodage des combinaisons tripartites du nouveau code IUPAC

Par un jeu de caractères gras, soulignés et de majuscules/minuscules, ce nouveau code permettrait de signifier la prépondérance des bases à une position donnée

1.6.2. Histogrammes et web logo

L'histogramme constitue le lien entre le motif et le profile (Figure 36). En effet, son abscisse représente la base ou la combinaison de base prédominante à chaque position alors que son ordonnée représente sa probabilité d'occurrence obtenue en examinant une matrice de profile.

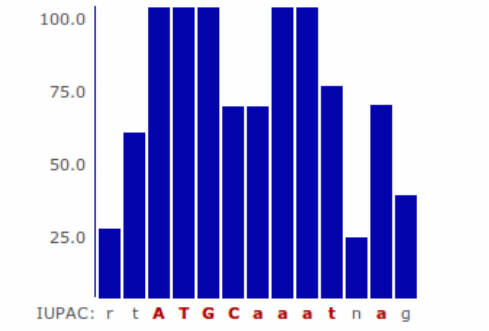


Figure 36 : Représentation en histogramme du profile d'un facteur de transcription par MatDefine.

Les barres de cet histogramme correspondent aux probabilités d'occurrence de la base ou la combinaison de bases la (les) plus représentée(s) à une position donnée.

Le web logo (Crooks *et al.*, 2004) est une représentation où les bases observées à une position sont représentées sous forme d'empilement de lettres dont les tailles différent selon les fréquences d'apparition de chacune des bases. De plus, la taille totale à chaque position est variable et sera le reflet de la conservation globale de cette position (Figure 37). Enfin, cet outil dépasse sa seule condition de visualisateur en permettant d'intégrer des informations sur la composition du contexte des séquences, et des couleurs en fonction des propriétés physicochimiques.

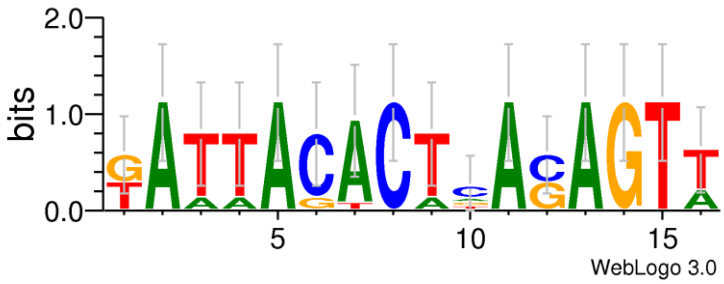


Figure 37 : Représentation en web logo d'une séquence nucléique

Cette représentation figure les probabilités de 4 bases à chaque position, accompagnée d'un écart-type pour la plus représentée.

2. Méthode de validation

La conservation durant l'évolution, qui indique fréquemment le maintien et donc l'importance d'une fonctionnalité, a été le postulat de diverses méthodes de validation de prédictions d'éléments régulateurs. La conservation de séquence entre l'homme et la souris fut le point de départ de nombreuses mises en évidence de régions fonctionnelles (Elnitski *et al.*, 2003) et il est apparu qu'à l'échelle du génome 51% des bases des promoteurs de l'homme, du rat, du chien et de la souris peuvent être alignées, en utilisant une fenêtre de 4000 pb autour du TSS.

Les méthodes de *phylogenetic footprinting* et *phylogenetic shadowing* se basent toutes deux sur la conservation différentielle entre un site de fixation et la séquence « inutile » qui l'entoure, cette dernière n'ayant pas de pression pour demeurer fonctionnelle, mutera donc plus facilement que la séquence du site de fixation au sein duquel une mutation pourrait s'avérer délétère.

2.1. Phylogenetic footprinting.

L'empreinte phylogénétique ou *phylogenetic footprinting* (Zhang and Gerstein, 2003) se base sur la comparaison de l'alignement de séquences non-codantes d'organismes suffisamment éloignés au regard de l'évolution. Par affichage sélectif des seules séquences conservées au moins dans deux organismes, on définit une empreinte correspondant aux régions possiblement fonctionnelles. Toutefois cela présuppose d'une part, de bien connaitre la phylogénie du facteur considéré, pour bien choisir les organismes à intégrer dans l'étude, et d'autre part, de disposer d'alignements de bonne qualité. Dans le cadre des régions promotrices, les exons des gènes souvent conservés permettent d'ancrer les alignements. Toutefois, l'existence de paralogues proches peut biaiser ces alignements. De plus, de micro-erreurs locales d'alignement peuvent se produire, faussant l'analyse base-à-base. Enfin, si les génomes « complets » de vertébrés se multiplient, atteignant aujourd'hui le nombre de 45, leur qualité est très inégale, ce qui limite fortement les études de conservation par alignement de génome. Les problèmes concernent aussi bien la couverture du séquençage (2.2X pour le chat et 7.6X pour le chien) que l'assemblage (absence de chromosome pour le chat, le xénope ou le cobaye).

2.2. Phylogenetic shadowing

Cette méthode (Boffelli *et al.*, 2003) se pose en *anti-footprinting* et se base sur la comparaison de séquences proches, initialement celles des primates. Le postulat cette fois est que ces séquences n'auront pas eu le temps de trop diverger. Ici, l'empreinte correspond désormais aux

nucléotides pour lesquels au moins un organisme présente une substitution. De fait, les régions masquées correspondent aux séquences totalement conservées que sont les exons et les éléments régulateurs. Les divers ordres, et en particulier les primates, commençant à se peupler d'un nombre suffisant de génomes séquencés (homme, macaque, marmoset, chimpanzé, orang-outang pour les primates), cette méthode ne pêche désormais plus que par le manque de données et pourra éventuellement s'avérer utile. Toutefois, les prudences énoncées pour le *phylogenetic footprinting* demeurent de rigueur ici, à ceci près que l'attention à apporter pour le choix des organismes à intégrer, du fait de la distance évolutive, n'est plus de mise.

3. Ressources bioinformatiques dédiées à l'analyse de promoteurs

3.1. Centres généralistes

Un certain nombre de centres généralistes offrent des ressources également dédiées aux promoteurs. Parmi eux, le NCBI ou l'UCSC sont les plus utilisés, principalement par la gamme de résultats d'expériences qu'ils fournissent, notamment des expériences de CAGE (Cap Analysis for Gene Expression), RACE ou de ChIP-Seq (voir chapitre 3, paragraphe 1.2.1). Ils permettent également la visualisation de la conservation de séquences génomiques de plusieurs génomes en se basant sur des alignements *pairwise* via BlastZ (Schwartz *et al.*, 2003) ou multiples via MultiZ (Blanchette *et al.*, 2004). Ceci ouvre de réelles perspectives pour les analyses de génomique comparative portant sur les éléments régulateurs (voir chapitre 2, paragraphe 2.1.) et permet d'afficher ou de récupérer n'importe quelle séquence génomique d'intérêt.

3.2. Bases de données spécialisées

Un certain nombre de bases de données couplées à des portails internet sont dédiées à l'analyse de promoteurs. Parmi elles, la banque DBTSS (http://dbtss.hgc.jp) fournit des annotations de qualité basées uniquement sur des données expérimentales et permet de délimiter, avec précision, le promoteur d'un gène. Sa version 2010 offre désormais des annotations de TSS alternatifs selon le tissu considéré (Yamashita *et al.*, 2010)

Plus précisément encore, certaines entreprises ont fait commerce de la connaissance accumulée sur les sites de fixation de protéines à l'ADN. De complétudes inégales, leurs offrent se basent en partie sur des données expérimentales, principalement par analyse de la littérature. Elles

proposent globalement des solutions avec les mêmes fonctionnalités, à savoir l'analyse d'une séquence par profile en utilisant la connaissance accumulée sur les divers facteurs de la base, et permettent également de définir un modèle exploratoire à partir d'un ensemble de séquences. La plus complète à ce jour, est MatBase de Genomatix (Cartharius *et al.,* 2005), qui fournit des outils dépassant même le cadre des éléments régulateurs. La base concurrente Transfac (Wingender *et al.*, 1996), qui est moins complète dans ses données, présente l'énorme avantage de proposer son architecture en local et de pouvoir ainsi construire sa propre analyse, avec possibilité de ne filtrer que sur le réel expérimental. Enfin, avec sa version 2010, Jaspar s'impose en challenger de poids face à ces entreprises avec ses 457 profils, corrigés et mis à jour avec les nouvelles données issues de Chip-on-chip et de ChIP-Seq. Toutefois, depuis la dernière version, ce ne sont que 12% qui furent mis à jour (Portales-Casamar *et al.*, 2010). Un point intéressant en revanche est la mise en place d'une nouvelle nomenclature pour les sites de fixation notamment une classification basée sur la structure.

Résultats

Chapitre 5 : Enjeux et contraintes

1. Enjeux

Nous avons abordé la mise en place d'une infrastructure à même de questionner de manière massive le contexte génomique humain comme un projet à part entière. Celui-ci avait pour but de fournir une analyse éclairée des données génomiques en les replaçant dans leur contexte informationnel le plus vaste possible et pour moyen l'intégration des données récentes, notamment issues du *deep-sequencing*. Il s'agit donc d'un projet sous contraintes devant à la fois tenir compte des réalités techniques, biologiques et statistiques. Par ailleurs, il s'agissait de proposer une nouvelle philosophie de construction des outils biologiques. Celle-ci repose initialement sur 4 piliers :

i. Disposer, à l'échelle du génome, des données et de statistiques de chaque élément contextuel et pouvoir les extraire rapidement
ii. Etre autorisé à comparer les données d'un utilisateur aux données moyennes observées sur le génome
iii. Offrir une analyse rapide et compréhensible des données
iv. Résumer au maximum les données en donnant accès à leur intégralité au besoin.

Par ailleurs, dans le cadre de notre thématique biologique, cette architecture devait permettre de répondre à la problématique des sites de fixation de hStaf qui a été l'élément moteur tout au long du développement et pour lequel nous savions qu'*in fine* nous disposerions de données expérimentales à haut-débit permettant d'exploiter pleinement notre architecture et, par *feedback*, d'améliorer cette dernière.

2. Contraintes et choix stratégiques

Les 4 points mentionnés plus haut appellent une solution couplant une base de données originale (point i), un moteur Statistique (point ii), un portail internet (point iii) et une collection d'outils graphiques innovants, permettant de résumer un nombre important de données pour en accélérer l'accès par l'utilisateur (point iv).

En considérant les points ii et iii, il apparait nécessaire de pré-calculer un maximum de données, les mesures à l'échelle du génome étant trop gourmandes en temps pour être faites à la volée et respecter le point iii. En croisant les points i et ii, on se rend compte de la nécessité d'utiliser les mêmes protocoles pour les données pré-calculées et les données calculées à la volée pour un utilisateur (voir paragraphe 2.2). Enfin, si l'on essaye de concilier les points iii et iv, la nécessité de représentations graphiques innovantes et orientées web apparait. Celles-ci, de par le point ii devront se baser sur les données statistiques retenues et en permettre la représentation. Ainsi, nous étoffons notre liste de piliers comme suit :

v. Pré-calculer un maximum de données
vi. Utiliser les mêmes protocoles pour les données pré-calculées et les données calculées à la volée pour un utilisateur
vii. Développer de nouveaux moyens de représentation

2.1. Contraintes techniques

2.1.1. Langages de Programmation

L'analyse contextuelle à l'échelle du génome implique de manipuler un nombre massif de données alphanumériques de manière séquentielle. Un langage de script était donc tout indiqué pour exécuter cette tâche. Notre solution se devant d'être utilisable de manière aisée par les non-informaticiens, le langage PHP fut privilégié du fait de sa double utilisation possible comme langage de script pur et pour gérer les applications internet graphiques. Afin de se laisser le choix du système de bases de données, nous avons utilisé PHP 5.2.6 disposant de l'API (*Application Programming Interface*) « PDO » (*Php Data Object*). Comme ses prédécesseurs, cette API permet de s'affranchir du système de gestion de bases de données (SGBD) sous-jacent. Elle présente en outre l'avantage d'être basée sur un code objet natif, bien moins gourmand en temps pour qui souhaite faire un nombre important de requêtes.

2.1.2. Systèmes de gestion de base de données

Nous devions disposer d'une architecture facilement utilisable en web. Pour ce faire, beaucoup sinon la plupart des moteurs de données sont permis. C'est cependant le format des données et les relativement bonnes performances de MySQL qui lui valurent d'être choisi. Les données que nous agrégeons étant pour la plupart disponibles au format MySQL, ceci limite les pertes de temps liées au développement de nouveaux parseurs dédiés à leur intégration. De surcroit, le format MySQL est géré par PDO que nous avons choisi. Conscients toutefois des limites de MySQL face à des volumes très importants de données, nous avons décidé d'adopter MySQL

pour les tables possédant moins d'un million d'enregistrements et le système BIRD (*Biological Integration and Retrieval Data*) pour les volumes supérieurs. Ce moteur développé au laboratoire par Ngoc Hoan Nguyen et largement exploité par la grille de calcul Décrypthon (Bard *et al.*, 2010), est spécialement dédié au stockage, à l'interrogation et à l'analyse de données biologiques hétérogènes. Disposant de son propre métalangage de haut-niveau et de bases de données hiérarchisées orientées-objet, il constitue la solution idéale pour mobiliser de très importants volumes de données. Il permet également de lancer des requêtes à distances via le réseau, ce qui peut s'avérer chronophage lors des accès à la machine. Nous avons donc limité l'utilisation du systèmes BIRD à ce qu'il sait faire de mieux : le traitement des volumes de données massifs.

2.2. Contraintes statistiques

Le but de notre architecture étant de replacer les résultats qu'elle produit dans leur contexte global, il s'avérait important de pouvoir comparer ces résultats aux valeurs standards du génome. Il était donc impératif de « comparer ce qui est comparable », ce qui en matière de statistiques impose certaines contraintes. Dans un souci de rigueur, notre architecture devait se servir de protocoles identiques pour sa construction et son exploitation et que le calcul de statistiques sur les données de l'utilisateur devait être fait en utilisant le même moteur que sur le génome. Concevoir des moteurs capables par les mêmes fonctions de concevoir un site web et de mettre à jour une base de données n'était pas trivial et apporta son lot de contraintes et de soins prodigués tant au niveau des algorithmes (voir Chapitre 6, paragraphe 2) qu'à la structuration des données.

L'application de notre méthode au domaine génomique nous fit écarter l'emploi pourtant classique des moyennes et des écart-types. En effet, les variations très importantes des valeurs de données mesurables, telles que les distances génomiques ou les décomptes d'entités, aboutissent à des valeurs d'écart-type pouvant atteindre deux fois la moyenne, très sensible aux valeurs extrêmes (Tableau 2).

feature	length_average	length_sd	length_q1	length_median	length_q3
exon	298	673	94	133	196
gene	50730	113693	2106	15050	50080
intron	6215	20194	527	1598	4320

Tableau 2 : Valeurs statistiques calculées à partir de la longueur des gènes, des exons et des introns

On voit bien ici que du fait de valeurs extrêmes dans le génome, l'écart-type peut atteindre près de 3 fois la valeur de la moyenne. De plus, ces valeurs extrêmes déportent la moyenne vers elles puisque, en observant les quartiles, on voit que ces moyennes sont plus proches du troisième quartile, ce qui correspond donc à une valeur supérieure aux trois quarts de celles du génome et non la moitié (médiane) comme on l'aurait pensé intuitivement.

Nous avons donc résolu d'adopter le système de percentiles, classant puis découpant les données selon leur rang. Le percentile d'ordre 25 constitue le premier quartile (Q1), celui d'ordre 75 constitue quant à lui le 3ème quartile (Q3). Enfin, le second quartile est aussi appelé médiane (Med). Nous considérons également les percentiles d'ordre 5 et 95 sur lesquels nous reviendrons plus longuement au paragraphe 3 du chapitre 6. Les trois quartiles, la valeur minimale et la maximale sont souvent représentés sous forme d'une "boite à moustaches" comme à la Figure 38.

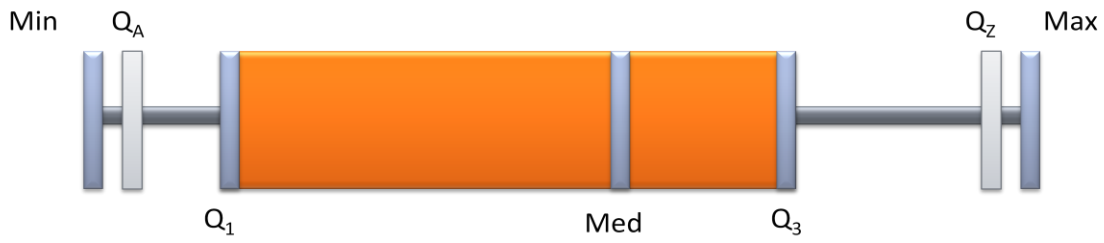


Figure 38 : Représentation en "boite à moustaches" représentant les quartiles particulier et les minimum et maximum décrivant une distribution statistique

La boite orange est bigorné par le premier quartile (Q_1). et par le 3ème quartile(Q_3). et contient la médiane. Les moustaches grises sont limitées par le minimum et le maximum de la distribution. De plus nous avons fait figurer ici les percentiles d'ordre 5 (Q_A) et 95 (Q_Z).

2.3. Contrainte linguistique

Pour notre étude, nous avons défini plusieurs concepts spécifiques, et mis en place des distinctions particulières appelant des termes au sens propre à notre architecture.

Mode : Notre protocole gérant à la fois le pré-calcul de données et un calcul de données à la volée, nous avons distingué ces deux approches en deux modes : le mode « batch » et le mode « interactif » respectivement.

Périmètre : En plus du contexte global à l'échelle du génome, nous avons considéré que le contexte propre aux gènes constituait une approche suffisamment spécifique pour disposer de ses propres statistiques. En ce sens, nous avons distingué deux angles : le premier se référant aux gènes , le second au génome.

Vecteurs informationnels: Il s'agit de caractéristiques biologiques définies par un certain nombre de descripteurs. Il peut s'agir des îlots CpG, des éléments répétés, des modifications d'histone, des zones d'euchromatine ou encore du positionnement de la polymérase II. A ces entités communes aux champs génomiques et géniques, certains vecteurs informationnels propres aux gènes ont également été intégrés, notamment des entités issues de la carte exonique (exon, intron, UTRs, CDS)

Descripteurs CLON : Afin de décrire les vecteurs informationnels, nous avons choisi quatre types de descripteurs. Il s'agit de leur nombre dans une région donnée, rapporté à la taille de ladite région (C, pour *count*), leur longueur (L pour *length*) en terme de nucléotides qu'elles représentent (cis-contexte) ou recouvrent (trans-contexte), leur occupation (O pour *occupancy*) de la région considérée, représentée par leur longueur L rapportée à celle de la région et enfin de leur distance minimale à chaque entité mesurable la plus proche (N, pour *nearest*).

Chapitre 6 : La base de données GeCo

Il s'agit d'une base de données semi-relationnelle contenant à la fois des données "mortes" (intégrées en l'espèce) de qualité, issues des grands centres de référence de chaque spécialité, ainsi que des données "vives", que GeCo génère elle-même pour compléter ou enrichir les données "mortes" avec des informations originales. Notre définition du contexte s'appuie pour l'heure sur les vecteurs informationnels que sont les gènes ainsi que les localisations d'îlots CpG, d'éléments répétés, de SNP, de la polymérase II, des zones d'euchromatine, des histones modifiées et des zones du génome humain conservées au sein de nombreux organismes. Chacun de ces aspects constitue une table majeure de notre base. La structuration de GeCo étant amplement détaillée dans le manuscrit fournit en annexe, nous allons, dans ce chapitre, nous intéresser aux choix "philosophiques" que nous avons fait dans GeCo et à l'utilisation qu'on peut faire de cette plate-forme.

1. Philosophie de structuration : un modèle semi-relationnel

Notre système repose sur une base de données hétérogènes qu'il a fallu structurer. Le besoin était de disposer de données alliant qualité et nouveauté. Pour la qualité, nous nous sommes basés sur les plus grands centres mondiaux ou disposant des annotations les plus complètes possible d'un domaine comme les ARNmi du Sanger Institute ou les piRNA de la piRNAbank (Figure 39). Pour s'assurer de la nouveauté, nous avons mis en place des protocoles automatisés vérifiant si les données dont nous disposons n'ont pas fait l'objet d'une mise-à-jour par une des sources distantes. Dans le cadre du projet EncODE, nous avons également intégré la notion de disponibilité des données du fait du délai légal d'utilisation de ces données dans une optique de publication.

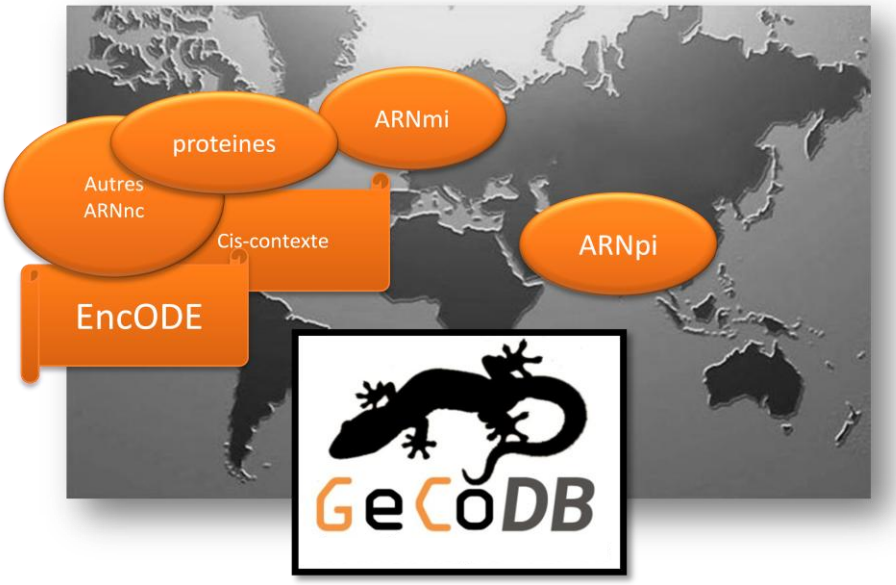


Figure 39 : Sources de données de la base de données GeCo

Les annotations de gènes viennent du NCBI pour les protéines (RefSeq) du Sanger Institute pour les ARNmi, de la piRNADatabank pour les piRNA et de l'UCSC pour les autres ARNnc. Le cis-contexte (SNP, éléments répétés, îlots CpG) vient de l'UCSC. Les données du projet mondial EncODE sont hébergées à l'UCSC.

Par ailleurs, au vu de l'importante diversité des données stockées, nous avons choisi de casser le modèle relationnel global en introduisant une part de redondance. Ceci permet de limiter les jointures de tables qui, si elles sont adaptées à une interrogation courante, peuvent par contre se montrer très pénalisantes pour une utilisation haut-débit de la base de données. Ainsi, lors de la construction des diverses tables, les données les plus fréquemment demandées sont en partie reportées en complément de la seule nécessaire clé-étrangère. En terme de base de données, il n'existe pas de solution toute faite et encore moins pour la gestion de si forts volumes. Ce sont les quantités de données stockées, le nombre de tables et l'articulation entre les données, couplés à leur usage qui doit dicter le bon sens de structuration. Dans notre cas, il s'agit d'un équilibre entre deux pénalités : les jointures externes et le nombre d'attributs décrivant une table. Nous avons limité la première à la portion congrue et minoré la seconde en se contentant d'une centaine d'attributs. Ainsi, en dehors de l'espace mémoire important requis pour stocker de nombreux index, cette seconde pénalité reste acceptable.

Le cœur de l'architecture GeCo étant son annotation de gènes auto-enrichie, une attention particulière a été portée à sa conception. L'idée était d'enrichir les annotations de gènes par les éléments de leur contexte afin d'en disposer rapidement par interrogation simple de la base de données. Ainsi, lors de la mise à jour, les annotations de gènes sont croisées entre elles et avec les éléments de contexte génomique pour que chaque gène de GeCo "connaisse" sont contexte et son contenant en vecteurs informationnels. Nous avons pour cela fait le choix d'introduire une certaine redondance dans nos données. La méthode classique aurait été de placer dans GeCo uniquement les clés pour retrouver les éléments de contexte d'un gène mais les jointures que cela implique auraient été trop pénalisantes au vu du nombre de vecteurs et du volume des données. Aussi, l'utilisation de GeCo par le portail a guidé nos choix et les données contextuelles de chaque gène faisant l'objet d'une utilisation systématique par le portail ont été reportées dans la table d'annotation. Ce faisant, le portail permet de récupérer plusieurs centaines de gènes par minute en mode rapide (affichage résumé).

2. Philosophie algorithmique : le vecteur informationnel au cœur de l'architecture

Afin de respecter la contrainte de réutilisabilité des protocoles, nous nous sommes fortement appuyés sur la modélisation objet. Les objets correspondant aux vecteurs informationnels sont créés par l'algorithme comme des coquilles vides héritant toutes du concept abstrait "vecteur informationnel". Pour remplir ces coquilles, nous avons externalisé le type d'accès SQL que fait chaque objet à ses données. Ainsi, lors de la génération des objets, l'algorithme personnalise chaque vecteur informationnel en lui assignant un type (CpG, polymérase…) lui indiquant où chercher ses données, et un cas de localisation lui expliquant comment le faire. De fait, l'algorithme exploitant un tel système n'est rien d'autre qu'une "couveuse" à objets, les enfantant, leur donnant une identité et les faisant communiquer entre eux. Il n'y avait donc aucune barrière à ce que cet algorithme soit à la fois un script de traitement de données utilisateur et un script de mise à jour de base de données. Il ne restait simplement qu'à doter les objets d'une collection de méthodes leur apprenant à exécuter les tâches relatives à ces finalités d'algorithme. Ce faisant, nous avons obtenu ce que nous décrivons succinctement dans le manuscrit de publication par "moteurs hybrides". La place centrale du vecteur informationnel au cœur de l'architecture GeCo est résumée à la Figure 40.

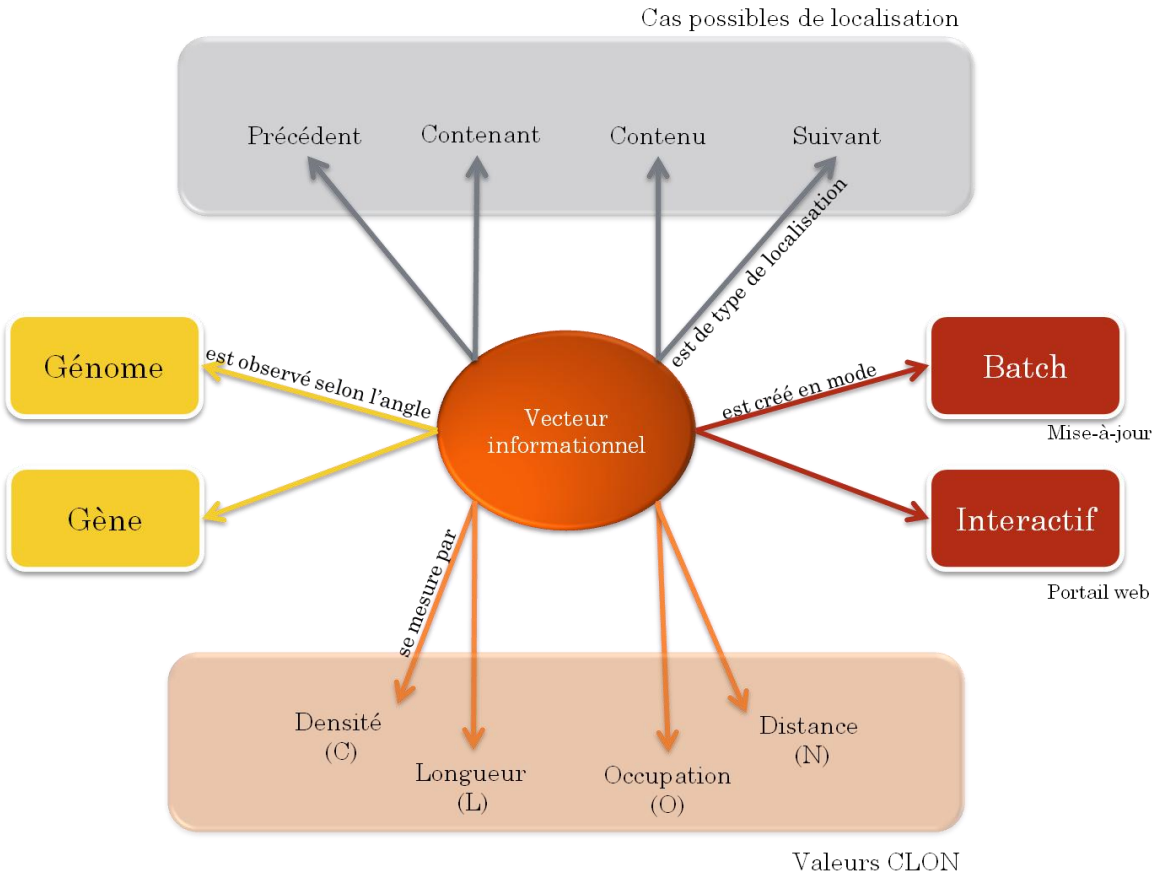


Figure 40 : Le vecteur informationnel au coeur du système GeCo

Le vecteur informationnel est l'objet central à la croisée des nombreuses dimensions qu'offrent GeCo. Qu'importe l'angle sous lequel on le regarde (jaune), le type de programme l'exploitant (rouge), la manière dont on leur donne une identité de localisation (gris) et les descripteurs qu'on attend de lui (saumon), son code est unique et facilite donc les comparaisons entre objets.

En revanche, la présence de deux périmètres a été problématique et a nécessité certains aménagements. En effet, GeCo demeure construite sur une table de gènes qui est centrale. Pour disposer de gènes stockant leur contexte informationnel, tout se passe comme si ces gènes avaient été utilisés comme entrée utilisateur de type coordonnées chromosomiques et que le résultat de la communication des vecteurs informationnels était venu enrichir les annotations des gènes. De fait, interroger GeCo selon l'angle "gène", même à haut-débit ne relève que d'un seul ordre SQL construit à la manière de Légo™ à partir de nombreuses conditions booléennes ("ET", "OU", "NE CONTENANT PAS", etc.). Par opposition, interroger GeCo en utilisant des coordonnées (de région ou de SNP) génère une requête pour chacune des entrées. On voit donc que dans le premier cas, c'est un seul objet "gène" aux multiples cardinalités (enregistrements de la table) qui est créé alors que dans le second, c'est autant d'objets que d'entrées utilisateur,

chacune correspondant à un enregistrement de la table. En intégrant cette dimension de cardinalité aux objets, nous n'utilisons toujours qu'un seul constructeur d'objet ce qui rend la maintenance plus aisée et le code plus léger. En revanche, l'impact sur l'algorithme de gestion de ces objets fut important car dans le premier cas, il s'agit du traitement multiple (n enregistrements) d'un unique produit (1 requête) issu de multiples entrées alors que dans le second il s'agit d'un traitement unitaire (1 enregistrement) de chaque entrée (n requêtes). C'est ici la seule fracture dans l'unicité de notre modèle, certaines méthodes devant être adaptées selon le mode, comme l'affichage que peuvent faire d'eux-mêmes les objets.

Ceci imposa également de passer en objet les statistiques issues de GeCo. Nous aurions pu apprendre à chaque vecteur informationnel à tirer des statistiques de son domaine, mais une contrainte étant qu'une partie chronophage de ces statistiques soit précalculée et stockées, il n'est besoin que d'y accéder une seule fois et non pour chaque objet. Ainsi, GeCo gère aussi son objet statistique, indépendant des autres mais à leur entière disposition pour leur fournir les données stockées ou procéder à des calculs à partir des instances de vecteurs informationnels. Cette manière de faire découple totalement les statistiques de la problématique de cardinalité de l'algorithme, cantonnée aux seules localisations et affichages.

3. Philosophie statistique : détecter les valeurs atypiques

Les statistiques dans GeCo sont un ensemble de concepts et d'application de méthodes visant à pouvoir replacer les résultats d'une analyse dans leur univers informationnel global, de manière compréhensible et rigoureuse.

Après une étape de localisation des objets les uns par rapport aux autres, l'algorithme récolte les descripteurs CLON correspondants. Notre système se basant sur les percentiles, nous déterminons les descripteurs CLON segmentant les distributions en 5%-95% (Q_A) et 95%-5%, (Q_Z), bornes couramment utilisées pour déterminer les intervalles de confiance (IC). Ceci est réalisé sur l'intégralité du génome et des gènes lors de chaque mise à jour de GeCo. Ce calcul est également effectué sur les données de l'utilisateur à la volée. Notre définition d'une valeur atypique est une valeur inférieure à Q_A ou supérieure à Q_Z. Ainsi, pour chaque entrée de l'utilisateur, les descripteurs CLON de chaque vecteur informationnel seront comparés à l'intervalle [Q_A - Q_Z] correspondant à ce vecteur (Figure 41).

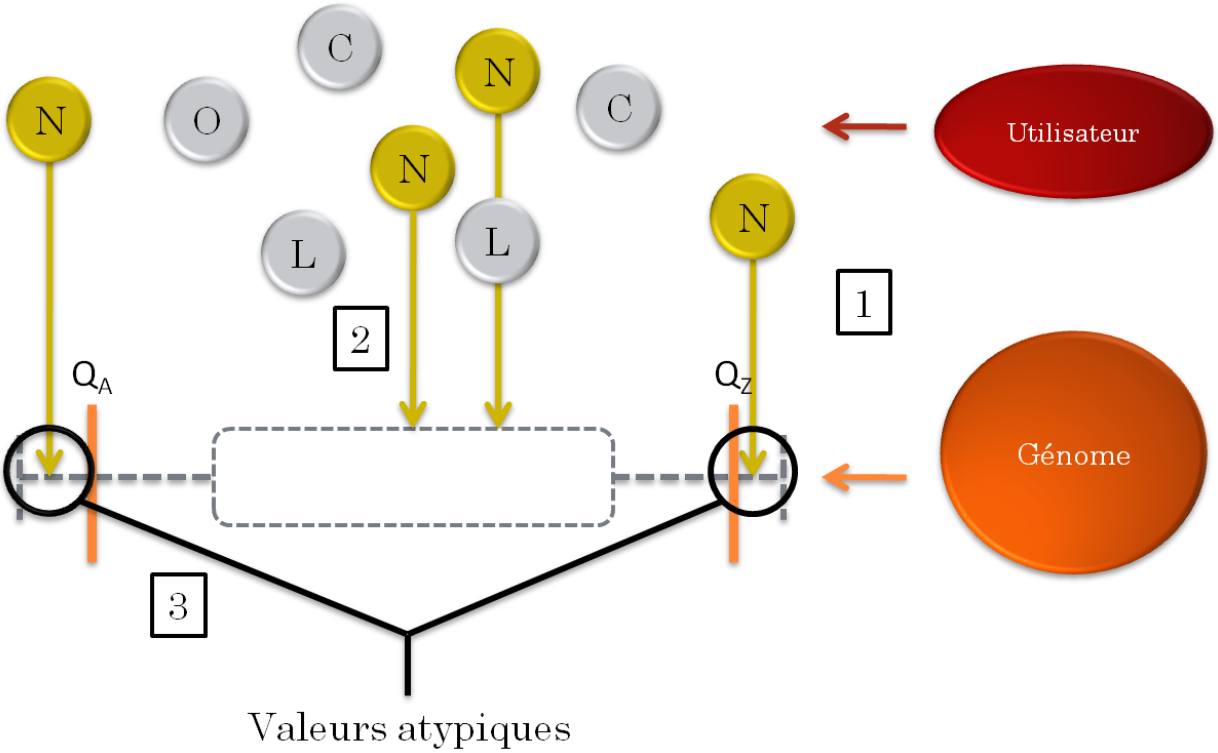


Figure 41 : Identification des valeurs atypiques générées par l'analyse de GeCo pour le descripteur "N" (distance) d'un seul vecteur informationnel

Les descripteurs CLON d'un vecteur informationnel sont générés à la volée à partir des entrées de l'utilisateur tandis que l'intervalle $[Q_A - Q_Z]$ de chaque descripteur CLON est généré pour tous les vecteurs informationnels à partir du génome. (1). Dans le cas présent, les valeurs de distance (N) de l'utilisateur sont comparées à l'intervalle $[Q_A - Q_Z]$ correspondant (2) et les valeurs en dehors de cet intervalle sont identifiées comme atypiques (3).

Une comparaison supplémentaire est également réalisée entre les descripteurs CLON de chaque entrée de l'utilisateur et l'intervalle $[Q_A - Q_Z]$ global déterminé à partir de l'ensemble des entrées de l'utilisateur. Ainsi, pour chaque vecteur informationnel de l'utilisateur, les descripteurs CLON atypiques au regard de 2 comparaisons sont connus. Enfin, la fraction de descripteurs CLON atypiques d'une entrée utilisateur (que nous appelons "divergence score" dans le manuscrit de publication) fournit un indicateur de la divergence de cette entrée par rapport aux valeurs standard.

3.1. Détermination sans *a priori* des lois

Notre intervalle $[Q_A - Q_Z]$ n'est pas à proprement parlé un IC de la moyenne. Ce dernier fait l'objet de lois et de formules précises et ne se base pas directement sur le rang occupé par une

valeur dans l'ensemble ordonné de la population de valeurs. Toutefois, le calcul des bornes d'un IC est intimement lié à la loi statistique sous-tendant les données. A titre d'exemple pour une loi normale, cet intervalle sera :

$$\left]\overline{x}-1,96\frac{\sigma(X)}{\sqrt{n}};\overline{x}+1,96\frac{\sigma(X)}{\sqrt{n}}\right[$$

avec σ(X) écart type, x barre = moyenne et n = effectif.

Pour une loi log-normale en revanche, une transformation de l'IC est nécessaire et le simple passage sous forme exponentielle des bornes de l'IC de la loi normale n'est qu'une approximation naïve encadrant en fait la médiane (e^{μ}) et non la moyenne($e^{\mu+\sigma^2/2}$), ce qui posera problème pour les écart-types grands, comme c'est le cas pour nos descripteurs CLON. De plus, le fait que certaines de nos valeurs d'Occupation (O) soient une proportion (taille du vecteur informationnel divisé par la taille de la région le contenant), cela oblige à calculer un IC de proportion et non plus un IC de la moyenne. Cet intervalle est donné par :

$$IC(p)=\left[\frac{f_n+\frac{u^2}{2n}-\frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n}+f_n(1-f_n)}}{1+\frac{u^2}{n}},\frac{f_n+\frac{u^2}{2n}+\frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n}+f_n(1-f_n)}}{1+\frac{u^2}{n}}\right]$$

simplifié, pour les grands effectifs par :

$$IC(p)=\left[f_n-u\sqrt{\frac{f_n(1-f_n)}{n}}\ ,\ \ f_n+u\sqrt{\frac{f_n(1-f_n)}{n}}\right]$$

Il est donc nécessaire, si l'on veut fournir un IC correct pour chaque descripteur CLON, de connaitre les lois statistiques décrivant les données et d'appliquer les formules d'IC qui en découlent, en prêtant attention à ce que ces formules soient applicables au type et au volume des données.

Afin de tenter de déterminer les lois statistiques sous-tendant la distribution des descripteurs CLON, nous avons développé des routines reposant sur des modèles de mélange. Se basant sur le programme R (http://www.r-project.org) et plus particulièrement le package bbmle, ces moteurs créent des mélanges de deux lois (exponentielle et log-normale, exponentielle et

gamma, exponentielle et normale) comme cela est utilisé dans certaines études sur la distribution des cadres de lecture ouverts dans le génome (McCoy *et al.*, 2009) ou la longueur des protéines (Zhang, 2000). Le moteur essaie de faire correspondre les données CLON aux modèles par maximisation de la vraisemblance. La part de chaque loi dans le modèle est évaluée et le modèle est éventuellement simplifié si l'une ou l'autre est négligeable. *In fine*, les modèles sont comparés par le critère d'Akaike (Akaike, 1974) et le meilleur est choisi. Connaissant les lois qui sous-tendent les données, le calcul de l'intervalle de confiance correspondant devient accessible. Toutefois, si certaines lois sont en parfaite adéquation avec les données (principalement la log-normale (Figure 42)), certains cas sont plus problématiques comme la distribution des longueurs de gènes qui semble bimodale (Figure 43).

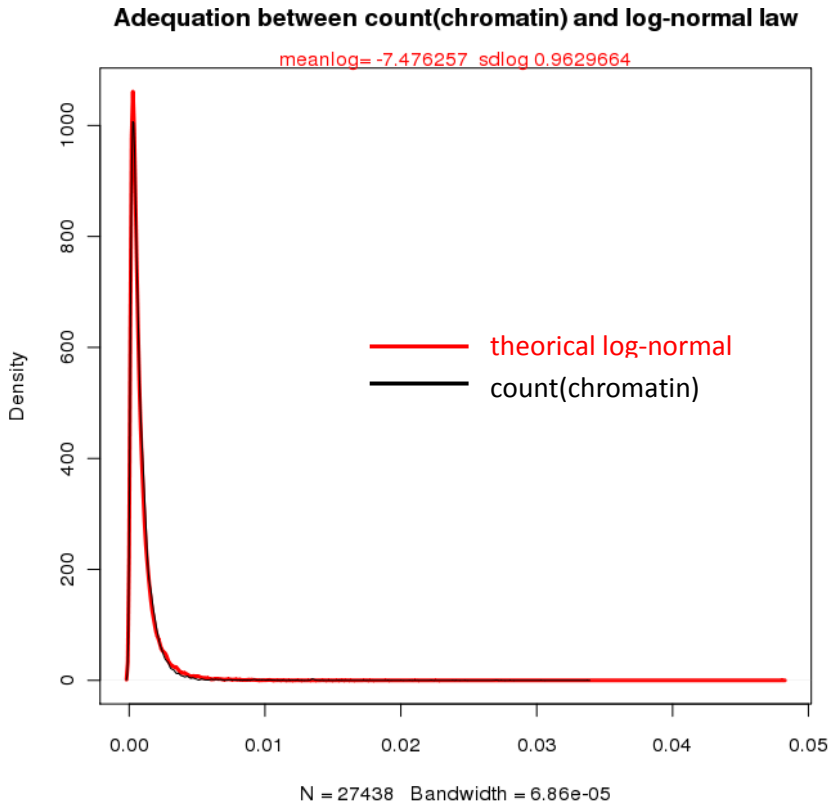


Figure 42 Adéquation entre la distribution du nombre de zones d'euchromatine à l'intérieur des gènes rapportée à la taille des gènes (CLON : C)

La distribution des nombres de zones d'euchromatine (rouge) s'avère en quasi-parfaite adéquation avec la loi log-normale (noire) (Données générées en automatique par application d'un modèle de mélange composé d'une part de loi gamma et d'une part de loi log-normale, raffiné en loi log-normale pure) à l'aide du logiciel R.

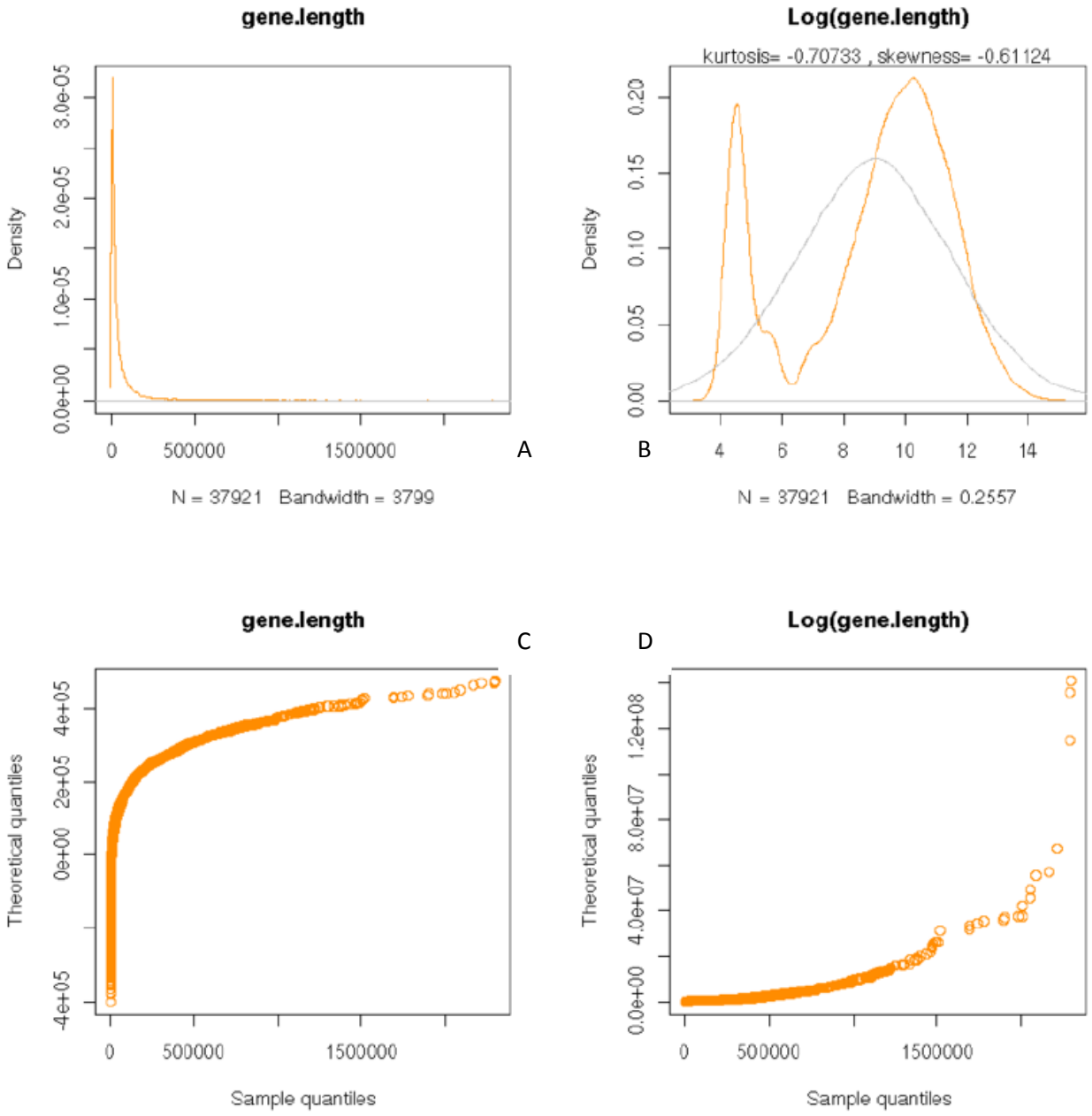


Figure 43 : Exemple de graphiques tracés en automatique par les sous-routines de mise-à-jour de GeCo. Cas de la longueur des gènes.

A et C correspondent aux données brutes alors que pour B et D elles ont été transformées en logarithmes. A et B correspondent à des tracés de densité de probabilité alors C et D sont des représentations en "QQplot" traçant les quantiles attendus pour chaque loi contre ceux observés en réalité, une ligne droite signifiant une adéquation avec la loi. Dans le cas précis de la longueur des gènes, on note une apparente bi-modalité de la densité de probabilité logarithmique (B). Ainsi la distribution semble composée d'un ensemble de deux populations de gènes, courts et longs, suivant chacune une loi log-normale. Ceci se reflète au niveau du QQplot (D) décrivant grossièrement deux droites avant et après 1500000 (1.5 Mpb).

Une minorité de cas, comme les distributions bimodales échappent donc encore à nos modèles et dans quelques autres cas, bien que semblant en adéquation avec les données, les modèles ne permettent pas de conclure à l'existence d'une adéquation car la maximisation de la vraisemblance ne converge pas. Ceci est dû à la sensibilité de la méthode aux paramètres initiaux. Il conviendrait donc de les déterminer un-à-un pour chaque modèle, représentant ainsi plus d'un millier d'optimisations manuelles, ce qui ne pouvait être réalisé dans le temps imparti. Ces routines préliminaires posent donc les premières bases vers une détermination systématique des lois sous-tendant des données génomiques. Ces bases devront être améliorées, notamment en intégrant plus de lois statistiques et en développant des outils automatiques à même de détecter et contrecarrer la non-convergence des fonctions de vraisemblance. En guide de validation préliminaire, nous avons dores et déjà déterminé quelques IC. Par exemple, nous avons déterminé l'IC à 5% encadrant la moyenne du nombre de bases couvertes par les histones H3K27Ac (H3 acétylées sur leur lysine[27]) à [3800;3938] pour un moyenne de 3869.

3.2. Echecs des tests statistiques classiques

Afin de comparer l'ensemble les descripteurs CLON moyens (déterminés à partir du set entier des entrées) de l'utilisateur avec les valeurs observées sur le génome, nous avons cherché à implémenter d'autres méthodes que l'intervalle [Q_A - Q_Z]. L'idée sous-jacente était de comparer la distribution statistique de ces descripteurs à celle du génome. Toutefois, du fait de l'anormalité des données et des très forts effectifs (18.000.000 de SNPs par exemple), de nombreux tests statistiques sont en échec et ont tendance à décréter systématiquement l'adéquation ou l'inadéquation des distributions puisque l'effectif des données entre systématiquement dans la valeur du test. Ainsi, le test t, le test de Wilcoxon ou encore celui reposant sur les coefficients d'aplatissement (*kurtosis)* ou d'asymétrie (*skewness)* n'étaient pas adaptés pour de tels échantillons. En effet, ils déclaraient "atypiques" des valeurs très proches de la moyenne/médiane calculée sur le génome ou, au contraire, "normales" des valeurs au delà de Q_Z. Le moteur statistique propose toutefois de réaliser les deux premiers tests via le portail web, aux risques et périls de l'utilisateur qui souhaiterait les utiliser pour comparer ses valeurs aux standards génomiques.

4. Philosophie graphique: accès facilité et personnalisé à l'information

Le contexte informationnel génomique, pourtant à ses débuts, constitue déjà une masse d'information importante pour appréhender le génome dans son intégralité et accéder

rapidement à la pertinence des entités manipulées. Nous avons à cet effet mis au point de nouveaux modes de représentation.

4.1. Accès granulaire aux données et carte exonique

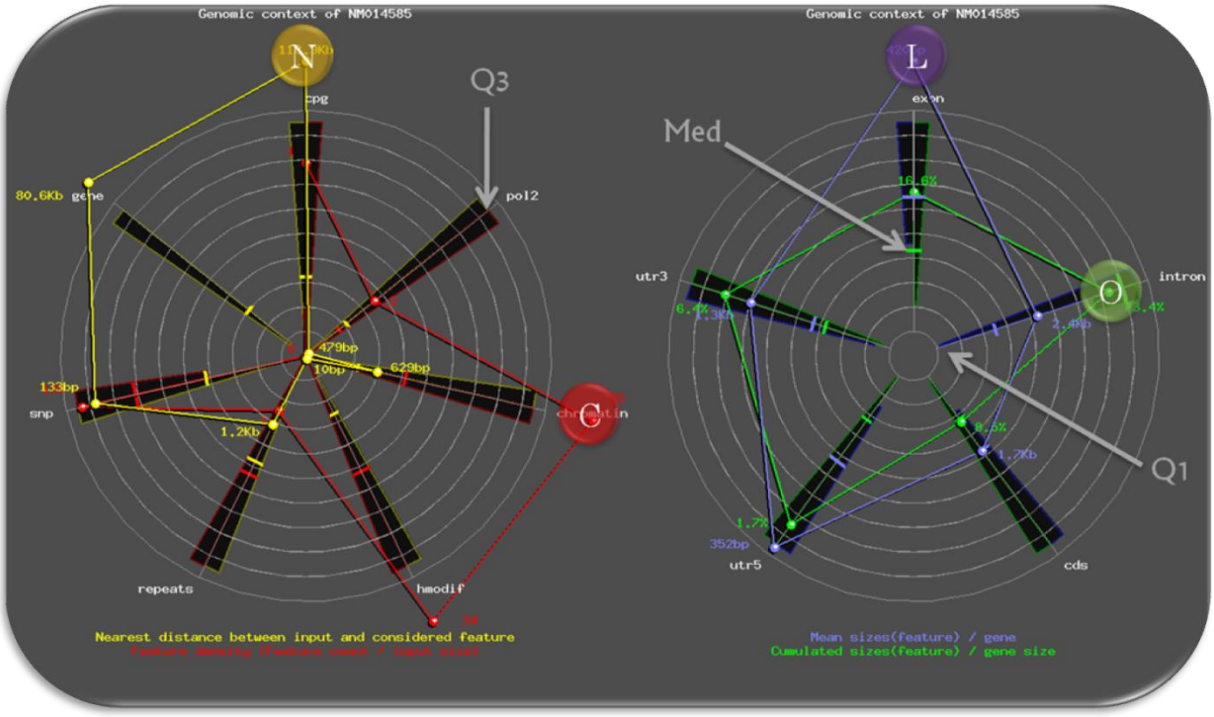
L'architecture même du portail web est une représentation prise en charge par le moteur graphique. Une attention toute particulière a été apportée à la structuration des données permettant, au besoin, une analyse en profondeurs progressives. Priorité a été donnée à la carte exonique des gènes protéiques. En effet, visualiser l'intégralité du contexte génomique d'un gène n'est pas chose triviale et pour l'heure, les solutions disponibles - essentiellement des *génome browsers*- ne permettent que d'observer correctement la distribution de vecteurs informationnels de tailles inégales sur le gène. Aussi une représentation en tableau nous a-t-elle semblé pertinente. Celle-ci détermine pour chaque exon et intron, le compte des vecteurs informationnels contenus à l'intérieur de cet exon/intron, permettant ainsi un accès immédiat à la distribution de ces entités sur le gène. De plus, le moteur graphique génère d'autres informations comme la détermination en termes de nombre d'exons des régions 5' et 3' non traduites (5'-UTR et 3'UTR, respectivement) ainsi que de la CDS, ces informations étant en général absentes des serveurs dédiés.

4.2. Radar de boites à moustaches (RBAM)

Replacer les éléments dans leur contexte ne suffit pas car, chaque donnée ayant un contexte conséquent, il faut pouvoir y accéder à l'œil et apprécier des centaines de contextes en une fois dans le cadre d'une étude à haut-débit. Un contexte se définissant par plusieurs vecteurs avec chacun ses descripteurs CLON, il fallait pouvoir à la fois les tracer tous et rendre compte du caractère atypique des données. Nous avons donc résolu d'ajouter une nouvelle dimension à la classique représentation en boite-à-moustaches, en les agrégeant sous forme de radars (Figure 44). Cette représentation permet de mémoriser un profil de distribution polaire, l'œil cherchant naturellement à rapprocher la forme de la distribution à une forme qu'il connait. Dans le détail, chaque entité mesurée constitue un axe du radar sur lequel figure une boite à moustache construite en utilisant les valeurs de Med, Q_1 et Q_3. Chaque boite est normalisée par son Q_3 (qui vaut donc 1) et rapportée à 90% de la mire du radar. Les valeurs de l'utilisateur sont alors replacées sur ces boites à moustaches, permettant d'évaluer leur spécificité par rapport au

génome. Les valeurs au-delà de 120% de la mire sont bornées à cette valeur dans un souci de délimitation graphique et peuvent d'emblée être considérées comme très atypiques.

Pour les radars liés à une entrée de l'utilisateur, les descripteurs CLON de type « N » et « C » sont groupées, de même que les valeurs de types « O » et « L », ces valeurs possédant les mêmes axes. Pour les radars moyens déterminés pour l'ensemble d'un set d'entrées de l'utilisateur, un seul descripteur CLON est figuré par radar.



A B

Jaune : Distance au plus proche vecteur informationnel

Rouge : Nombre de vecteurs informationnels par taille de région interrogée (ici : taille du gène)

Violet : Taille des vecteurs informationnels considérés

Vert : Somme des tailles des vecteurs considérés, ramenée à la taille de la région.

Figure 44 : Représentation en radar de boite à moustache du gène SLC40A1

Pour chaque entrée de l'utilisateur, le radar repositionne les descripteurs CLON sur les boites à moustaches calculées en mode batch sur le génome complet. Les valeurs N et C sont calculées pour tout type d'entrée (A) alors que les valeurs O et L sont spécifiques des gènes (B). Par exemple, on voit ici que le nombre d'entités (tracé rouge) est atypique pour les modifications d'histones et l'euchromatine et que la portion de la taille du gène (tracé vert) que représente la cds est très proche de la médiane déterminée sur l'ensemble des gènes du génome.

4.3. Gène informationnel

Afin de visualiser les coordonnées génomiques fournies par l'utilisateur par rapport à leur gène le plus proche, nous avons développé une représentation d'un gène reportant toutes ces localisations. Nous avons limité cette représentation aux parties statistiquement plus informationnelles des gènes, notamment du point de vue de la transcription. La partie en amont du TSS et la partie 5' du gène ont un fort potentiel informatif du fait de la fréquence d'éléments régulateurs. Notamment, le début de l'intron 1 présente un enrichissement important dans ces éléments (Finocchiaro *et al.*, 2007). De même, comme cela a été décrit au paragraphe 1.1 du chapitre 3, un nombre important d'éléments régulateurs est retrouvé en fin de gène, principalement dans le dernier exon. Nous avons donc inclus cette partie 3' dans notre gène informationnel. Dans le détail, nous représentons donc en 5' : la partie en amont des gènes, leur premier exon, premier intron et second exon. Parallèlement, nous représentons en 3' : le dernier intron, le dernier exon et la partie en aval du gène.

La taille des introns, et parfois des exons, pouvant s'avérer relativement grande, nous avons choisi de ne considérer que la partie la plus informationnelle de ceux-ci. Nous nous sommes donc focalisés sur les régions à proximité des bornes des exons et introns et en avons représenté une fraction de leur taille moyenne (Tableau 3). Par exemple, l'intron 1 faisant en moyenne 13kb, nous en représentons 10% à chaque borne, soit 20% des 13 kpb (2600 pb). (Figure 45).

Entité	Fraction informationnelle (%)	Taille informationnelle (pb)
Exon 1	2*50	328
Intron 1	2*10	2660
Exon 2	2*50	348
Dernier intron	2*20	2266
Dernier exon	2*50	1326

Tableau 3 : Fraction et taille de la région informationnelle des exons 1, 2 et dernier et des introns 1, 2 et dernier.

Dans le cadre des exons, la fraction informationnelle constituera 100% de la taille moyenne alors que dans le cadre des introns, ce seront de 20 à 40%

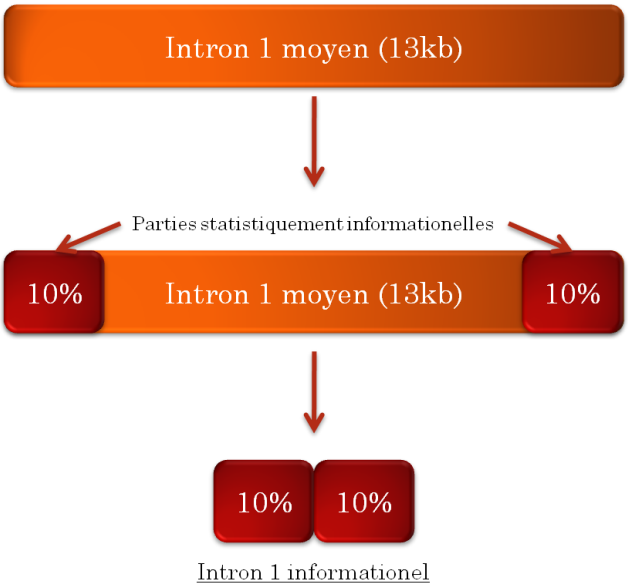


Figure 45 : Constitution de l'intron 1 informationnel représenté dans GeCo.

Les 10 premiers et derniers pourcents de la taille d'un intron moyen (13 kbp) sont identifiés et la partie intermédiaire est retirée de l'intron pour former l'intron informationnel qui sera représenté dans GeCo et mesurera donc 2600 pb.

Améliorant et automatisant une méthode suggérée par Blanchette (Blanchette *et al.*, 2006) et fournie à la Figure 46, nous décomptons pour chaque base du gène informationnel, le nombre de fois où elle est comprise dans une entrée de l'utilisateur. Notre choix de ne représenter qu'une partie des exons et introns nous a forcés à attribuer à la borne la plus proche, les bases impliquées dans une entrée de l'utilisateur.

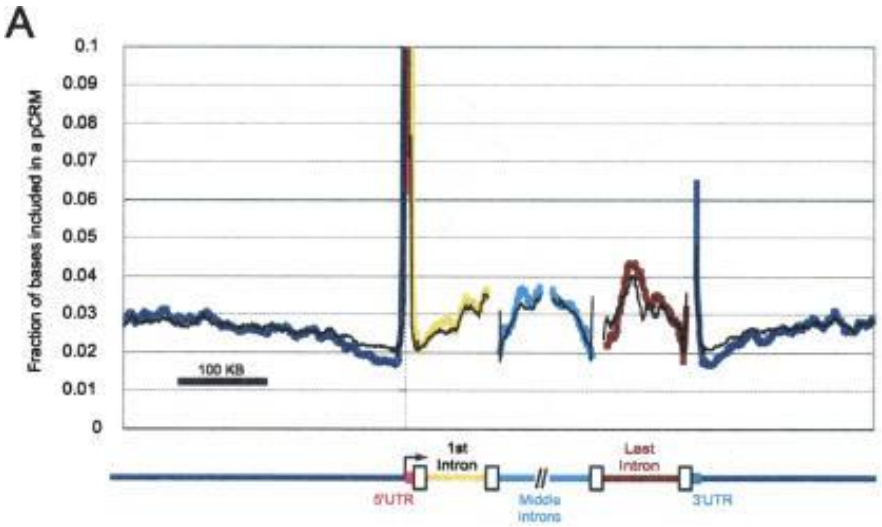


Figure 46 : Représentation en gène cumulatif selon (Blanchette *et al.*, 2006)

Cette représentation fait état de la fraction de bases impliquée dans un module cis-régulateur pour les entités en amont du gène, région 5' non-traduite (5'-UTR), intron1, autres introns,3' dernier intron, région 3' non-traduite (3'-UTR) et aval du gène.

Cas d'erreur et correctif

Le principe du gène informationnel est de décompter le nombre d'entrées de l'utilisateur incluant la $i^{ème}$ base en partant du début ou de la fin d'un exon ou d'un intron (ce que nous appellerons l'implication de Bi). Considérons le cas d'un exon E de longueur *l*, de borne de début Bα et de borne de fin Bω. Enfin, posons l^{inf} comme longueur de l'exon informationnel.

Si $i > l^{inf}/2$ et que $i<l/2$, , on se trouve devant une aberration où Bi, pourtant affectée correctement à la borne Bα ($i<l/2$), devra être représentée plus loin que le milieu de l'exon informationnel ($i > l^{inf}/2$) En ce cas, cette base n'est pas représentée. Ainsi, des données seront perdues à mesure que l'on s'éloigne des bornes. Le problème soulevé est encore plus flagrant si l'on pose que l'exon E (trop court) ne possède pas de $i^{ème}$ base. Là encore, c'est une donnée perdue, quand bien même l'entrée utilisateur couvrirait tout l'exon E. Ainsi, il y aura statistiquement plus de chance que les bases proches des bornes de l'exon informationnel soient incluses dans les entrées de l'utilisateur comparées aux bases plus éloignées. Ceci se traduira par un effet de bord dans la représentation et explique notre emploi d'un correctif des implications pour atténuer ce biais.

$$Implication\ normalisée(Bi) = \frac{\sum_{k=0} séquences_{\mathrm{k}}\ \ni\ Bi}{\sum_{\mathrm{p}=1} \mathrm{gènes_p} \mid d_{\mathrm{Bi\text{-}B\alpha}} < d_{\mathrm{Bi\text{-}B\omega}}}$$

En divisant l'implication de Bi par le nombre de fois où Bi existe dans le génome et est plus proche de Bα, on appliquera une pénalité à toutes les bases, mais celle-ci sera de moins en moins forte à mesure que l'on s'éloigne des bornes.

Un raisonnement similaire est appliqué en amont et en aval du gène informationnel. Cette fois, le correctif apporté tient compte du nombre de gènes où la base Bi n'est pas plus proche d'un autre gène que de celui considéré.

Pour illustrer l'utilisation de cet outil, nous avons relocalisé par GeCo les 1175 sites de fixation de hStaf/ZNF143 identifiés récemment dans la littérature (Myslinski *et al.*, 2006). Et nous sommes concentrés sur la partie 5'. Les résultats sont présentés à la Figure 47.

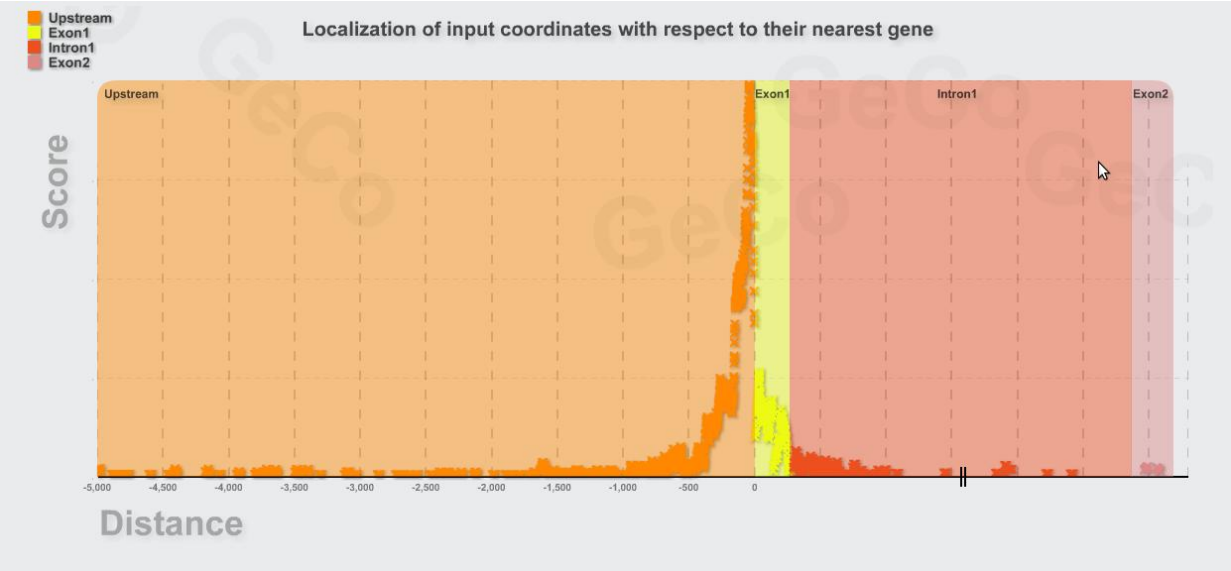


Figure 47 : Représentation en gène informationnel figurant les localisations de sites de fixations connus pour le facteur hStaf/ZNF143.

On note un fort enrichissement juste en amont du TSS puis, un effondrement dans le premier exon qui se poursuit dans le premier intron. L'exon 2 ne contient pas ou peu de sites de fixation.

5. Fonctionnalités du portail

L'interface d'interrogation de GeCo, GPS (pour *GeCo Positioning System*), est accessible via un système de gestion de contenu, (CMS pour *Content Ressource Manager*) permettant également de présenter les évolutions des outils, assurer un suivi des bugs et proposer des tutoriels ou le téléchargement de données (Figure 48). Celui-ci est accessible à l'adresse : http://gps.igbmc.fr.

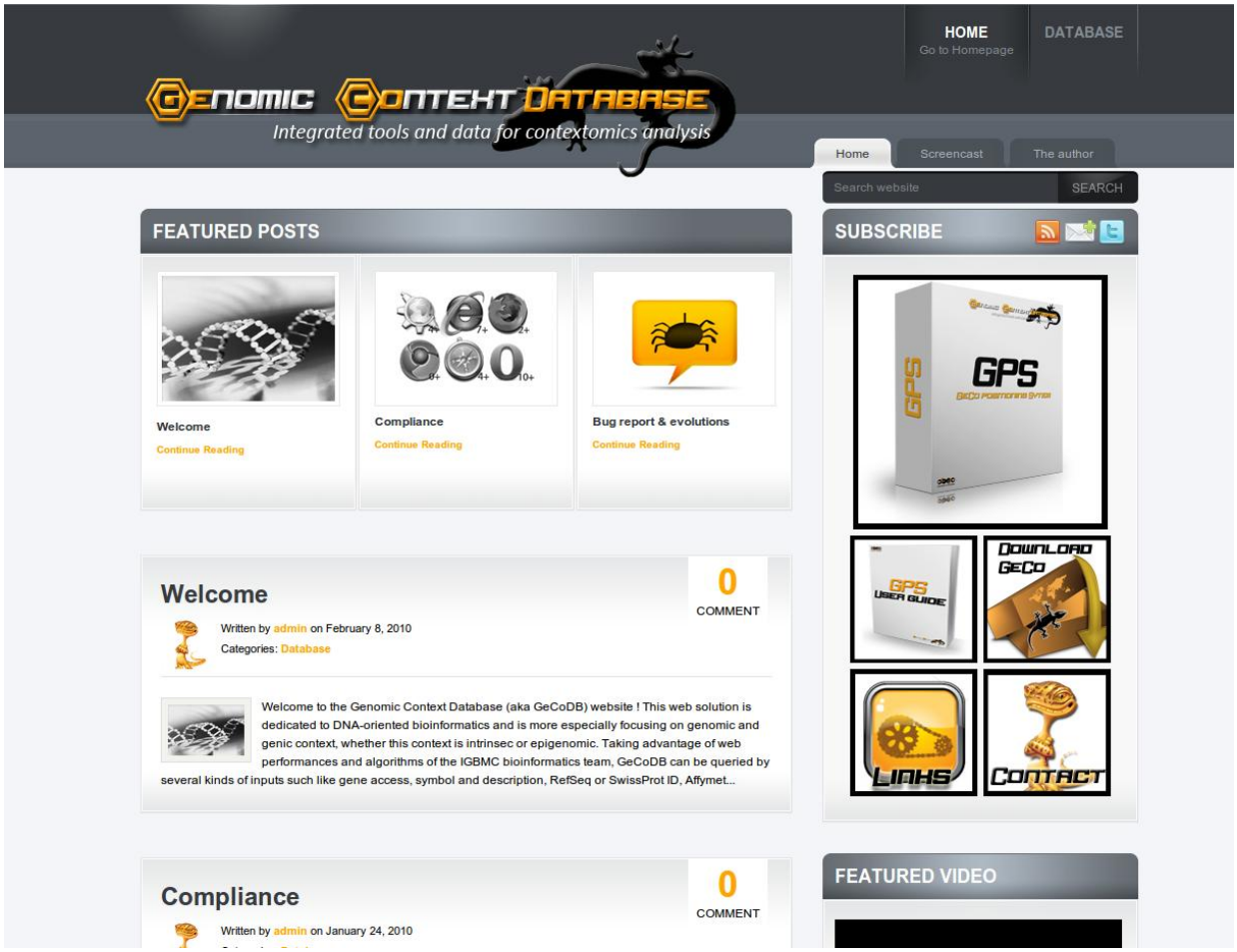


Figure 48: Portail d'accès à la database GeCo.

La partie gauche du CMS permet d'afficher des nouveautés alors que le menu contextuel à droite permet d'accéder à l'interface d'interrogation et aux fonctions classiques de ce genre de sites.

5.1. Données en entrée

Nous avons voulu rendre l'interface d'utilisation la plus simple possible, même au niveau du formulaire de saisie. Ainsi les paramètres spécifiques sont de prime abord cachés et n'apparaissent que sur demande de l'utilisateur. Le portail permet d'interroger la base de

données selon les deux périmètres précédemment définis : le génome et le gène et tous les types d'entrées admises par le portail seront-elles converties en l'un ou l'autre type de ces périmètres. Il pourra admettre comme entrée des numéro d'accès d'ARNm ou de protéines, RefSeq ou SwissProt, voire des probesets Affymetrix identifiés dans la table HG-U133A. De même, on peut choisir de rentrer par une définition de gène, comme mot-clé, en excluant au besoin d'autres mots pour affiner la requête. Enfin, en cas d'incertitude, une option permet de préciser que la donnée saisie est partielle et l'algorithme de requête se chargera de trouver les entrées la contenant. En parallèle, le périmètre "génome" permet de saisir des éléments contextuels se rapportant aux coordonnées génomiques. Ainsi l'utilisateur peut-il saisir des coordonnées ou des SNP (Figure 49).

Figure 49 : Interface de saisie du *Geco Positioning System*

L'interface affiche les nombreux moyens orientés gènes ou génome d'interroger la base de données. Certains choix comme le type de gène ou la recherche phylogénétique, donneront lieu à l'apparition de nouvelles options qui leur sont dédiées.

L'utilisateur peut également ne s'intéresser qu'à une partie de la base d'annotation de gènes et pourra en conséquence ne localiser ses entrées qu'en regard d'un certain type de gène, par exemple les ARNt. Au niveau phylogénétique, l'utilisateur peut choisir de se servir de GPS

comme récupérateur de séquences homologues de 23 vertébrés ou comme filtre préalable à l'analyse, interdisant aux données non-conservées de ne pas être affichées.

5.2. Interface graphique granulaire

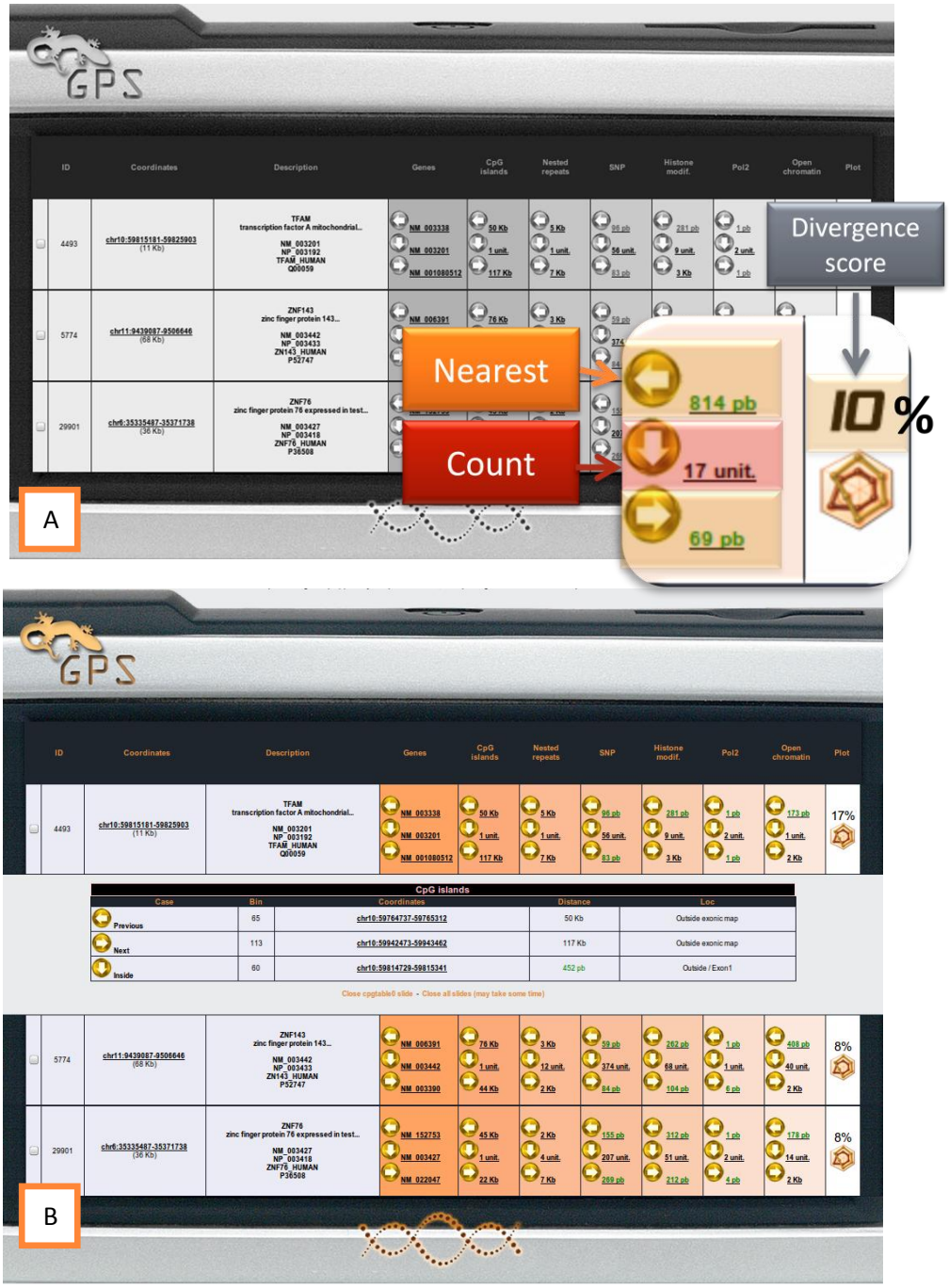


Figure 50 : Exemple d'accès granulaire aux données.

En lançant l'analyse sur 3 gènes, un GPS global affiche les données résumées de ces gènes en termes de distance et de décompte d'élément du contexte génomique (A). En cliquant sur un élément du résumé, un accès plus détaillé aux données est permis (B).

Au vu de la grande quantité de données à afficher, nous avons voulu une interface de résultat la plus ludique possible. Celle-ci file la métaphore de l'acronyme GPS en présentant les données sous forme de l'appareil du même nom. Son écran est un tableau contenant une entrée utilisateur par ligne en périmètre génomique et un gène par ligne en périmètre "gènes" (voir paragraphe 2.3). Les colonnes de ce tableau correspondent aux différentes entités mesurées. Dans la plupart des cellules du tableau, les localisations sont figurées par de petits boutons-flèches comme le ferait le localisateur automobile. On y trouve trois directions : la flèche gauche correspond au cas "précédent", la droite au cas "suivant" et la flèche vers le bas couvre à elle seule les cas "contenant" et "contenu", et, au lieu de fournir des distances, propose un décompte. Via cette apparence aux allures de table de mixage aux multiples potentiomètres, et aidé d'un code couleur catégorisant la proximité des éléments (« vert » si inférieur à 1kb, « rouge » si supérieur à 1Mb, « noir » entre les deux), l'utilisateur peut donc apprécier les valeurs chiffrées du contexte génomique de ses entrées (Figure 50A).

Pour plus de détail, il peut cliquer sur n'importe quelle valeur pour que le GPS se désolidarise à cet endroit et fasse apparaitre une table de détails jusqu'alors masquée, lui permettant d'affiner son tableau associé. Par exemple, si seule la localisation par rapport à un type d'histone modifiée donnée ou dans un type cellulaire précis l'intéressent, c'est par ce biais qu'il pourra y accéder et, pour certaines sous-tables une désolidarisation est encore possible pour arriver aux données brutes.

D'un point de vue technique, cet affichage granulaire progressif est rendu possible par le fait que les "lignes" d'un tableau ne sont en réalité pas des lignes mais des tableaux à part entière. Il n'est donc pas difficile d'y insérer d'autres tableaux cachés à faire apparaitre sur demande via des fonctions javaScript. De la même manière, seules les informations d'identifiants sont fournies pour les gènes et l'accès à la carte exonique se réalise en un clic sur l'un de ceux-ci (Figure 50B).

5.3. Score de divergence

Fournir ces résultats, même guidés par l'interface originale n'est cependant pas suffisant, l'utilisateur pouvant certes accéder à la quantité d'information qu'il désire mais se retrouvant au final avec un ensemble de valeurs qu'il ne peut pas apprécier. Aussi, pour faciliter l'évaluation de la spécificité du contexte de chaque entrée, nous fournissons un score de divergence, sous forme de pourcentage. Il s'agit du nombre de descripteurs CLON de l'entrée en deçà de Q_A ou au delà de Q_Z dans leur champ génomique respectif, rapporté au nombre d'entrées mesurées. Dans le cas des gènes, les descripteurs CLON sont au nombre de 24. Ainsi, à la Figure 50B, le troisième gène analysé, ZNF76, présente 8% de score de divergence, ce qui correspond à deux descripteurs atypiques sur 24.

Si le score de divergence fournit une bonne idée de l'écart des entrées de l'utilisateur par rapport au contexte génomique global, celui-ci ne rend pas compte de l'identité de chaque entité mesurée. Aussi peut-on si on le désire afficher les RBAM en cliquant sur le score divergence Ceci permet ainsi d'afficher un tracé de radar par descripteurs CLON et de voir quelles sont les entités divergentes, celles-ci se trouvant en deçà ou au delà de la boite à moustache.

5.4. Déceler les causes de divergence : utilisation des radars de boites-à-moustache sur un ensemble de gène

Nous avons interrogé GeCo avec un ensemble de 635 gènes identifiés dans la littérature pour présenter au moins un variant de transcrit tissu-spécifique (de la Grange *et al.*, 2010). Pour les analyses des descripteurs CLON d'un ensemble de gène, GeCO fournit les RBAM sous forme de polaroïdes à raison d'un descripteur par RBAM (Figure 51). En cliquant sur chacun d'eux, on fait apparaitre le radar se focalisant sur le descripteur CLON correspondant (Figure 52, Figure 53, Figure 54 et Figure 55)

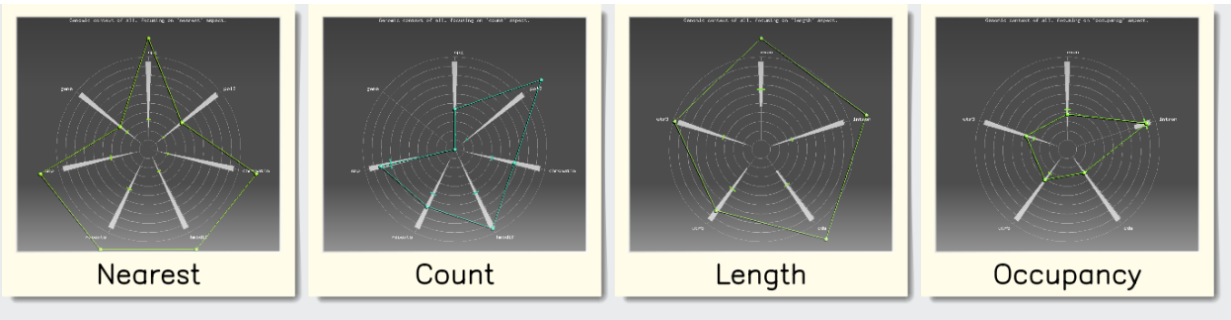


Figure 51 : Exemple d'affichage de RBAM moyens pour un ensemble de gènes via GeCo.

.

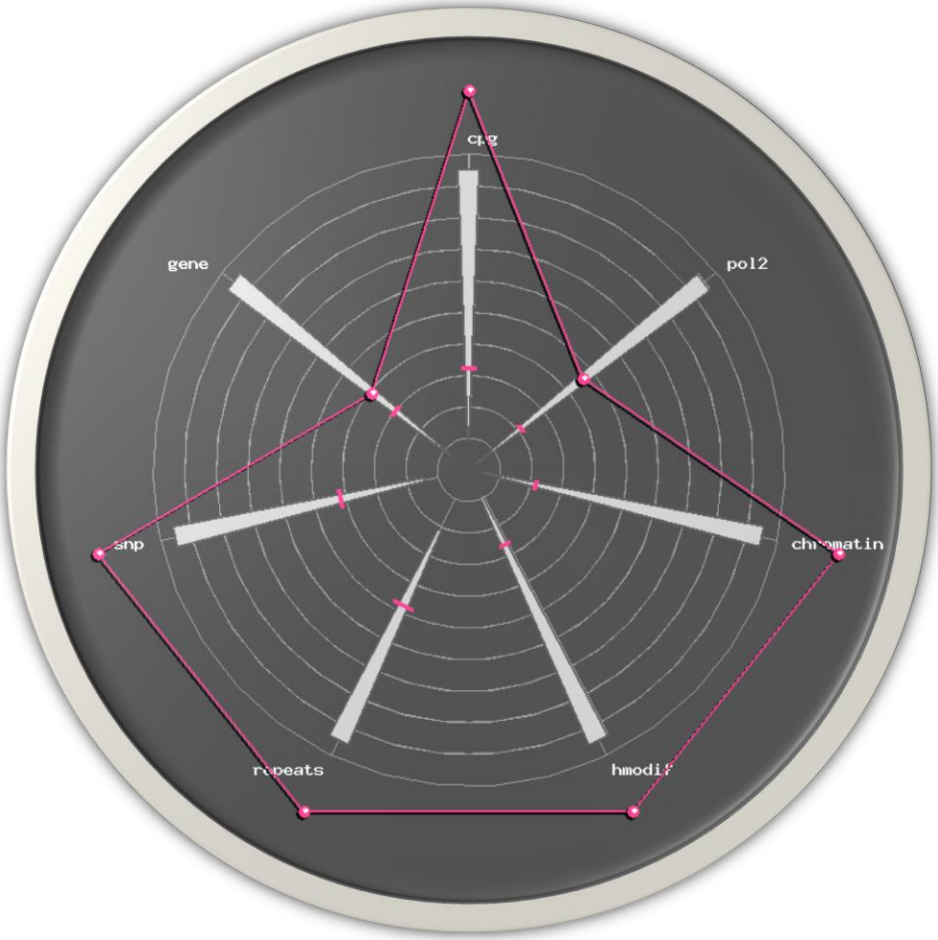


Figure 52 : RBAM moyen du descripteur "Nearest" de l'ensemble des 635 gènes à variant d'épissage tissu-spécifique.

Les vecteurs informationnels sont ici très loin des gènes, à l'exception de la polymérase II et du gène le plus proche. Ainsi, ces gènes ne se trouvent ni dans un cluster de gènes proches, ni dans une zone non-transcrite. L'absence de modification d'histone proche et d'euchromatine peut surprendre. Si l'on considère que ce sont des gènes à expression tissu-spécifique, il se peut que le contexte informationnel requis soit spécifique du tissu et donc de la lignée cellulaire qui diffèrera de notre moyenne sur l'ensemble des lignées cellulaires présentes dans GeCo.

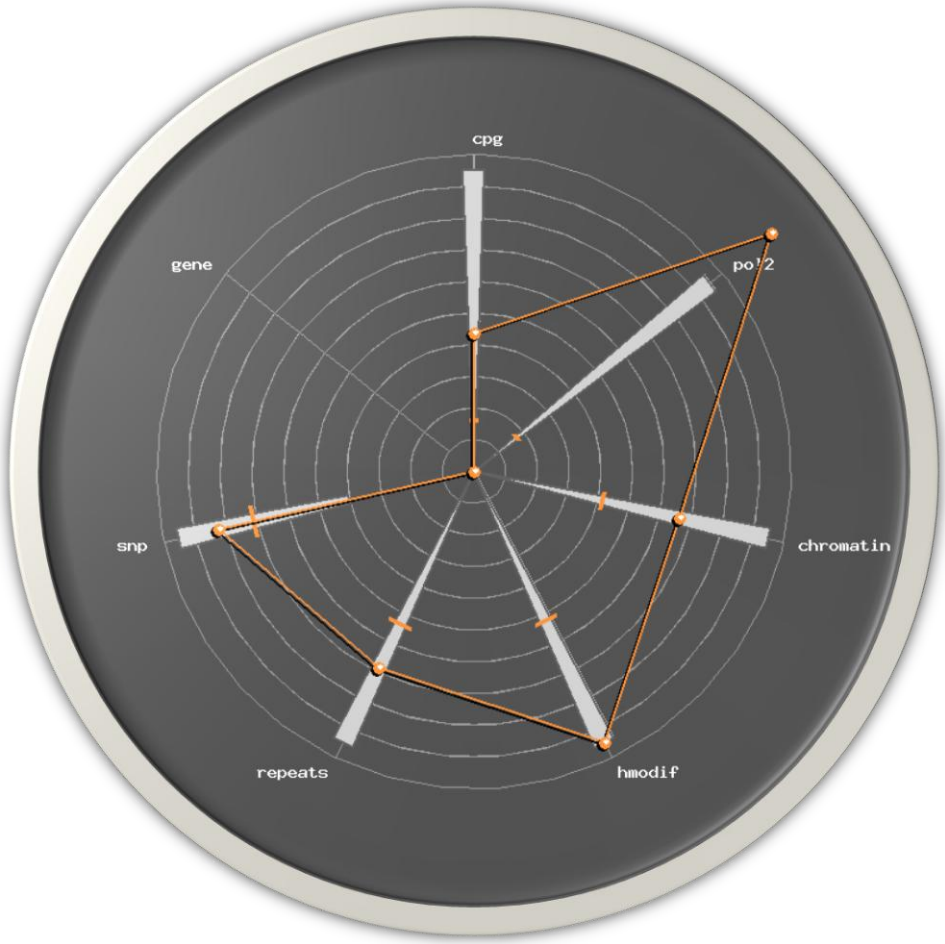


Figure 53 : Radar de boites à moustaches du descripteur "Count" de l'ensemble des 635 gènes à variant d'épissage tissu-spécifique.

Ce radar recense la décompte des vecteurs informationnels par taille de gène ou, plus simplement, leur densité. On voit ici que les gènes ne contiennent pas d'autres gènes, qu'ils possèdent tout de même parfois des îlots CpG à l'intérieur du gène, signe que ces gènes ne sont sans doute pas tous purement tissu-spécifique ou qu'un promoteur alternatif d'un des variants est de type large. Le nombre de modifications d'histones et surtout de présences de la polymérase, ramenés à la taille du gène sont très importants, signe encore de la possibilité de transcription initiée et régulée à l'intérieur du gène et corroborés par un nombre supérieur à la médiane de zones d'euchromatine. Concernant le nombre de SNP et d'éléments répétés, on est globalement dans les standards du génome.

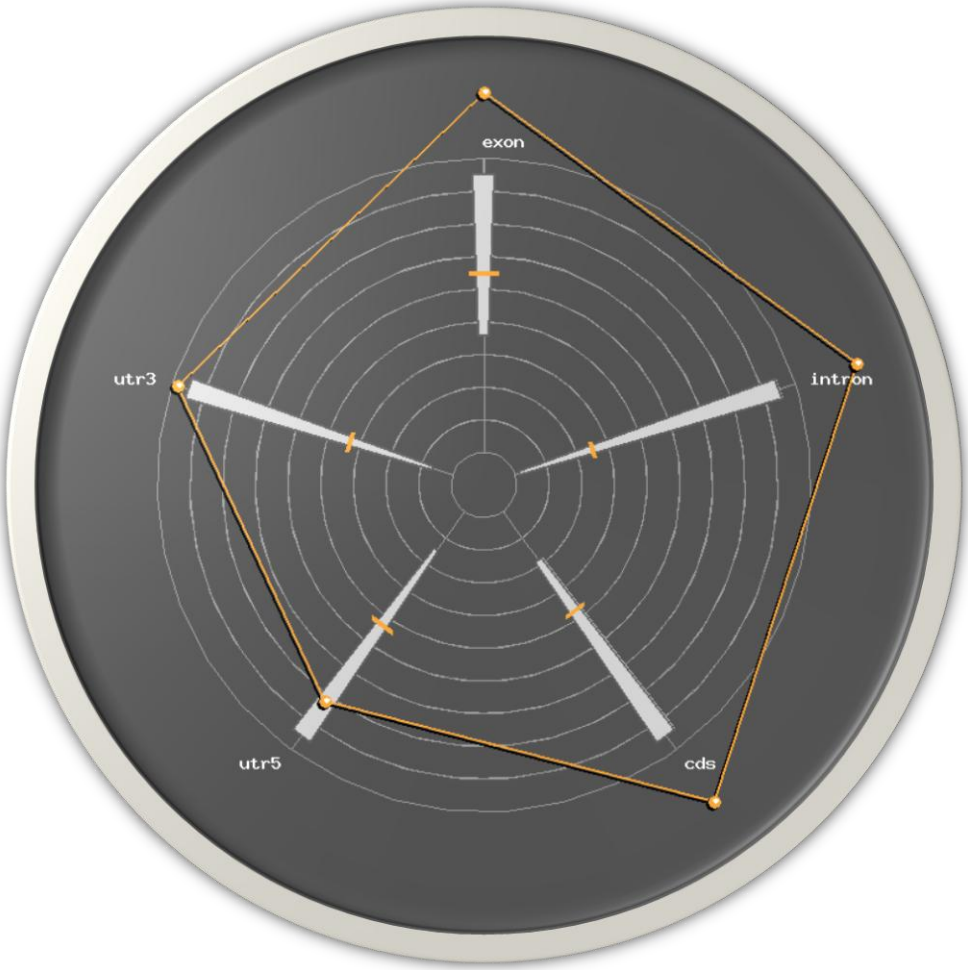


Figure 54 : Radar de boites à moustaches du descripteur "Length" de l'ensemble des 635 gènes à variant d'épissage tissu-spécifique.

On note ici que globalement , tout dans ces gènes est excessivement grand, que ce soit les exons et les introns et, manifestement, ce sont surtout les exons de la séquence codante (CDS) qui contribuent à cette divergence exonique au delà de la limite.

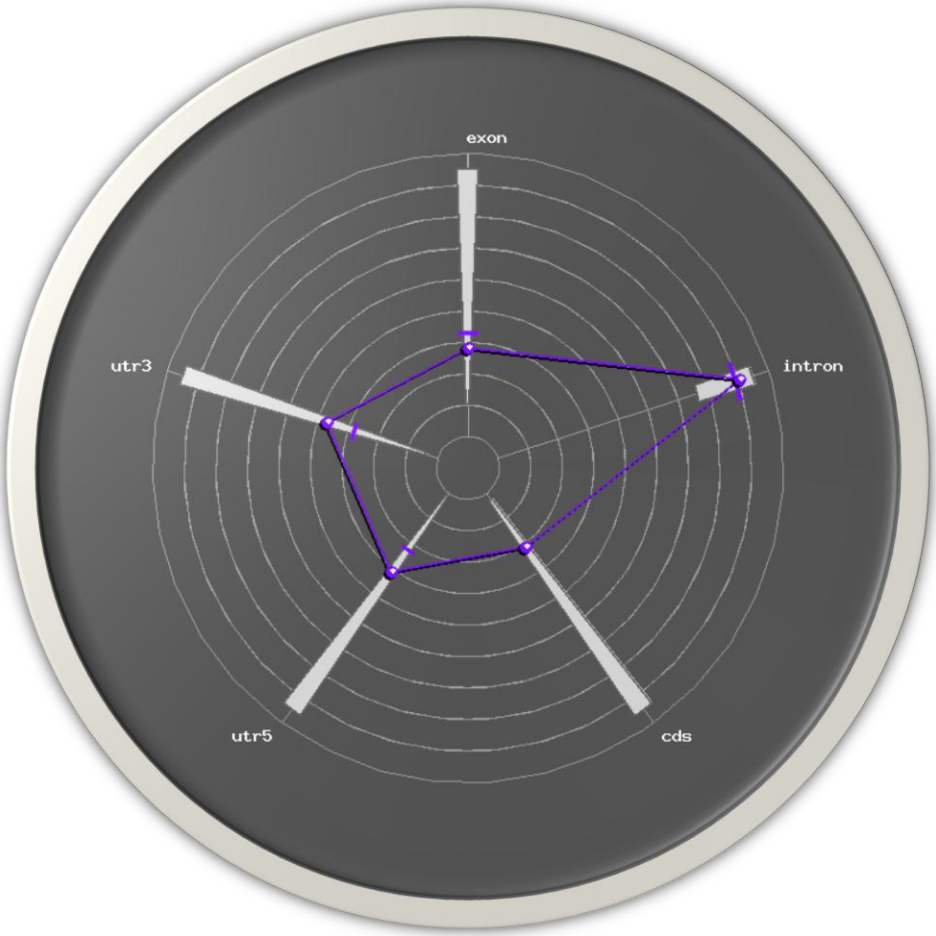


Figure 55 : Radar de boites à moustaches du descripteur "Occupancy" de l'ensemble des 635 gènes à variant d'épissage tissu-spécifique.

Toutes les occupations que peuvent faire les introns ou les exons (qu'ils soient en région 5'-UTR, 3'-UTR ou dans la CDS) sont absolument standards par rapport au génome. En comparant à la Figure 54, on en déduit que, si ces gènes possèdent des exons et des introns longs mais que ceux-ci n'occupent pas plus de place sur les gènes, c'est que ces gènes sont globalement très longs.

Par examen des radars, nous avons pu mettre en évidence une propension significative de ces gènes à contenir de nombreuses zones fixant la polymérase II, un nombre important de modifications d'histones et des zones d'euchromatine un peu supérieures en nombre à la médiane (Figure 53), ce qui est cohérent avec des gènes à variant pouvant soit posséder un TSS alternatif, soit freiner la polymérase aux jonctions exon-intron différentiellement épissées. Ainsi, par une simple analyse, on a pu vérifier la sensibilité de GeCo et son aptitude à détecter un message biologique présupposé. De manière surprenante, nous avons pu également observer que les gènes testés étaient globalement très grands sans pour autant qu'un déséquilibre de taille entre les éléments de tous types ne soit observé (Figure 54, Figure 55).

De plus, comme tous les gènes ne contribuent pas de la manière même au RBAM moyen via leurs descripteurs CLON, on peut éditer une planche-contact de tous les RBAM et chercher à l'œil, les profils communs à plusieurs gènes (Figure 56).

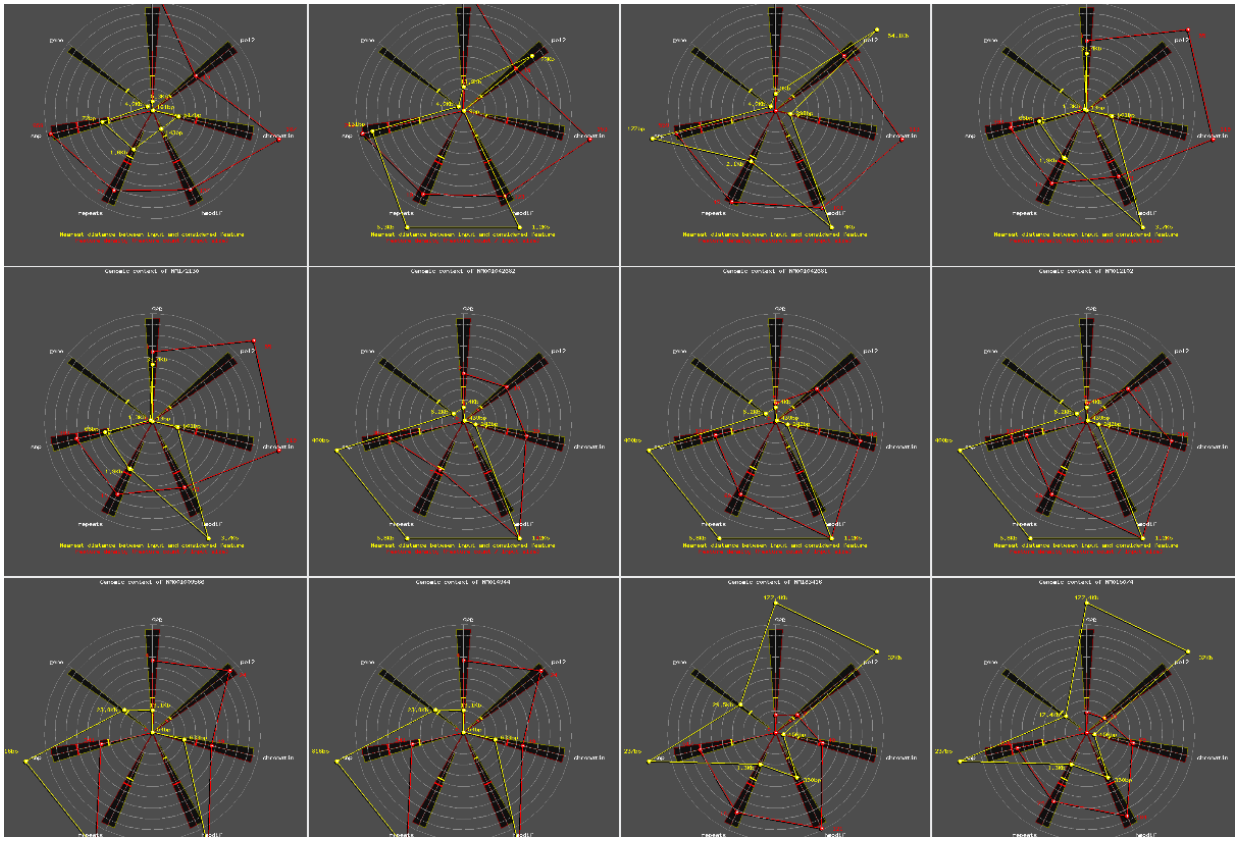


Figure 56 : Planche contact de l'ensemble des RBAM des 635 gènes à variant tissu-spécifique (12 gènes montrés).

Dans ces représentations de RBAM où figurent à la fois les radars pour les descripteurs N et C, certains profils informationnels semblent partagés entre plusieurs gènes.

Enfin, dans les cas de gènes peu documentés, voir apparaitre un tel profil de contexte informationnel pourra éventuellement s'interpréter de manière similaire, à savoir comme soulignant la possibilité de variants de transcrit. On pourra alors guider les analyses comme la recherche de voies de régulation, non pas en se cantonnant au promoteur du gène mais en s'attachant également aux régions correspondant à un contexte informationnel similaire.

Au final, cette représentation est appelée à évoluer et le peut facilement puisque l'ajout d'une entité mesurée n'est qu'une cellule de tableau à créer et, par le jeu des modifications d'histones et des lignées cellulaires, ce sont 80 entités/axes qui sont déjà prêtes à être intégrées. Il faudra cependant mener une réflexion pour prendre en compte la redondance de ce type de données similaires dans le score de divergence.

The GeCo database : a new way of making sense of genomic context

Anno YN[1*], Harle E[1], Poch O[1], Carbon P[2], Lecompte O[1]

1: Department of Structural Biology and Genomics, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Institut National de la Santé et de la Recherche Médicale (INSERM), The Centre National de la Recherche Scientifique (CNRS), UMR7104, F-67400 Illkirch, Université Louis Pasteur, F-67000 Strasbourg, France

2: Architecture et réactivité de l'ARN, Université de Strasbourg, The Centre National de la Recherche Scientifique (CNRS), IBMC, 15 rue René Descartes, F-67084 Strasbourg, France

* Address for correspondence: tel: (33) 3 88653266, e-mail: anno@igbmc.fr

ABSTRACT

Recent accumulation of genomic and epigenetic data mean we are able to draw a more and more precise picture of the genomic context, but pieces of the puzzle remain to be assembled. Whatever the nature of genomic features (phylogenetics, epigenetics, gene types, gene exonic or nucleotidic composition, etc.), they help give a more accurate view of the genome at every level of focus (gene family, chromosomal region, organism). By considering the context of a set, specific signatures can be detected - such as the epigenetic pattern of histone post-translational modifications, outliers can be eliminated and behaviors predicted. We propose that such underlying information can be extracted and rebuilt in an integrative manner for a better understanding of the entities being manipulated. At this end, we present GeCo, the Genomic Context database and its coupled web portal, providing rich and up-to-date access to most recent contextual data and building new knowledge out of it.

BACKGROUND

Interaction networks and context constraints are central in the understanding of biological processes. It is therefore crucial to address the genome and genome features with a multidimensional look-up and dress-up regions of interest with flanking and intrinsic information coming from every possible source.

As technique evolves previously limited experiments can nowadays be completed on a genome-wide high-throughput fashion. On a similar manner as the huge jump from ChIP to ChIP-on-chip, access to ChIP-Seq [1], DNAse-Seq [2] or Methyl-Seq [3] data of the EncODE project [4] provides incomparable insights on protein-DNA interactions, chromatin structure and DNA methylation and allow to reach higher steps of epigenetic knowledge on genomic context, assessing that primary sequence would be meaningless without any information about its accessibility or ability of being transcribed.

Yet, since the beginning of the deep-sequencing era, data has accumulated so exponentially that reminding even standard values like sizes and numbers of classic genomic features began to overflow our mnemonic capabilities . The limiting step for our understanding biological processes has therefore switched from determining primary sequences to global apprehension of their meaning .This leads to a frequent incapacity for users to interpret software outputs. The latter can no longer afford to only provide processed data but have to replace it with respect to its field, in order to give users proportions of what they’re manipulating and help them making sense of their

data. Such a methodology can be applied to many fields of biology, and especially to genomic data.

Several efforts have been made to give access to the information. Leading the way, raw data providing consortia tried to give easier access to their data. For instance EncODE offered EncODEdb [5], a web interface during their pilot phase to easily access and query the data. Unfortunately, this solution didn't allow to query on high-throughput fashion and has not been maintained since the start of production phase, letting UCSC genome browser providing graphic access to data. Among the plethora of genomic localization tools, mainly focusing on regulatory elements, most of them either tend to be the most exhaustive every-day tool box such like Galaxy [6] without integrative view of their data. Other can't provide high-throughput querying or focus on specific aspect only - phylogenetics for DoOPSearch [7] epigenetic statistics for Epigraph [8] - giving up crucial feature of the genomic context. On the other hand, several ones tend to upgrade algorithms by including epigenetic data. For example the Cis-regulatory Element Annotation System [9] now offers to analyze files in search or visualization of enriched peaks signal [10] without exhibiting any original gene annotation set though, and offers to detect local enrichments in a given set to assess a particular behavior to it (promoter, gene, downstream). But for now the tool can't provide any access to whole genome statistics. However, a more integrative and broader approach is needed to query the genome because providing data, even contextualized one, or at best raw results would be meaningless without any efficient structuration and tools to query and build knowledge out of it (Roos 2001; Philippi 2008).or guidelines to light user's way and make him understand his results..

At this aim, we constructed GeCo, a human-oriented flexible database, queriable with size-independent genomic coordinates or entire genes, single nucleotide polymorphisms or even full microarray probesets, in order to extract, structure and summarize inside/outside contextual information of input queries and easily compare it to pre-calculated genomic statistics, with possible choice of any type of genes.

CONSTRUCTION AND CONTENT

The aim of the database is to provide contextual data to characterize genes or genomic regions (whatever their size and nature). Contextual data include sequence conservation as well as genomic and epigenetic features: CpG islands, repeated elements, polymerase II presence, histone post-traductional modifications, single-nucleotide polymorphism and euchromatin areas.

Additional features specific to protein genes are also taking into account: exon, intron, 5'-untranslated-region (5'-UTR), coding sequence (CDS) and 3'-untranslated-region (3'-UTR). Four descriptors, the CLON values, can be associated to each feature: the feature count (C), *e.g.* the number of the considered feature found in a given region, its length (L), its occupancy (O) of a given region (size of the feature divided by the region size) and the distance to the closest entity of the feature (N). At the gene level, all CLON values are calculated whereas only C and N values are calculated for a raw genomic region. The CLON values obtained for a given region are compared to the corresponding average values calculated on the whole genome. More precisely, C and N statistics are drawn in comparison to the whole genome in the case of a raw genomic region whereas CLON values are compared to a second set of pre-calculated statistics established exclusively on the gene collection.

The GeCo database collects and combines freely available genomic contextual data, but is also a knowledge generating data warehouse, which means that the database can refine its own information by cross-comparisons. Doing so, it generates unique knowledge which is stored and ready to be refined again if needed.

Construction and query of the GeCo database relies on the same dedicated engines. These hybrid engines allow to use exact same protocols to build the database and perform user's analysis. This was the statistical prerequisite to be allowed to compare user's data and genomic ones. This constraints implies that an important part of statistics are calculated offline during database update (batch mode) to ensure users to interactively query and display a large quantity of geanomic features in a reasonable time, according to scenarii and user's desiderata.

Data sources

GeCo raw data come from 4 public sources:, the University of Santa-Cruz California (UCSC) [11], the EncODE project, the mirBase [12] from the Sanger Institute and the piRNA databank [13].

Protein genes annotation set is the human hg18-related refGenes table provided by UCSC. Extra data on gene are also present in a distinct UCSC table, KgXref, which contains description and equivalence between NCBI and RefSeq accesses. This table serves as annotation complement to GeCo's refGenes tracks. tRNA, rRNA, snRNA, snoRNA, scRNA gene annotations are part of RNAgene table, as well as CpG islands and nestedRepeats data. By convenience, NCBI dbSNP's (snp130) annotations are also downloaded here. Conservation data are also integrated with blastZ [14] whole genome pairwise alignmentsbetween human and 23 chordates: cat (felCat3), chicken (galGal3), chimp (panTro2), cow (bosTau4), dog (canFam2), fugu (fr2), guineapig (cavPor3)

horse (equCab1), lamprey (petMar1), lancelet (braFlo1), lizard (anoCar1), marmoset (calJac1), medaka (oryLat1), mouse (mm9), opossum (monDom4), orangutan (ponAbe2), platypus (ornAna1), rat (rn4), rhesus (rheMac2), stickleback (gasAcu1), tetraodon (tetNig1), xenopus (xenTro2) and zebrafish (danRer5).

Epigenetics data are directly extracted from EncODE project data hosted by UCSC and include ChIP-seq generated post-traductionnal histone modifications mapping, polymerase mapping and DNAse-generated euchromatin mapping. miRNA's annotations are provided by the mirBase [12] at the Sanger Institute and enriched by co-annotation tables mined in literature and experiments, also available at the Sanger Institute. piRNA clusters [15] are provided by piRNA databank [13].

Database architecture and querying systems

Since relational schemes are time consuming for huge volume of data, GeCo follows a half-relational database model that introduces redundancy instead of using foreign keys. To build gene collection of GeCo, we used two types of input files, according to their mySQL compatibility: UCSC and Sanger institute data pre-formatted to be imported into a mySQL architecture are imported in their native form whereas other data needed to be parsed.

We developed parsers to integrate EncODE data as thematic tables, one for each type of feature (histone modifications, euchromatin mapping and polymerase II mapping). Doing so, we centralize related data and therefore limit database access number. Other dedicated parsers have been developed to process piRNA clusters annotation and blastZ-generated genome-VS-genome pairwise alignment files. Computing data files as databases provides extreme query flexibility but the volume of generated tables requires database optimization. To improve the response time of our application, we adopted two different querying systems: mySQL for small-size tables (i.e. $< 10^6$ records) and DB2 for bigger ones, taking advantage of the in-house database engine, BIRD Biological Integration and Retrieval of Data) (Cora, 2008). The BIRD engine and its high-level dedicated querying language, BIRD-QL, were especially designed to manage huge volume of heterogeneous biological data. This system relies on an IBM architecture running on a quadcore processor associated to 16Gb of RAM and its dedicated tools for storing and exploiting data. As an example, the alignment table containing 40 millions of records (including full-length sequences of several thousands of bases) can be queried in 40 milliseconds.

Database update

To ensure automated maintenance of the GeCo database, we endowed its architecture with additional CRON jobs able to detect remote source updates and to integrate new data on the fly. These jobs can detect cyclic updates like gene annotations (ca. 2 updates / months) as well as punctual ones such like new experiments in the EnCODE project, with respect to privacy conditions.

Knowledge extraction

The knowledge extraction in the GeCo database relies on 4 interconnected engines dedicated to the features localization, statistical treatments, phylogenic analysis and finally graphical display which is an integrant part of the coupled portal.

Localization engine

The localizing engine is a program considering each feature as objects and localizing them on an all-versus-all fashion. After determining their length (CLON value: L), this engine builds four objects per feature: “previous”, “next”, “surrounding” and “inside”. The “Previous” and the “Next” objects correspond to immediate upstream and downstream features, regardless of their orientation. The distances between the feature of interest and the “Previous” and “Next” features are calculated (CLON value : N). "Inside" and "Surrounding" objects allow to retrieve features localized inside a chromosomal region and features surrounding small regions, respectively (CLON values : C, O). If the considered feature is located inside a protein gene, GeCo provides additional information. It calculates the location into the exonic map of the gene, , the distance to the nearest exonic/intronic bound and to the transcription start siteand attributes a 5’-UTR, CDS or 3’UTR flag if the features relies in one of these regions. Supplementary CLON values are calculated for exons, introns,CDS and UTR regions. For the batch mode, these values are stored in the database for ulterior quick access as they are calculated on users data to whom they will be compared in interactive mode.

Statistical engine

The statistical engine is the global comparator of genomic context. It stores CLON values, draw statistics out of them. In batch mode, this function is performed on whole genome and gene set, whereas in interactive mode, the routine is executed on user’s subset and then compared to batch-generated ones with respect to the genic or genomic mode. Due to presence of extreme CLON values (e.g. Nearest feature), very high standard deviation emerges, making the basic use of the

average and standard deviation meaningless. We therefore decided to focus on percentiles (5%, 25%, 50%, 75% and 95%) which are less sensitive to extreme values. In each mode, for each feature, every CLON values are used to extract such percentiles. The 25% (first quartile), 50% (median) and 75% (3^{rd} quartile) are provided to graphical engine to draw box plots, while 5% and 95% percentiles are being used as follows.

The engine provides user with two ways of evaluating the divergence of his data with genomic/genic ones. The first one evaluates the divergence by comparing user's CLON values to a [5 percentiles – 95 percentiles] interval ($I_{5\text{-}95}$), considering that CLON values above and beyond are atypical ones. Here a divergence percentage (DP) is calculated for each user entry by counting divergent CLON values and dividing this sum by the total of evaluated CLON values of the user's entry. An average DP is also calculated for the whole user set. As a second way of determining atypical behaviors_at the subset level we introduced in this statistical engine, two non-parametric tests: Wilcoxon and t tests using the R software. The first method allows to dispose of a quick and distribution-independent method whereas the second one is a bit more stringent with obvious risks of systematic divergence assessment due to large amount of data.

By extending a method proposed by M. Blanchette [16], the statistical engine is also in charge of normalizations for the graphical engine (see below). In order to provide an unbiased localization of user's inputs with respect to an average gene, this representation summarizes user's output and exhibits classical regions of a gene (upstream, exons, introns and downstream),. The statistical engine first extracts first, second and last exon sizes (E1, E2, EL respectively) as well as first, second and last intron sizes (I1, I2 and IL respectively) of each protein gene stored in GeCo and calculates average size for each of them. For simplification, we considered that an E1 and an E2 can't be an EL and that I1 and I2 can't be IL (AE, for average entity) and each AE is virtually cut into two parts, according to their sizes. We only model percentages in length of the exons and introns, starting from their bound. For instance, we only consider20% of the bases (2660) of average I1, 10% (1330) at each bound..). Resulting values are presented in Table 1.

Entity	Average size (bp)	Modeled part of each bound	Cumulated sizes of two modeled parts (bp)
Exon1	327	50%	327
Intron1	13298	10%	2660
Exon2	348	50%	348
Intron2	15369	10%	3074
IntronLast	11325	10%	2265
ExonLast	1326	50%	1326

*Table 1**: Normalization values calculated by the statistical engine to prepare the mapping on an average gene.***

Splitting AE's and attributing their bases to one or the other bound could result in double counting, considering the variable size of AE's, especially small ones. These double counting would lead to graphical side effects. To avoid these side effects , an implication coefficient (IS), is attached to each base of these AE, corresponding to the trend of each base to be closer to start or end boundary in the corresponding collection items. A similar operation is led for intergenic region, to attach each base to its closest gene and evaluate when a given base of the collection, upstream or downstream a gene, is no-longer associated to this gene because of a closest gene in its vicinity. Doing so, we introduce a bound-related gradient, modeling the likelihood of a base to be correctly assigned to a given bound.

Phylogenetic engine

Whole genome pairwise alignments are used to estimate the sequence conservation among chordates of a human genomic region provided by the user. As the batch part deals with the filling of the database with flat blastZ-generated files, the interactive part allows precise gene-based - localization and sequence retrieving of holomologs by using sequence, coordinates and predicted gaps. Finally, every homolog is re-localized with respect to its nearest gene using the localization engine in order to evaluate the conservation of the distance between inputs and their nearest genes across evolution.

Graphical engine

We developed a web portal relying on a localizing pipeline called the GeCo Positioning System (GPS) developed using object-oriented PHP 5.2.12 running on a Apache 2.2 server and querying a local mySQL 5 server as well as a distant IBM DB2-based server. Local database queries take advantage of PHP Data Object extension, allowing to easily switching of database resource

manager without affecting the code. This module tightly linked to the web portal will be further described in the corresponding section.

UTILITY AND DISCUSSION

The aim of this portal is to analyze a set of genomic or genes-related inputs on a high-throughput fashion, using GeCo. With this GPS-shaped tool, the user will be able to easily visualize the localization of his inputs with respect to the previously described features. Moreover, the portal gives access to GeCo-generated knowledge such like divergence percentage to highlight outliers or enriched exonic maps of protein genes. Finally, innovating graphics are also implemented in order to summarize a maximum of data and to compare them to the whole-genome context.

Input data

	Coordinates	*SNP*	*mRNA, geneSymbol, ProteinID*	*Gene description*	*probesets id*
Genomic localization	X	X	X	X	X
Average-gene mapping	X	*NA*	*NA*	*NA*	*NA*
Homolog retrieval	X	*NA*	*NA*	*NA*	*NA*
Genome-related radar	X	X	X	X	X
Gene-related radar	*NA*	*NA*	X	X	X
Individual statistics	X	X	X	X	X
Whole user's set statistics	X	X	X	X	X

Table 2 : Available functionalities and visualizations of the GeCo portal with respect to the type of input data.

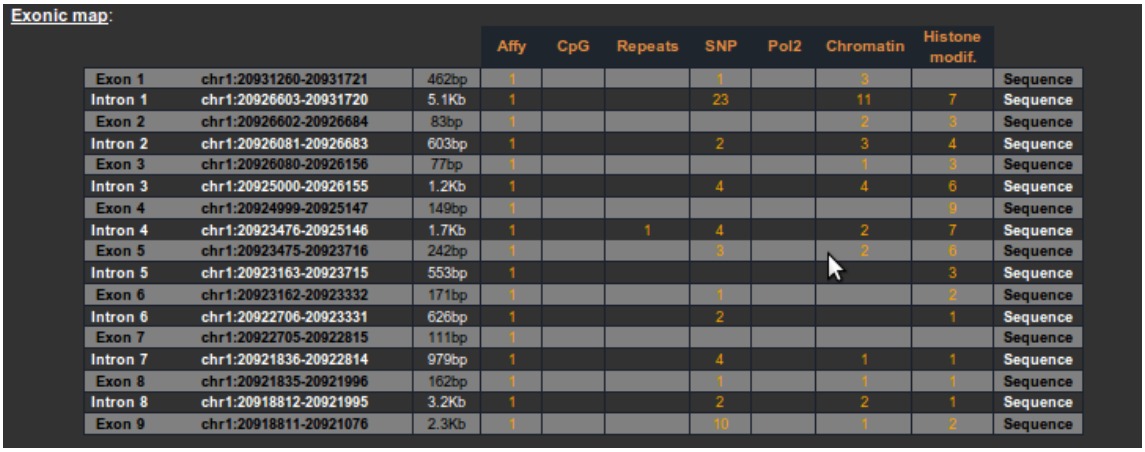
User can feed the web portal according to both genic and genomic (coordinates-based) focus. Among gene-related possibilities, either genes/protein refSeq accesses, swissprot id’s, Affymetrix 44k probesets or gene description (with possible forbidden keywords) can be used. On the other hand, coordinates-oriented inputs include genomic coordinates and SNP's. Data can be provided unitary or in a high-throughput way and file upload is allowed.

Output and visualization

In order to propose complete and summarized data in a guided easy-to-understand way. For this purpose we've developed innovating graphical charts consuming formated data coming for the statistical engine.

Exonic map table

Every gene in GeCo disposes of its precise ID card in which one can find a table-shaped exonic map in case of protein genes. This map details for every exon or intron of the considered gene, which and how many genomic feature it contains. This is the ground level of easy-to-detect distribution variations of these features, directly using pre-calculated localization data. Moreover, the ID card allows appreciating UTR and CDS occupancies in term of sizes and number of exons (Figure 3).



Exonic map:

			Affy	CpG	Repeats	SNP	Pol2	Chromatin	Histone modif.	
Exon 1	chr1:20931260-20931721	462bp	1			1		3		Sequence
Intron 1	chr1:20926603-20931720	5.1Kb	1			23		11	7	Sequence
Exon 2	chr1:20926602-20926684	83bp	1					2	3	Sequence
Intron 2	chr1:20926081-20926683	603bp	1			2		3	4	Sequence
Exon 3	chr1:20926080-20926156	77bp	1					1	3	Sequence
Intron 3	chr1:20925000-20926155	1.2Kb	1			4		4	6	Sequence
Exon 4	chr1:20924999-20925147	149bp	1						9	Sequence
Intron 4	chr1:20923476-20925146	1.7Kb	1		1	4		2	7	Sequence
Exon 5	chr1:20923475-20923716	242bp	1			3		2	6	Sequence
Intron 5	chr1:20923163-20923715	553bp	1						3	Sequence
Exon 6	chr1:20923162-20923332	171bp	1			1			2	Sequence
Intron 6	chr1:20922706-20923331	626bp	1			2			1	Sequence
Exon 7	chr1:20922705-20922815	111bp	1							Sequence
Intron 7	chr1:20921836-20922814	979bp	1			4		1	1	Sequence
Exon 8	chr1:20921835-20921996	162bp	1			1		1	1	Sequence
Intron 8	chr1:20918812-20921995	3.2Kb	1			2		2	1	Sequence
Exon 9	chr1:20918811-20921076	2.3Kb	1			10		1	2	Sequence

Figure 3: Table-shaped exonic map display provides summarized distribution of each genomic feature on its related gene.

Average gene mapping

With multiple coordinates querying, one may want to visualize their distribution with respect to their nearest gene, in the shape of an average gene representation, exhibiting classical region of a protein gene. Normalized data coming from the statistical engine are used by the graphical engine of GPS to build an XML file interpreted by the flash API XML/SWF, in charge of the final interactive display of the average gene. As mentioned before, exact localizations are converted in base-oriented implications, meaning that for each base of the average gene, an associated count is realized and normalized. This transformation implies to plot counts as a function of bases positions instead of distances and assures to use width-fixed entities while preventing side effects.

The need of such a display comes typically with high-throughput sequencing data, like ChIP-Seq. However, since the size of ChIP-Seq data reach nowadays several tenth of millions of inputs, interactive mode is no longer possible. We therefore propose a mapping-oriented very-high-throughput version of GPS. The GPS batch mapper uses the same engines as the classic ones but has been lighted of all contextual data gathering. The size of upstream and downstream regions can be set from 1 to 10kb..

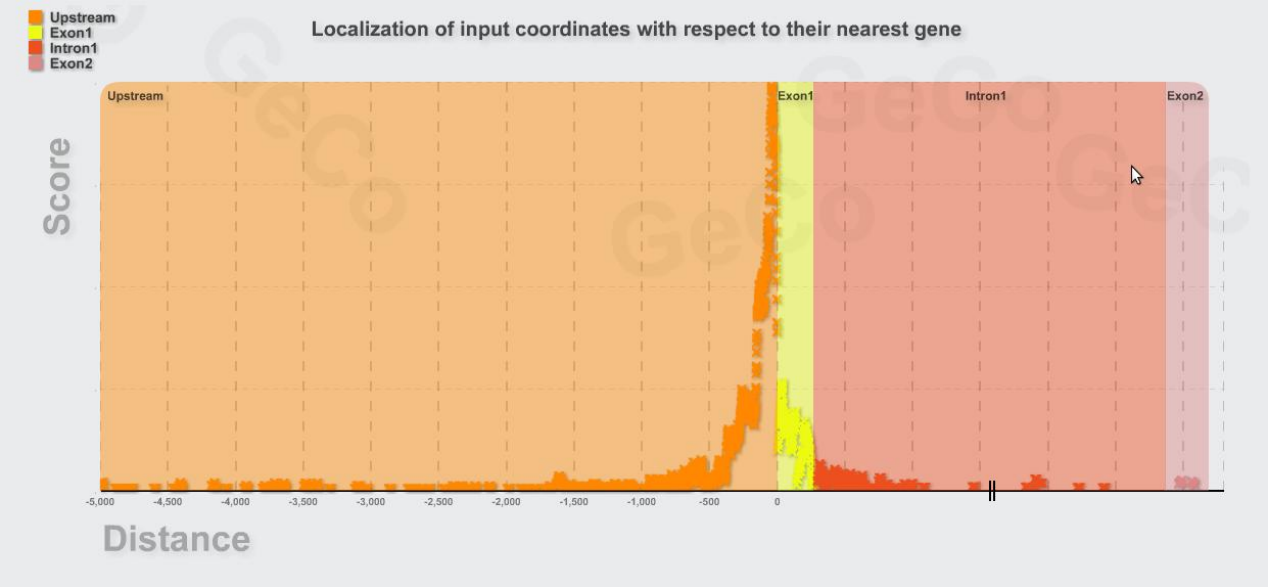


Figure 4*: Graphical mapping of ZNF143 binding site identified by Myslinski et al.** [17] ***The graphs shows a strong enrichment in the region surrounding the transcription start site.

Box plot radar

To easily visualize context features, particularities and adequacy with genomic one, we developed a statistical radar representation exhibiting on each axis a derived box-plot on which input set is plotted. Doing so, user can instantaneously compare his queries with genomic-related Q1, median and Q3 (Figure 5a). On this radar, we plot the distances (N-values, yellow) and, we superpose a second plot showing the per-base number of features (C-values, red).

If queried with protein genes, the portal provides an additional radar, showing classical protein-related features (exon, intron, UTR, CDS...) on axis, while lengths (L-values) and per-base occupancies (O-values) are plotted in blue and green respectively. This box-plot radar is available for each entry. To easily detect by eye outliers or patterns, a global view of all radars is also possible. Finally an average radar for the entire set of entries can be drawn.

FUTURE DIRECTIONS

Putting one's data on in front of the genomic landscape was indispensible. However, being able to precisely replace it using true confidence intervals has still to be done. Indeed, such law-driven confidence intervals depend on the statistical law describing data. Our statistical engine is already able to detect normal, log-normal, and exponential and gamma distributions using mixture models. Such models were previously used to appreciate genomic distances [18] or protein lengths [19]. Yet, the convergence of a mixture models strongly depending on input parameters, the success of such an approach, using so far 1200 models, depends on the type of data and can't therefore be determined automatically for now.

Replacing genomic data in their field was the mandatory pre-requisite for further interpretation of data. With evergrowing features number, the next step should be to provide efficient way of making sense of contextualized data, using on-board data mining technologies. Moreover, since there was no doubt contexual data wouldn't remain human-related only, GeCo is ready to integrate new organisms as soon as data of quality is available. Extension of EncODE project to other organisms [20] may speed up this evolution and shall open fields of whole genome "comparative contextomics", enhancing our knowledge of genomic behavior in the light of systems biology [21].

CONCLUSION

Here we presented the GeCo database, a new efficient way to query the human genome based on heterogeneous contextual features. With its collection of fully-automated engines, GeCo enables the user to compare and easily understand many concepts and proportions to quickly access the message. We highlighted the power of constructing a hybrid database, that allows to draw innovative correlations and to compare individual data to the whole genome. As we saw, our protocol pinpoints more cryptic messages and therefore reveals unsuspected mechanisms. Moreover, the absence of human intervention in the updating and knowledge generating systems in GeCo makes it reliable and perennial, and promises to keep providing access to most recent data. Finally, we believe that reproducing our data structuration protocol for other organisms opens new whole-genome "*comparative contextomics*" and enhances our knowledge of genomic behavior in the light of systems biology.

1. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.
2. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.
3. Brunner, A.L., et al., *Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver.* Genome Res, 2009. **19**(6): p. 1044-56.
4. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
5. Elnitski, L.L., et al., *The ENCODEdb portal: simplified access to ENCODE Consortium data.* Genome Res, 2007. **17**(6): p. 954-9.
6. Blankenberg, D., et al., *A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly.* Genome Res, 2007. **17**(6): p. 960-4.
7. Sebestyen, E., et al., *DoOPSearch: a web-based tool for finding and analysing common conserved motifs in the promoter regions of different chordate and plant genes.* BMC Bioinformatics, 2009. **10 Suppl 6**: p. S6.
8. Bock, C., et al., *EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data.* Genome Biol, 2009. **10**(2): p. R14.
9. Ji, X., et al., *CEAS: cis-regulatory element annotation system.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W551-4.
10. Shin, H., et al., *CEAS: cis-regulatory element annotation system.* Bioinformatics, 2009. **25**(19): p. 2605-6.
11. Karolchik, D., et al., *The UCSC Genome Browser Database.* Nucleic Acids Res, 2003. **31**(1): p. 51-4.
12. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.
13. Sai Lakshmi, S. and S. Agrawal, *piRNABank: a web resource on classified and clustered Piwi-interacting RNAs.* Nucleic Acids Res, 2008. **36**(Database issue): p. D173-7.
14. Schwartz, S., et al., *Human-mouse alignments with BLASTZ.* Genome Res, 2003. **13**(1): p. 103-7.
15. Aravin, A. and T. Tuschl, *Identification and characterization of small RNAs involved in RNA silencing.* FEBS Lett, 2005. **579**(26): p. 5830-40.
16. Blanchette, M., et al., *Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.* Genome Res, 2006. **16**(5): p. 656-68.
17. Myslinski, E., et al., *A genome scale location analysis of human Staf/ZNF143-binding sites suggests a widespread role for human Staf/ZNF143 in mammalian promoters.* J Biol Chem, 2006. **281**(52): p. 39953-62.
18. McCoy, M.W., A.P. Allen, and J.F. Gillooly, *The random nature of genome architecture: predicting open reading frame distributions.* PLoS One, 2009. **4**(7): p. e6456.
19. Zhang, J., *Protein-length distributions for the three domains of life.* Trends Genet, 2000. **16**(3): p. 107-9.
20. Rosenbloom, K.R., et al., *ENCODE whole-genome data in the UCSC Genome Browser.* Nucleic Acids Res, 2010. **38**(Database issue): p. D620-5.
21. Kitano, H., *Systems biology: a brief overview.* Science, 2002. **295**(5560): p. 1662-4.

Chapitre 7 : Application aux éléments de réponses du récepteur nucléaire à l'acide rétinoïque

En dépit de la bonne connaissance des récepteurs nucléaires, en particulier de celui à l'acide rétinoïque (AR), les éléments de réponse (RAR-REs, pour *Retinoic Acid Receptor Response Elements*) connus, à même de le fixer, sont peu nombreux et souvent déterminés de manière indirecte par transcriptomique en condition d'ajout d'acide rétinoïque suivie d'une recherche de ces éléments de réponse au niveau du promoteur (par *gel shift assay/emsa*). Ces sites de fixation sont des agencements répétitifs ou palindromiques de six paires de bases (hemi-site). On trouve des répétitions directes, des répétitions inversées, des palindromes voire même des agencements plus complexes mettant en jeu un triplet d'éléments de 6 nucléotides. Les mieux documentés sont les éléments répétés directs, séparés de 2 et 5 paires de bases (respectivement nommés DR2 et DR5), pour lesquels le consensus est respectivement: RGKTCAnnRGKTCA et RGKTCAnnnnnRGKTCA. Un des buts de cette étude était de vérifier l'hypothèse d'une certaine souplesse sur l'élément répété RGKTCA, suspecté de pouvoir évoluer vers la séquence RGKTSA.
Nous allons voir ici que la gestion des paramètres à privilégier dépend fortement du type de facteur étudié.

Nous avons recherché par script les sites de fixation correspondant au consensus RGKTCAnnnnnRGKTCA sur l'ensemble du génome humain. Afin de définir le périmètre de notre analyse, nous avons d'abord été le moins stricts possible en autorisant jusqu'à deux erreurs au consensus par hémi-site et en autorisant également les chevauchements, le tout sur le génome humain non-masqué. Il en a résulté 800.000.000 de candidats pour les DR2 et autant pour les DR5, nous forçant à revoir drastiquement nos ambitions. Nous nous sommes donc cantonnés aux prédictions ne présentant aucune erreur, sur le génome humain masqué. Il en a résulté 15.000 candidats. Comme précédemment, nous avons localisé ces candidats par rapport à leur gène le plus proche en utilisant GeCo et avons généré le gène informationnel correspondant (Figure 57).

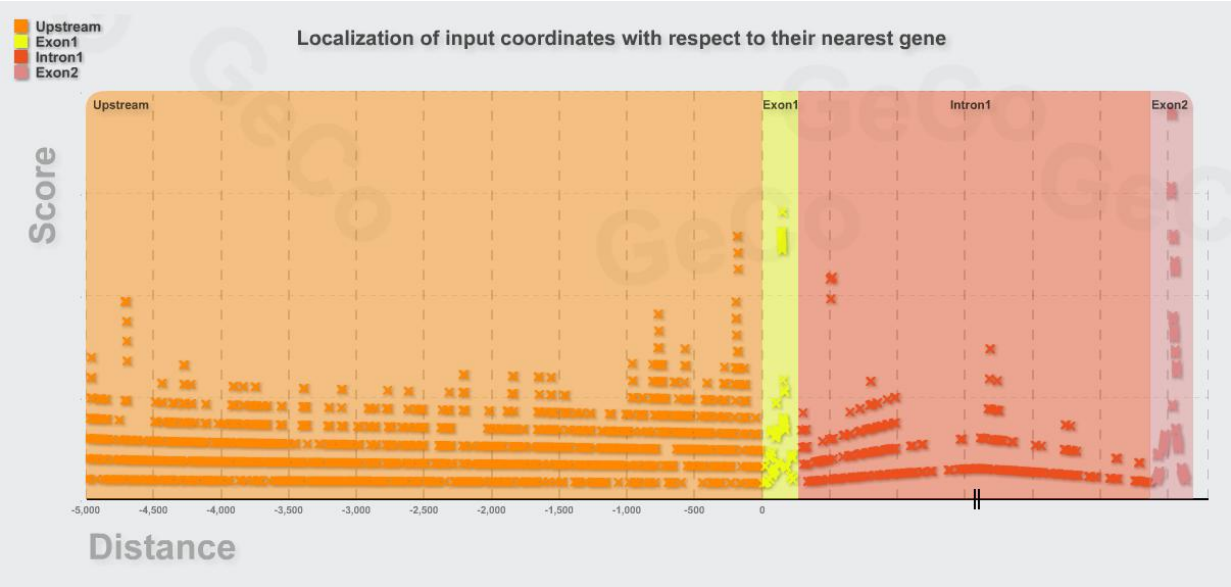


Figure 57 : Gène informationnel figurant les prédictions de RAR-RE

Ici, le peu de différences entre les implications des bases à proximité d'un gène conduit à un ensemble de lignes parallèles dues au nombre de fois équivalent où une base du gène informationnel est incluse dans un site. L'ensemble des bases partagera un nombre équivalent de sites et auront donc des ordonnées voisines.

Nous avons vu ici que les localisations de ces sites potentiels du récepteur à l'AR (ou RAR , pour *Retinoic Acid Receptor*) étaient bien moins attachées au TSS que ceux de hStaf que nous donnions en exemple au paragraphe 4.3 du chapitre 6. De fait, notre outil n'était pas adapté pour des localisations à si grande échelle. Nous avons étudié des distances de l'ordre de 100 kbp par rapport aux gènes et avons pu mettre en évidence l'enrichissement de ces sites pour une zone de 10kpb autour de chaque borne de gène (Figure 58). Si la localisation en fin de gène n'est pas surprenante (voir chapitre 3, paragraphe 2.5), l'enrichissement en revanche est important et demandera à être investigué pour comprendre la nature réelle de ces sites potentiels, possiblement impliqués dans la transactivation de gènes proches, la production d'ARN anti-sens ou simplement pour raison structurale mettant à proximité spatiale des éléments éloignés au niveau de la séquence.

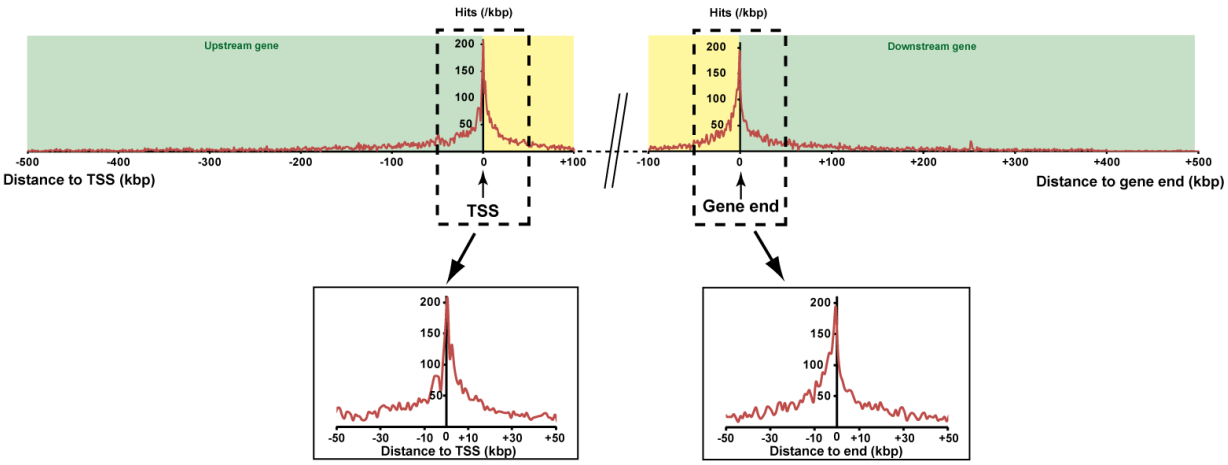


Figure 58 : Localisation des RAR-RE par rapport au gène le plus proche chez l'humain

Un fort enrichissement dans les 50 kpb entourant le TSS et les 50kpb de part et d'autres de la fin du gène.

La prédiction semblant être un maillon faible, nous nous sommes appuyés sur la littérature faisant état de la présence de ce récepteur nucléaire chez les métazoaires (Laudet *et al.*, 1992) pour choisir la conservation comme critère robuste de filtre.

Nous avons donc exigé de GeCo qu'elle ne conserve que les éléments présents dans un fragment aligné chez au moins 6 des 23 vertébrés proposés par notre architecture. Les séquences homologues ont été scorées par nos soins à l'aide du consensus évoqué plus haut. Les 100 meilleurs scores ont été gardés. Parmi eux, la moitié correspondait à des éléments de réponse déjà connus et d'autres, inconnus, mais à proximité d'un gène dont l'expression était connue pour être dépendante de l'acide rétinoïque. L'autre moitié était associée à des gènes non connus comme étant régulés par l'AR, mais présentant un enrichissement fonctionnel en gène impliqués dans le développement mis en évidence par une analyse à l'aide du logiciel DAVID (*Database for Annotation, Visualization and Integrated Discovery*, http://david.abcc.ncifcrf.gov/), ce qui était rassurant du fait du rôle important de l'acide rétinoïque dans le développement. Parmi les sites inconnus, 80% ont été testés par Sebastien Lalevée au laboratoire de C. Rochette-Egly à l'IGBMC. Des expériences de QPCR ont été réalisées dans quatre types cellulaires et 35% ont répondu positivement (dans au moins une lignée cellulaire). Quelques unes des activation à l'AR connues ou découvertes par notre étude sont présentées à la Figure 59.

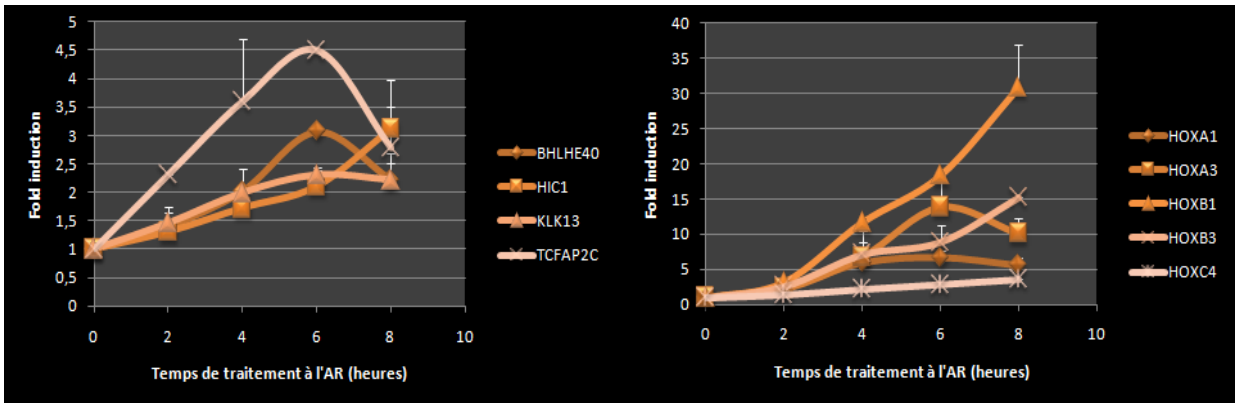


Figure 59 : Induction par l'AR des gènes à dépendance connue et hypothétique, de 2 à 8h après traitement de cellules de la lignée F9

A/ Activation par l'AR de cibles identifiées par notre analyse. On note une induction croissante sur les 8 heures à l'exception de TCFAP2C donc l'expression chute après 6 heures. B/ Activation par l'AR des gènes connus comme étant sous la dépendance de RAR. Tous voient leur expression augmenter de manière croissante, sauf HOXB3 dont l'expression diminue après 6h.On note également que l'induction par l'AR des nouveaux gènes-cibles est moins importante que pour les gènes du cluster HOX qui sont un contrôle positif classique répondant très bien à l'AR.

Ces sites validés nous ont permis de redéfinir le consensus du site qui valide notre hypothèse première que le consensus classique devait être assoupli à RGKTSAnnRGKTSA et RGKTSAnnnnnRGKTSA (Figure 60)

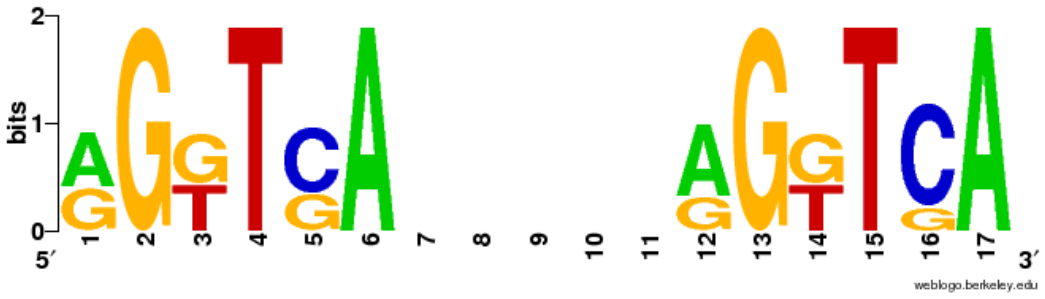


Figure 60 : Nouveau web logo redéfini à partir des RAR-RE de type DR5 connus et inconnus identifiés par notre analyse

On note que les cytosines supposées obligatoires en position 5 et 16 sont en fait substituable par des guanine.

Au vu du nombre de sites validés et du nombre restant de candidats potentiels, une expérience de RNA-Seq a été réalisée, dont les résultats sont en cours de traitement. Les résultats préliminaires ont déjà permis de retrouver 100% des activations observées par QPCR.

A la lumière de cette étude, notre analyse a permis de proposer au laboratoire expérimental de nouveaux gènes cibles de l'AR et de mettre en évidence la localisation de RAR-RE à des distances parfois très grandes des gènes. Ne pouvant nous appuyer sur un modèle mathématique plus solide que le consensus des quelques cibles déjà identifiées pour ce récepteur nucléaire, nous avons couplé nos prédictions avec un critère drastique qu'est la conservation dans 6 organismes. Ceci a été rendu possible en tirant partie de l'information la plus importante dont nous disposions sur ce facteur, à savoir sa large distribution parmi les métazoaires. On voit ici que c'est la biologie même d'une protéine régulatrice et les connaissances accumulées qui doivent guider l'analyse et appuyer le choix de la méthode de prédiction et le poids des différents critères.

Cette étude à l'échelle du génome humain des éléments de réponse au récepteur nucléaire à l'acide rétinoïque nous permit de calibrer notre méthode. Partant d'un consensus et n'exploitant que quelques vecteurs du contexte informationnel, la proximité aux gènes et la conservation phylogénétique, nous avons pu retrouver les données biologiques connues et proposer de nouvelles cibles pour ce récepteur.

Nous allons voir au chapitre suivant dans quelle mesure une quantité plus importante de données biologiques initiales nous a permis de tirer pleinement partie de l'architecture du contexte informationnel fourni par GeCo lors de l'analyse des sites du facteur de transcription hStaf/ZNF143 dans le génome humain. Ce facteur aux sites beaucoup plus dévoyés que RAR nous a contraint à affiner notre méthode de prédiction en nous appuyant sur les connaissances biologiques accumulées jusqu'alors. En l'absence première d'étude expérimentale à l'échelle du génome, nous avons d'abord entrepris de prédire les sites de ce facteur afin d'évaluer les cibles potentielles et d'en dresser les grandes lignes caractéristiques. Au vu des résultats très encourageants, une expérience de ChIP-Seq a été menée et nous permit d'accéder au répertoire effectif des sites de fixation du facteur. Son analyse fut réalisée en couplant les modèles mathématiques à l'analyse contextuelle offerte par GeCo, en interaction permanente avec le laboratoire expérimental. En approchant le problèmes sous ces angles complémentaires, nous avons pu passer au delà des limites de chaque domaine et mettre en évidence le rôle majeur du facteur hStaf/ZNF143 dans le génome humain.

Chapitre 8 : Application aux sites de fixation du facteur hStaf

En tant que facteur aux nombreux sites supposés et au contexte informationnel profondément ancré dans la régulation de la transcription, le facteur hStaf/ZNF143 s'imposait comme un exemple de choix pour tirer pleinement partie de notre architecture. Contrairement aux éléments de réponse à l'acide rétinoïque (voir Chapitre 8), les sites de fixation d'un facteur de transcription sont très dévoyés, *a fortiori* ceux du facteur hStaf (voir Chapite 1, paragraphe 4). De plus, nous disposions cette fois des connaissances conséquentes accumulées sur ce facteur et de l'expertise du laboratoire de biologie moléculaire pour guider nos pas dans l'analyse des cibles de ce facteur. Nous allons voir ici comment GeCo révéla un nombre très important de sites de fixations, notamment dans les agencements de gènes en promoteurs bidirectionnels dont 94% des sites testés expérimentalement ont été validés. Nous verrons aussi dans quelle mesure l'approche par le contexte qu'offre GeCo, couplée aux modèles mathématiques et à la biologie permit d'identifier la quasi totalité des sites cryptiques, montrant au total 4088 pics.de ChIP-Seq faisant de hStaf un régulateur de nombreux gènes du génome humain.

1. Données non-biaisées : recensement des sites de fixation de hStaf (SBS) prédits dans le génome humain

Les recherches antérieures de SBS (Myslinski *et al.*, 2006) ont révélé un grand nombre de sites potentiels. De plus, les auteurs se sont appuyés sur la base dbTSS qui est une base de données orientées exclusivement sur les régions promotrices ou localisées autour du site d'initiation (-2kb-+2kb). Or, au vu des données récentes et notamment, issues des `omics, (voir chapitre 3, paragraphe 2.5), il est clair que des sites de fixation fonctionnels de facteurs de transcription peuvent se trouver globalement partout sur le génome. Ceci augurait qu'un nombre potentiellement important de sites restait à découvrir à l'échelle du génome. Nous avons donc commencé notre analyse par la prédiction du répertoire de ces sites de fixation au sein du génome humain.

1.1. Choix de la méthode

Disposant de 400 sites validés expérimentalement, nous avons résolu de construire un modèle à partir de ceux-ci et d'utiliser ce modèle pour explorer le génome (voir chapitre 4, paragraphe 1.1). Avant toute chose, il a fallu actualiser les coordonnées chromosomiques et les séquences correspondant aux sites connus puisque la version du génome était passée entre temps de hg16 à hg18. La limite du programme BLAT étant des séquences de 30 nucléotides, contre 18 pour le SBS, nous avons opté pour le programme BlastN, 500 fois moins rapide mais à même de gérer de si courtes séquences. Il permit de relocaliser 397 des 400 séquences.

Pour construire le modèle mathématique, les deux méthodes les plus couramment utilisées étaient la méthode par profil et celle à base de HMM (voir chapitre 4, paragraphe 1.1.2) , nous avons testé ces deux approches afin de déterminer laquelle serait la plus pertinente dans notre cas. Notre critère de validation de la méthode de prédiction fut l'efficacité de ce modèle à retrouver sur un chromosome donné (le chromosome 19) les sites expérimentalement validés. En théorie, ces sites ayant servi à construire le modèle, ils devaient être tous retrouvés mais force est de constater que ce ne fut pas le cas.

Nous avons commencé par tester un programme se basant sur des HMM. En termes de promotologie, le programme HMMsearch (à l'époque en version 2) fait référence dans le domaine et nous l'avons donc testé. Il utilise les séquences en entrée pour construire un modèle de Markov et en calculer les états de transition cachés entre les séquences. Celui-ci fonctionne d'autant mieux qu'on lui donne un jeu de données supplémentaire censé constituer le *background* des sites afin de mieux mettre en évidence les HMM décrivant le passage de ce *background* au site. Il était évidemment compliqué de fournir le génome comme *background*. Nous avons donc utilisé une absence de *background* et deux *backgrounds* qui ont été définis respectivement à partir de séquences aléatoirement générées et à partir de séquences flanquant les SBS expérimentalement vérifiés. Le premier présente l'avantage de ne pas biaiser l'analyse en ciblant sur des promoteurs alors que le second se base sur une réalité biologique et non une pure aléatoire pouvant être dépourvue de sens. La méthode retrouvant le plus de sites connus fut évidemment celle se basant sur un background « promoteur » mais, contre toute attente ne dépassa pas les 40%. Elle put être augmentée à 78% en utilisant un HMM décrivant partiellement le site, mais ceci mit en évidence plus d'un million de sites pour le seul chromosome 19, ce qui était inacceptable car tout simplement ingérable à l'échelle du génome entier et présentant certainement un nombre prohibitif de faux positifs.

Nous avons testé par ailleurs la méthode basée sur une PSSM, ou méthode de profil (voir Chapitre 4, paragraphe 1.1.1). Nous avons pour cela utilisé la suite de Genomatix et en particulier MatDefine (Cartharius *et al.*, 2005) permettant de définir une matrice à partir d'un ensemble de séquences. Le programme constitue en premier lieu une matrice d'occurrences (Figure 61) de laquelle il calcule les probabilités et les log-score associés pour constituer sa matrice de scores position-spécifique: De plus MatDefine propose d'éliminer les séquences redondantes ou peu informatives afin de fournir un modèle le plus robuste possible. A ce titre, il ignora 50 séquences et n'utilisa donc que 347 séquences pour construire sa matrice.

A

Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	32	12	0	0	0	294	35	127	233	46	32	3	100	48	19	44	10	44
C	122	25	340	347	347	1	47	130	7	24	40	337	150	149	192	9	255	16
G	36	5	0	0	0	50	219	74	101	120	253	5	19	71	19	273	70	255
T	156	305	7	0	0	2	46	16	6	157	22	2	78	79	117	21	12	32
IUPAC	Y	T	C	C	C	A	G	N	A	K	G	C	N	N	Y	G	C	G

B

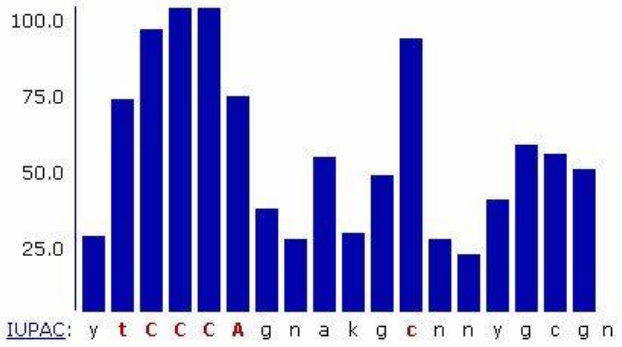


Figure 61 : Représentation du profil défini par MatDefine à partir des 397 SBS validés expérimentalement.

A/ Matrice d'occurences à chaque position des 4 bases azotées et consensus IUPAC associé pour les 18 bases que compte le SBS. B/ Représentation en histogramme associé au profil. Celui-ci évalue la fraction de séquences vérifiant le consensus à chaque base.

Par examen de la matrice, on se rend compte de la prévalence des cytosines en 3ème, 4ème, 5ème et 12ème position, la 4ème et la 5ème semblant obligatoire. Nous avons ensuite fourni cette matrice a un autre programme de la suite logiciel Genomatix, MatInspector, permettant d'analyser une ou plusieurs séquences de la taille d'un chromosome à la recherche de nouveaux sites. En gardant le seuil optimal suggéré par le programme, ceci permit de retrouver 83% des sites

connus sur le chromosome 19, ce qui était relativement acceptable et nous fit choisir cette méthode pour l'exploration du génome complet.

Ce test de performance des méthodes nous fit découvrir un fait important inconnu jusqu'alors. Dès la première recherche, nous nous aperçûmes que le SBS avait été intégré au cours de l'évolution par des éléments répétés de la famille Alu. Ces SBS présentent certainement un aspect intéressant d'étude dans le cadre du contexte génomique, mais s'avèrent tout aussi certainement très particuliers, en nombre massif de surcroit. De fait, nous avons résolu de les écarter de notre analyse et avons réalisé l'ensemble du test de performance et la recherche des SBS sur un génome humain masqué, c'est-à-dire en utilisant une version du génome où les éléments répétés de toute nature ont été remplacés par des « N » (au niveau de la séquence) par le programme repeatMasker.

1.2. Prédiction du répertoire des sites de fixation

Nous avons donc utilisé MatInspector pour cribler le génome humain masqué dans son intégralité. Ceci nous permit de mettre en évidence 230000 candidats pourvus d'un score supérieur au seuil optimal de 0.81 Par examen manuel des séquences nous avons vu que les séquences au score inférieur à 0.82 n'avaient que peu de chances d'être de réels SBS et avons donc réévalué le seuil optimal à 0.82. Ceci permit de réduire le nombre de candidats à 165000. Malgré notre réévaluation, le nombre aléatoire attendu de notre matrice était de 100000. Le nombre de 165000 candidats semblant évidemment irréaliste, nous avons tenu, dans un premier temps, à vérifier notre méthode en nous appuyant sur GeCo.

1.3. Validation de l'adéquation de la méthode de prédiction avec le génome complet

Nous avons voulu valider notre méthode en nous assurant de l'enrichissement attendu aux alentours des sites TSS de gènes. A cette fin, nous avons localisé l'intégralité des 165000 SBS candidats en utilisant GeCo et sa représentation non-biaisée sur un gène moyen.

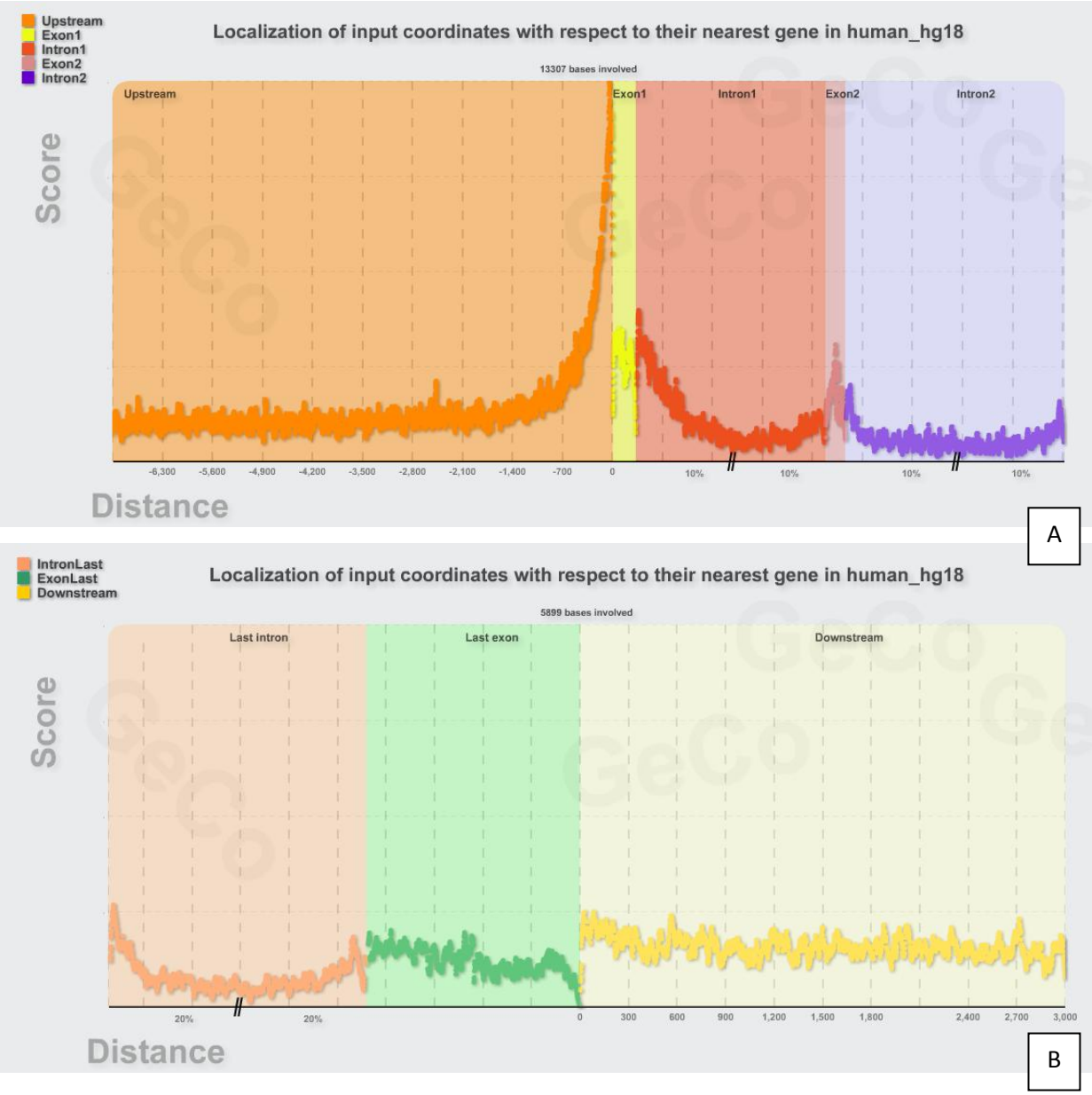


Figure 62 : Représentation en gène informationnel des 165000 SBS localisés par rapport à leur gène protéique le plus proche

A/ Un fort enrichissement est observé dans la région immédiatement en amont des gènes, puis vient un effondrement dans le premier exon, précédant un épaulement en début de premier intron. B/ En 3' des gènes, rien n'émerge du bruit.

Ceci permit de mettre en évidence un enrichissement massif peu avant le TSS des gènes puis un effondrement dès le premier exon. Ceci souligna également un enrichissement en tout début de premier intron du fait d'un épaulement dans la représentation graphique à cet endroit. En sélectionnant une fenêtre incluant les deux enrichissements, de 2 kpb de part et d'autre du TSS, nous avons pu observer que 12% des sites prédits s'y trouvaient, ce qui allait dans le sens d'une validation de notre méthode (Figure 62).

1.4. Séquences flanquantes.

Nous avons analysé les séquences immédiatement en amont et en aval des 165000 SBS prédits de détecter un éventuel motif enrichi. Ce volume de séquence étant trop important pour le logiciel MEME (voir Chapitre 4, 1.3.2), nous avons donc développé l'outil SignalMiner, recensant tous les mots de longueur N à une position fixe ou variable d'un ensemble de séquences. En recherchant les motifs de 4, 5 ou 6 pb de longueur variable, sans position particulière, aucun signal n'émerge. En revanche, en fixant les positions et en décomptant les motifs débutant à chaque position, on détecte un signal pour les motifs de 4pb. à partir de la position -7 et possiblement -9 par rapport au SBS et ce jusqu'en -4 (Figure 63A). Enfin, en aval du SBS aucun signal n'émerge du bruit (Figure 63B).

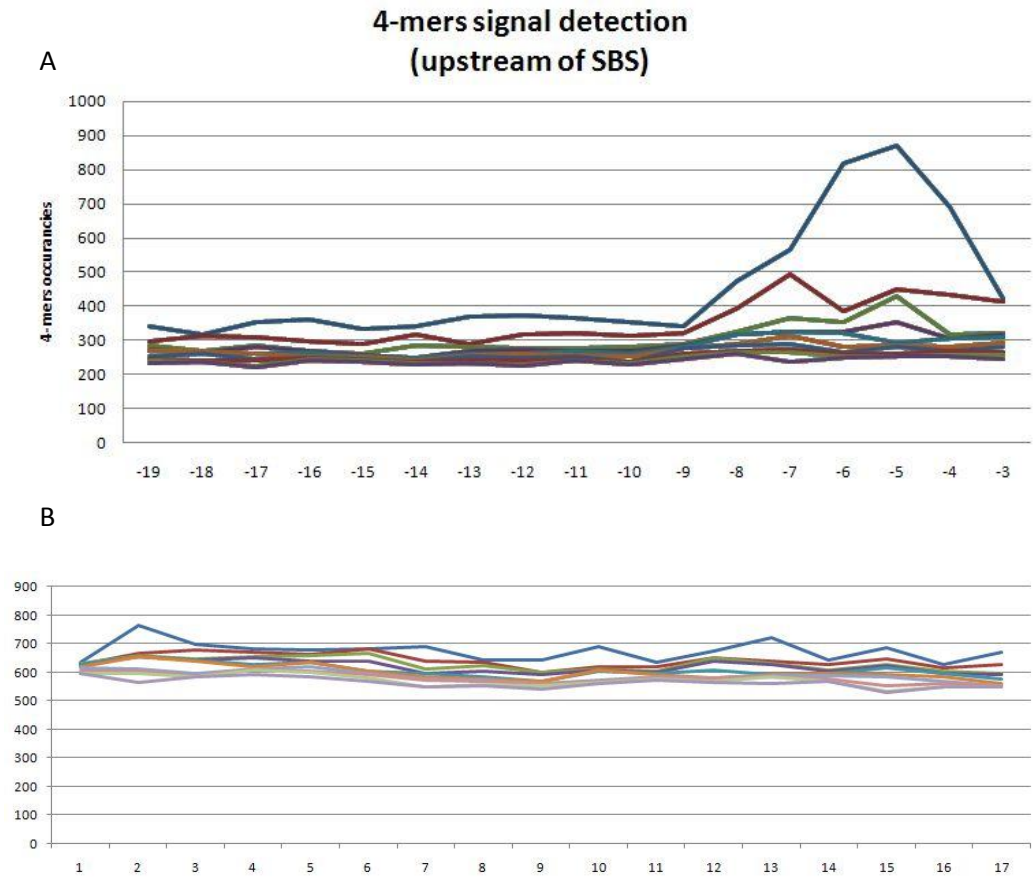


Figure 63 : Détection de signaux de 4 bases en amont et en aval du SBS

A/ En amont, une recrudescence de signaux de 4pb apparait de -7 à -3 par rapport au SBS. B/ En aval du SBS, aucun signal n'émerge du bruit.

Par rapport au SBS, c'est la séquence ACTA s'étalant de la position -7 à -4 qui présente le plus fort signal, puis CTAC de -6 à -4 et TACA en position -5 à -2. Ainsi, on retrouve bien ce motif

ACTACA en amont d'un certain nombre de sites mais ce motif ne doit pas être aussi stricte et certains écarts au consensus demeurent possibles puisqu'en examinant les 10 meilleurs motifs de 6 bases localisés en amont d'un SBS, nous avons mis en évidence un certain nombre de séquences ressemblantes parmi le plus représentées (Figure 64).

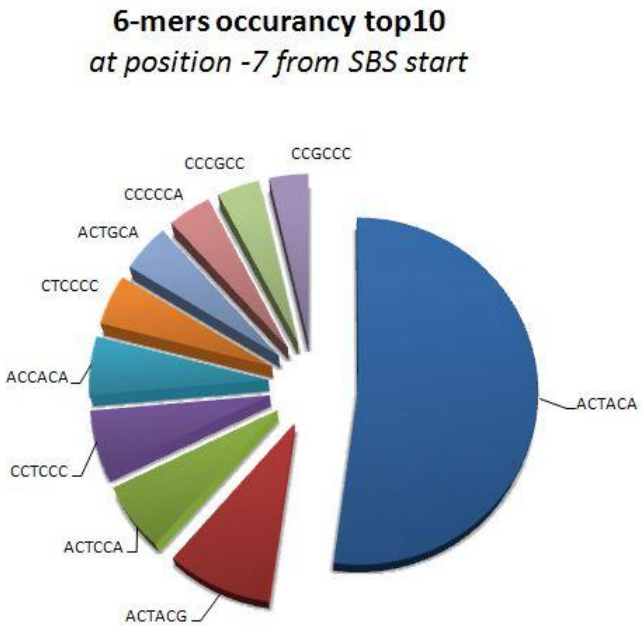


Figure 64 : Top 10 des mots de 6 bases retrouvés en position -7 par rapport au SBS.

Le motif ACTACA est retrouvé en plus forte proportion. Dans une moindre mesure, d'autres motifs similaires comme ACTAG et ACTTCA suggèrent une certaine flexibilité quant au motif retrouvé en 5' du SBS.

2. Une sous-famille de SBS localisés dans les promoteurs bidirectionnels

Devant le nombre important de promoteurs, nous avons choisi de cloisonner les populations. En interrogeant GeCo avec les gènes associés à un SBS prédit à proximité d'un TSS, nous avons observé une propension de ceux-ci à se trouver proche d'un autre gène. En étudiant de plus près, nous avons vu que ces gènes étaient souvent en orientation inverse et à une très courte distance de leur gène associé. Nous avons voulu vérifier ce possible rôle de hStaf dans les promoteurs bidirectionnels.

2.1. Les promoteurs bidirectionnels

2.1.1. Définition opérationnelle

Le consensus admis pour un promoteur bidirectionnel est une région intergénique de moins de 1 kpb séparant deux gènes en orientation inverse, présentant leur coté 5' à cette région (Trinklein *et al.*, 2004). La définition s'est toutefois étoffée. Il est aujourd'hui communément admis que de manière générale la transcription est bidirectionnelle à partir d'un TSS mais que dans le sens opposé au gène, celle-ci s'arrête très rapidement. En conséquence, le terme de transcription bidirectionnelle tend à être remplacé par transcription divergente, et commence à être réservé aux seuls cas où les gènes impliqués seraient co-régulés. Dans notre analyse, nous n'avons inclus que les gènes partageant un promoteur en 5' (que nous appellerons toujours promoteurs bidirectionnels) et n'avons pas considéré les gènes avec un intergène liant leur 3' ou 3' -5' (Figure 65)

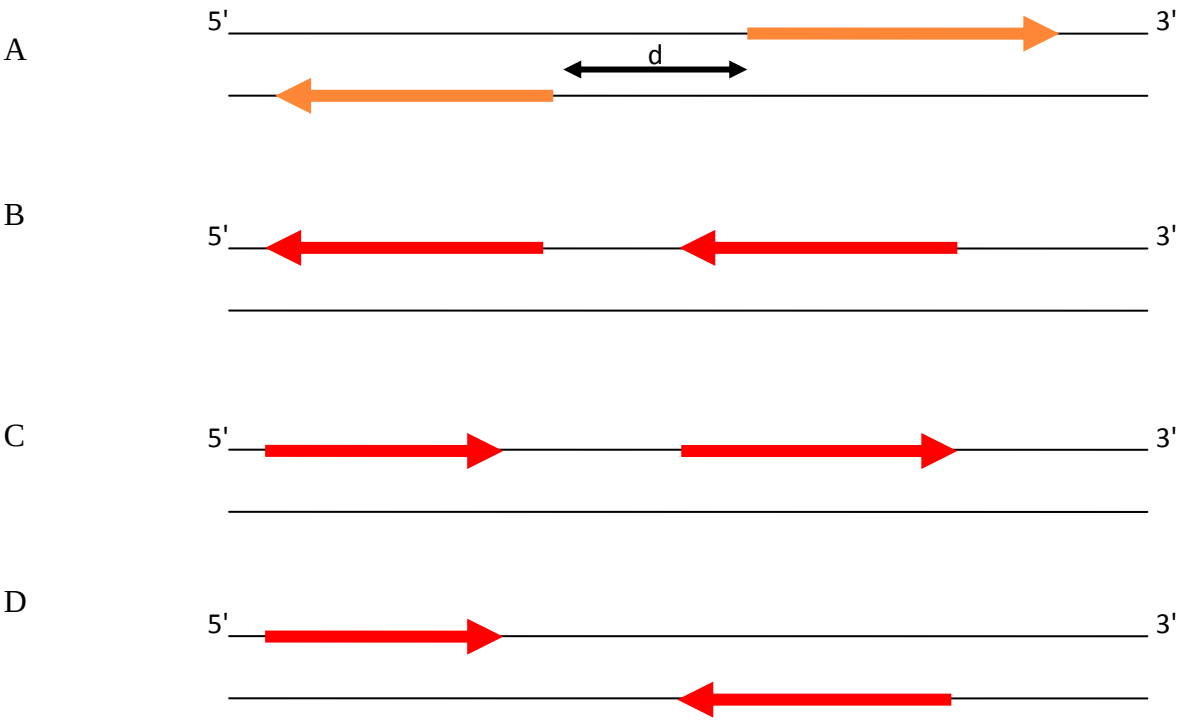


Figure 65 : Cas possible d'agencements des gènes.

Seul le cas "A" a été inclus dans définition opérationnelle de promoteur "bidirectionnel", pour peu que la distance d soit inférieure à 1 kpb.

2.1.2. Abondance chez les métazoaires

Nous avons développé un script permettant de récupérer tous les intergènes satisfaisant cette définition. Ce programme se basant sur le set d'annotation de GeCo, il permit de rechercher les promoteurs de ce type pour tous types de gènes. Nous avons pu retrouver des proportions précédemment identifiées dans la littérature, notamment 847 promoteurs ainsi agencés pour l'humain. Afin d'en étudier l'abondance dans d'autres génomes, nous avons tout d'abord étudié la complétude des annotations de gènes. Parmi les 16 organismes possédant une annotation refGene, pourtant d'excellente qualité chez l'humain, seuls 5 (l'homme, le chimpanzé, la souris, le ver et la drosophile) présentaient un répertoire de gène supérieur à 15000 gènes. Nous avons donc choisi d'étudier ces organismes en y ajoutant le poisson-zèbre aux 14000 gènes annotés puisqu'il s'agissait d'un organisme modèle courant. Nous avons ainsi pu mettre en évidence un taux relativement constant flirtant avec les 5% de gènes impliqués dans des promoteurs bidirectionnels chez les vertébrés. En revanche, chez les insectes et les nématodes, les structuration très particulières des génomes (duplications, opérons...) font atteindre le triple.

Organisme	Génome	Nombre de gènes refGene	Fraction de gènes impliqués dans un promoteur bidirectionnel
Chat	felCat3	437	
Cheval	equCab1	357	
Chien	canfam2	904	
Chimpanzée	**panTro2**	**26846**	**5,00%**
Humain	**hg18**	**26636**	**5,60%**
Macaque	rheMac2	498	
Opossum	monDom4	167	
Ornythorinque	ornAna1	24	
Poisson-zèbre	**danrer5**	**14090**	**4,10%**
Poulet	galGal3	4314	
Rat	rn4	847	
Souris	**mm9**	**21703**	**6,00%**
Vache	bosTau4	10428	
Xénope	xentro2	8797	
Drosophile	**droMer5**	**21158**	**18,90%**
Ver	**ce6**	**24260**	**13,10%**

2.1.3. Distribution des tailles chez les métazoaires

Nous nous sommes intéressés aux distributions fines des taille de promoteurs satisfaisant notre définition chez les métazoaires précédemment sélectionnés. De précédentes études on analysé ce facteur pour tout type de transcription divergente sans limite de taille et sont de fait restées assez floues sur les faibles tailles qui nous intéressent. Nous avons donc utilisé les promoteurs récoltés pour en tracer la distribution de taille par classe et en recenser les effectifs. Il en découle des distribution radicalement différentes chez les métazoaires avec, chez l'humain, un enrichissement aux alentours de 130-150 pb (population P1) et un seconde, plus faible aux alentours de 240-300 pb (population P2) et possiblement une troisième aux alentours de 380-410 bp. Une analyse fonctionnelle réalisée avec le logiciel DAVID a par ailleurs révélé que la population P1 présentait un enrichissement en gènes impliqués dans le phénomène de réparation alors que la population P2 était plutôt impliquée dans la structuration du nucléosome (Figure 66).

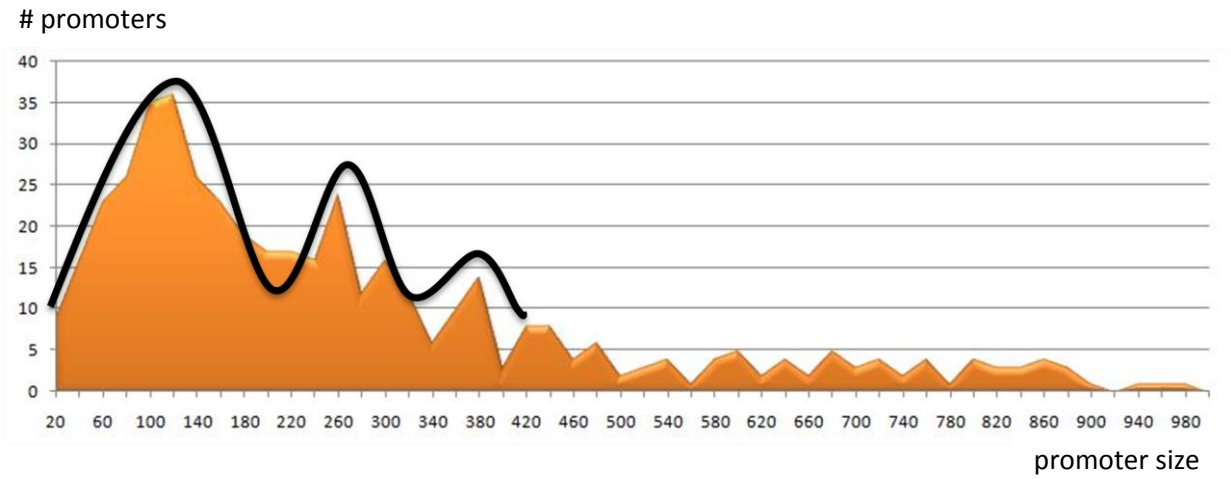


Figure 66 : Distribution des tailles de promoteurs bidirectionnels chez l'humain

Deux à trois populations de tailles semblent émerger et sont centrées autour de 140, 280 et 320 pb, suggérant une apparente fréquence dans la distribution de ces tailles.

Chez le chimpanzé, cette tendance se maintient, alors que chez la souris, les proportions entre les deux populations sont moins marquées. Chez le poisson-zèbre, sans doute du fait du manque d'annotation, il est difficile de conclure à une quelconque tendance. Chez les invertébrés en revanche, il semble au contraire que la population P1 soit proche de néant et que la population P2 se dilue jusqu'à la limite opérationnelle de 1 kpb. Ainsi, selon l'ordre, on

observe une sélection préférentielle de distances pour les promoteurs bidirectionnels. (Figure 67)

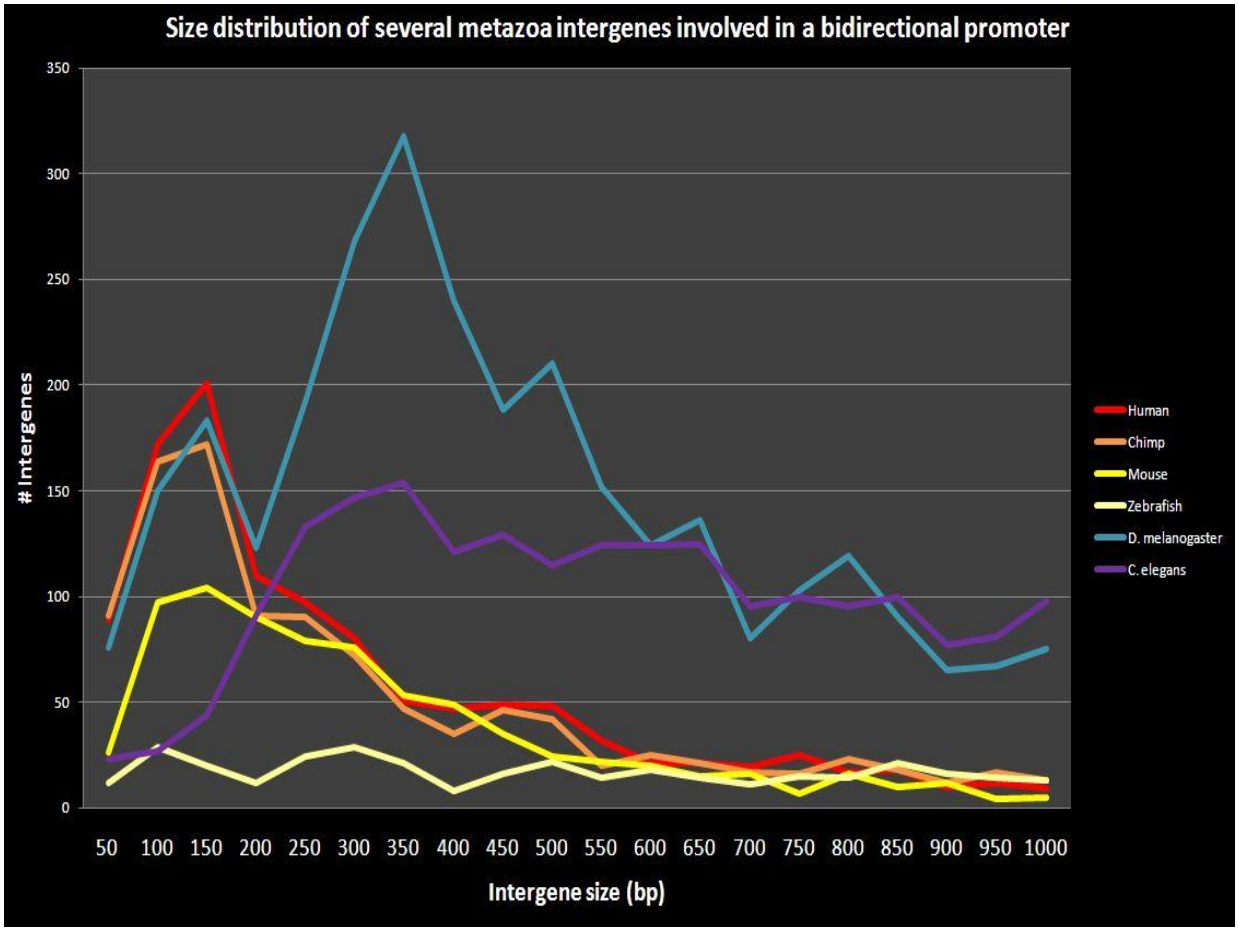


Figure 67 : Abondance et distribution des tailles des promoteurs bidirectionnels parmi 6 métazoaires

Selon les grands ordres, des préférences de taille de promoteurs différentes sont observées. De plus, cet arrangement particulier de gènes semble important chez les invertébrés.

1.1.4. Analyse du contexte informationnel des gènes impliqués dans un promoteur bidirectionnel chez humain

Nous appuyant sur les données fournies par GeCo, nous avons cherché à déterminer des caractéristiques propres aux gènes impliqués dans des promoteurs bidirectionnels et avons estimé les descripteurs CLON de ces gènes. Parmi les densités en vecteurs informationnels à l'intérieur des gènes (Figure 68), on note une recrudescence de la polymérase II, signe de gènes activement transcrits possédant possiblement des variants. De même, le nombre d'îlots CpG est

important (plus de 1 par gène, pouvant aller jusque 6). Ceux-ci sont en général caractéristiques de gènes ubiquitaires (voir Chapitre 2, paragraphe 2.4) et nécessaires au fonctionnement basal de toute cellule, ce qui est cohérent avec notre analyse fonctionnelle mettant en évidence des enrichissements en gènes de réparation et de structuration du nucléosome.

L'apparente fréquence de 150 pb entre les populations nous fit examiner la distance à l'histone modifiée la plus proche. Il s'avère que celle-ci est significativement plus loin que ce que l'on serait en droit d'attendre pour d'autres gènes. Il se pourrait donc que la fréquence de tailles observées corresponde à la taille minimale d'enroulement autour de un à deux histones, possiblement trois et du fait de région activement transcrite, il s'agirait plutôt de l'absence de une à plusieurs histones. Par ailleurs, au regard du descripteur CLON de distance aux autres vecteurs informationnels, il apparait que ces gènes sont également loin de tout SNP ou élément répété. De manière plus surprenante, la zone d'euchromatine la plus proche de ces gènes est généralement retrouvée bien loin plus qu'attendu dans l'ensemble des gènes du génome. (Figure 70)

En étudiant la taille des entités de la carte exonique (Figure 69A), nous avons pu remarquer que les exons de ces gènes étaient beaucoup plus grand que ceux de gènes médians avec une part importante apportée par les régions non-traduites, notamment la 5'-UTR qui occupe une proportion importante de la taille du gène (Figure 69B). Enfin, les introns étaient également relativement grands mais sans que ce soit atypique au regard de GeCo.

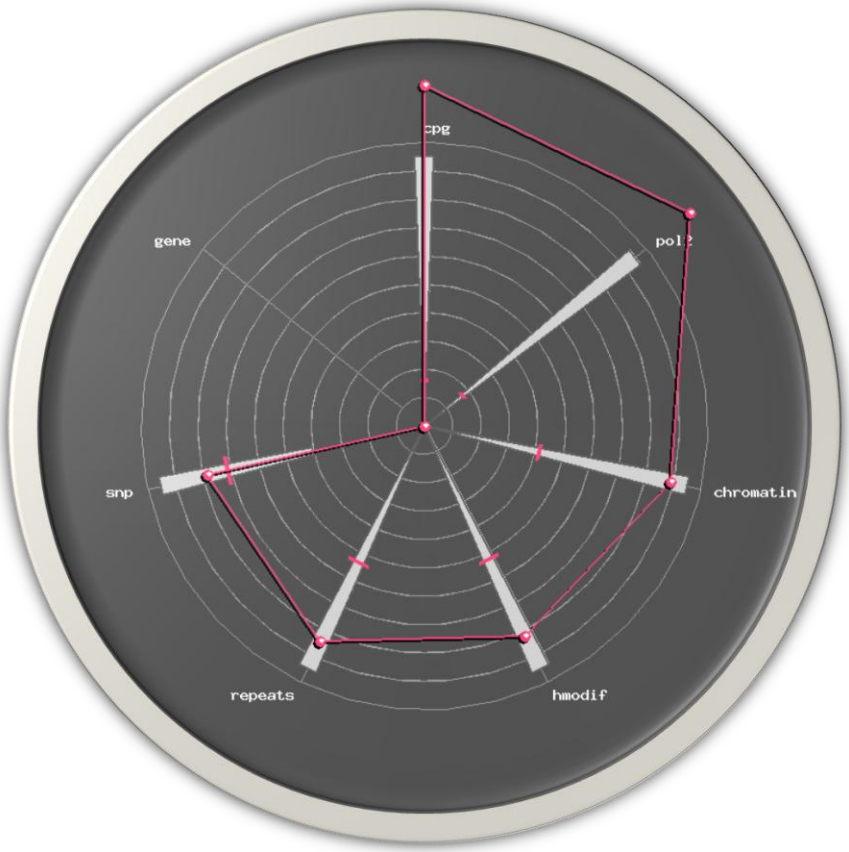
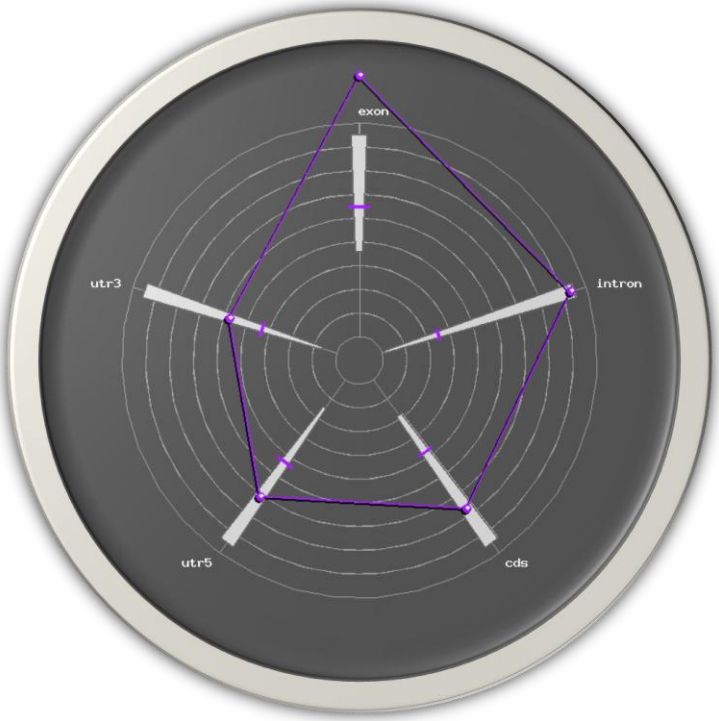
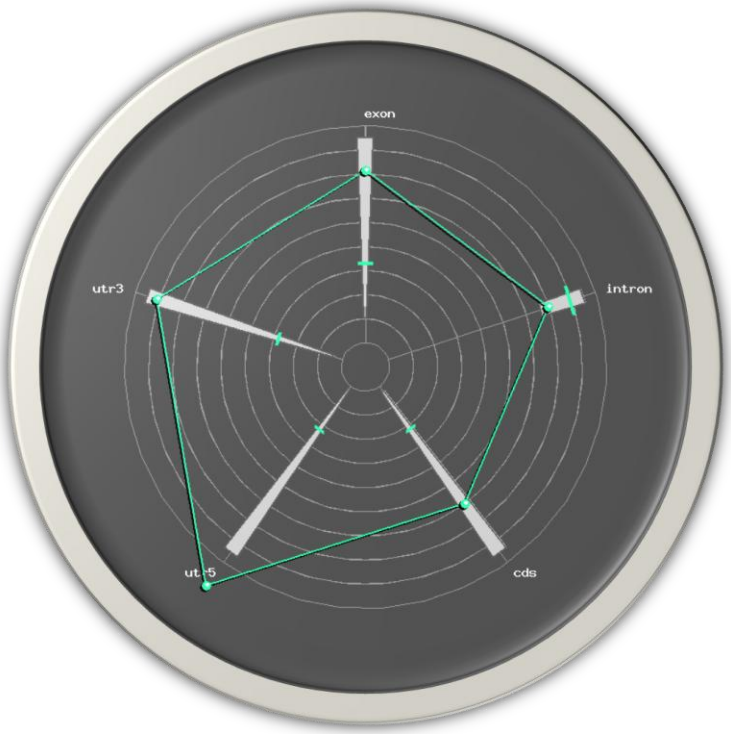


Figure 68 : RBAM représentant les vecteurs informationnels des gènes impliqués dans un promoteur bidirectionnel selon le descripteur CLON de densité (C)

Le nombre de CpG par gène est plus important que celui attendu à l'échelle de l'ensemble des gènes, de même que la polymérase. Globalement le nombre des différents vecteurs informationnels est commun par rapport à l'ensemble des gènes (intérieur de la boite à moustache) avec une légère tendance à s'approcher du troisième quartile donc d'une densité en vecteurs plus importante.



A



B

Figure 69 : RBAM représentant les vecteurs informationnels des gènes impliqués dans un promoteur bidirectionnel selon le descripteur CLON de longueur (L) et d'occupation (O)

A/ La taille des exons est atypique et le décalage de la taille de la 5'-UTR vers Q3 dans une plus forte mesure que la CDS laisse à penser que ce sont les exons de cette partie du gène qui concourent majoritairement à l'atypicité de la taille exonique. B/ L'occupation que fait la 5'-UTR du gène confirme que ce serait bien cette région qui serait très atypique de par sa taille et son occupation du gène.

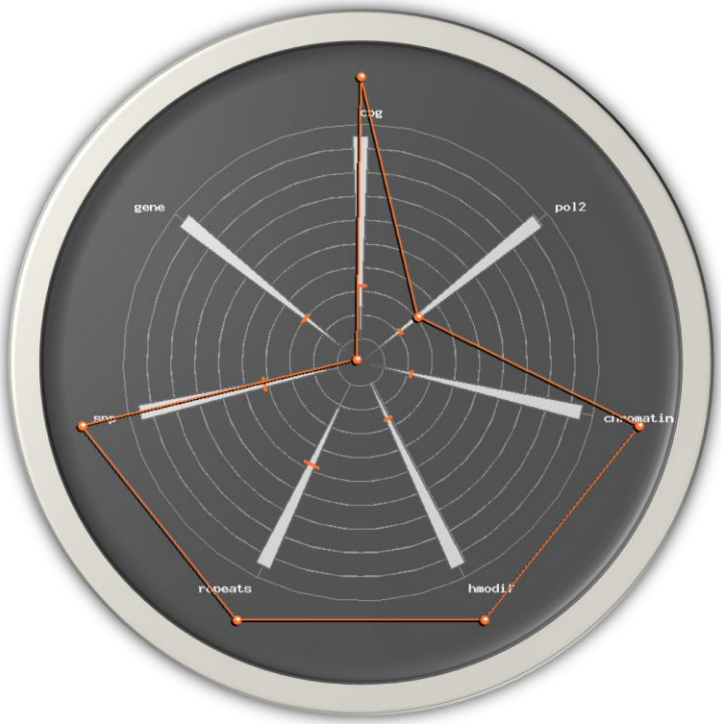


Figure 70 RBAM représentant les vecteurs informationnels des gènes impliqués dans un promoteur bidirectionnel selon le descripteur CLON de distance (N)

Hormis la distance très courte au gène le plus proche du fait de l'agencement en promoteurs bidirectionnels, et de la présence à distance classique de la polymérase, tous les vecteurs informationnels sont , de manière atypique, plus loin qu'attendu sur le génome.

2.2. Abondance de hStaf dans les promoteurs bidirectionnels humains

Nous avons cherché à apprécier la fraction de promoteurs bidirectionnels possédant un SBS prédit. Au sein des 847 promoteurs identifiés, au moins un SBS était présent dans 20% des cas. Ne sachant pas si ce nombre constituait un réel enrichissement, nous avons généré un nombre équivalent de promoteurs unidirectionnels de taille égale un-à-un aux bidirectionnels et répété l'opération 10 fois pour éviter tout biais. Il s'est avéré que globalement, un SBS était présent dans 10% des cas pour les unidirectionnels. Par utilisation du test du Khi2 dédié à la comparaison de deux proportions, nous avons pu valider cet enrichissement en SBS dans les promoteurs bidirectionnels. Enfin, les promoteurs bidirectionnels à SBS avaient en moyenne près de 2 SBS.

Nous avons ensuite cherché à déterminer si la présence d'un ou plusieurs SBS avait un rôle dans l'appartenance à P1 ou P2 mais il semblerait au vu des distributions correspondantes que cela n'ait aucun effet (Figure 71).

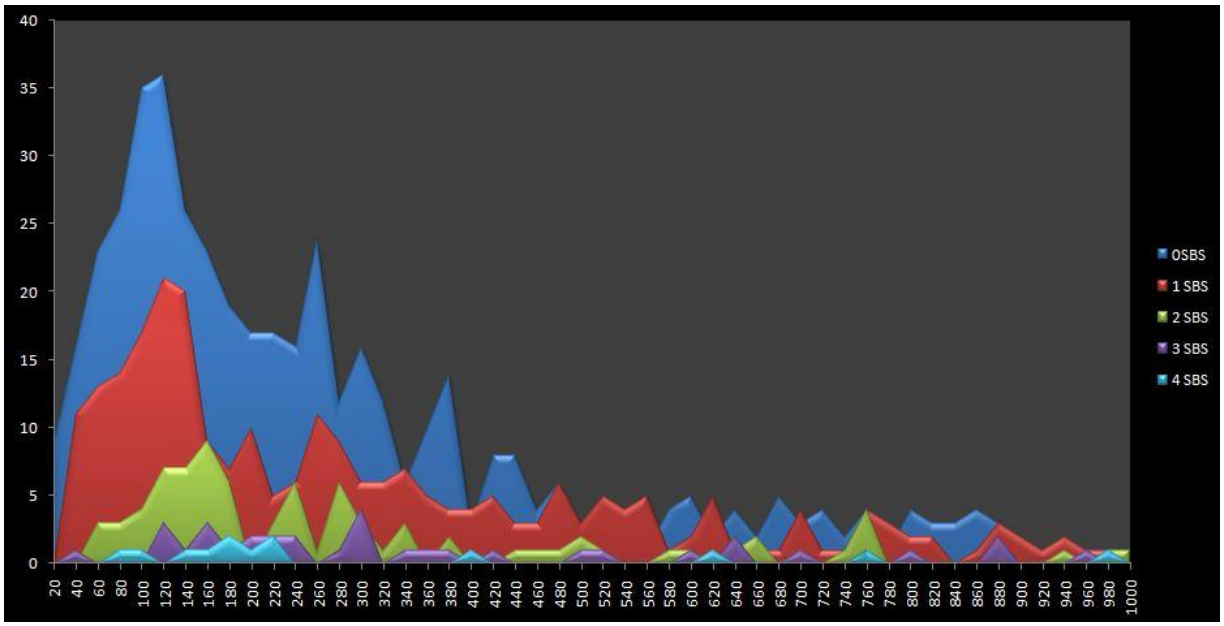


Figure 71 : Distribution de la taille des promoteurs bidirectionnels selon qu'ils aient un, deux, trois ou quatre SBS.

2.2.1. Abondance et raffinage de la séquence ACTACA

Au moment de nos analyses, Xie raffina son modèle et montra que le motif le plus représenté dans les promoteurs conservés de 4 mammifères était (RR)ACTAYR. Ainsi ce motif est plus souple que l'original ACTACA, comme notre étude à l'échelle du génome l'avait suggéré. Nous avons donc intégré cette donnée et recherché ce motif, en lui autorisant une erreur, dans les séquences ne 5' du SBS et il s'est avéré que 22% d'entre elles en possédaient ce motif en 5' signifiant une préférence de ce SBS bipartite pour les promoteurs bidirectionnels.

2.2.2. Validation expérimentale

Parmi les promoteurs bidirectionnels possédant au moins un SBS prédit, 72 ont été testés, se répartissant en 69 paires de gènes protéiques et 3 de gènes d'ARN non-codants. La gamme de promoteurs choisis permit d'inclure des cas présentant jusqu'à quatre SBS prédits. Sur les 72 promoteurs, 70 montrèrent une fixation de hStaf par ChIP. Parmi ces 70 promoteurs, 10 ont été testés en les plaçant entre deux gènes rapporteurs de luciférases différentes et en mutant un ou

plusieurs sites dans les promoteurs. Une activation de la transcription fut démontrée pour 94% des sites. En revanche dans le cas de l'ARNsn U6, un phénomène intéressant se produit. En présence de hStaf, la transcription est activée. Si l'on mute les deux SBS, celle-ci est abolie. En revanche, si l'on mute l'un des deux sites, la transcription est encore meilleure, laissant supposer une dynamique complexe quant à l'occupation optimale de ces sites. Ainsi, la fixation de hStaf et son aptitude à activer la transcription fut démontrée dans la quasi-totalité des promoteurs testés, suggérant une confirmation du rôle majeur prédit pour hStaf dans les promoteurs bidirectionnels et donc possiblement dans l'expression à partir des promoteurs bidirectionnels.

Genome-wide evidence for an essential role of the human Staf/ZNF143 transcription factor in bidirectional transcription

Yannick-Noël Anno*(1,2), Evelyne Myslinski*(2), Alain Krol(2), Olivier Poch(1), Odile Lecompte(1) and Philippe Carbon**(2)

(1): Department of Structural Biology and Genomics, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Institut National de la Santé et de la Recherche Médicale (INSERM), The Centre National de la Recherche Scientifique (CNRS), UMR7104, F-67400 Illkirch, Université de Strasbourg, F-67000 Strasbourg, France

(2): Architecture et Réactivité de l'ARN, Université de Strasbourg, CNRS, IBMC, 15 rue René Descartes, 67084 Strasbourg, France

* These two authors contributed equally to this work

** Address for correspondence: Tel: (33) 3 88417064 ; fax: (33) 3 88602218. E-mail: P.Carbon@ibmc-cnrs.unistra.fr

ABSTRACT

In the human genome, approximately 10 % of the genes are arranged head to head so that their transcription start sites (TSS) are located less than 1 kbp on opposite strands. In this configuration, the TSS are separated by a bidirectional promoter that generally drive expression of the two genes. Here by a combination of in silico and biochemical approaches, we demonstrate that functional binding sites (SBS) for the transcription factor hStaf/ZNF143 are overrepresented in bidirectional versus unidirectional promoters. Furthermore, in some cases, functional SBS are located in bidirectional promoters of gene pairs encoding a noncoding RNA and a protein gene. Chromatin immunoprecipitation assays with a significant set of bidirectional promoters containing putative SBS revealed that more than 97% of them are associated with hStaf/ZNF143. Expression assays and knock-down experiments demonstrated that hStaf/ZNF143 is indeed a transcription factor positively controlling the expression of divergent protein-protein and protein-noncoding RNA gene pairs. Furthermore, we also showed that hStaf/ZNF143 per se exhibits an inherently bidirectional transcription activity.

INTRODUCTION

Despite the vast genomic space, a substantial fraction of human genes (11%) are arranged as bidirectional gene pairs in a head to head, divergent fashion and controlled by bidirectional promoters (1,2). A bidirectional promoter was defined as an intergenic region containing less than 1kb of DNA which is flanked by the transcription start sites of two genes in opposite directions(1,2). Thus the closely located bidirectional pairs are overrepresented in the human genome and this abundance has been observed across several mammalian genomes (3). For a vast majority of human bidirectional promoters where the associated genes have mammalian orthologs the bidirectional arrangement is conserved suggesting that it is functionally important (1,2). Indeed, the expression pattern of divergent gene pairs are more correlated than those of randomly paired genes (2). Most of the bidirectional promoters lack TATA element, are GC rich and exhibit enriched colocalization with CpG islands (2,4). Bidirectional promoters display also a mirror sequence composition with an overrepresentation of G and T on one side of the promoter, and C and A on the other, on the genomic plus strand (5). Divergent promoters are enriched in binding site consensus sequences of a small group of transcription factors including GABPA/NRF2, NRF1, NFY and YY1 while the majority of known vertebrate binding motifs are underrepresented in bidirectional promoters (6).

The work described in this manuscript establishes that the human transcription factor Staf (hStaf, also called ZNF143) acts also as an esssential factor in transcription from bidirectional promoters. Staf, originally identified in Xenopus laevis as the transcription activator of the $tRNA^{Sec}$ gene, also controls expression of snRNA and snRNA-type genes (7-9). Furthermore, hStaf/ZNF143 has been reported to regulate transcription of the SCARNA2 gene and of ten protein coding genes (10-13) and references therein). Lastly, a genome-wide analysis led us to identify 1175 hStaf/ZNF143 binding sites (SBS) distributed in 938 mammalian protein gene promoters, strongly suggesting that hStaf/ZNF143 is involved in the transcriptional regulation of a very large number of protein coding genes. Seven contigous zinc fingers of the C2-H2 type contain the Staf DNA binding domain and only zinc fingers 1 to 6 establish base-specific contacts with 18 bp in Staf-DNA complexes (14-16). In the present study, we conducted bioinformatic and biochemical analysis at the whole genome scale which demonstrated that functional SBS are overrepresented in bidirectional compared to unidirectional

promoters. Furthermore, in some cases, functional SBS are located in promoters of pairs comprising a noncoding RNA (nc) and a protein coding gene (c). Chromatin Immunoprecipitation assays (ChIP) on 72 representative bidirectional promoters containing 69 c-c and 3 c-nc putative SBS revealed that more than 97% of them are associated with hStaf/ZNF143. Using a combination of luciferase and RTase assays, and hStaf silencing, we demontrated that hStaf/ZNF143 is clearly involved in the bidirectional expression directed by the promoters of 10 protein-protein (c-c) and 3 protein-ncRNA gene pairs (c-nc). Furthermore we also shown that hStaf/ZNF143 per se exhibits an inherently bidirectional transcriptional activity.

Materials and Methods

Identification of bidirectional gene pairs

Bidirectional gene pairs were identified by skimming through the gene annotation set constituted by Genomic Context database (GeCo) (YNA et al manuscript in preparation) and extracting all pairs of genes in inverted orientations separated by less than 1kbp, regardless of their gene type. The Genomic Context database is related to human genome NCBI build 36 (hg18) and was built by computing 'refGene' (proteins), 'rnaGene' (snRNA, snoRNA, tRNA, rRNA, scaRNA), kgXref tables from the University of Santa Cruz California (UCSC, www.genome.ucsc.edu), 'mirna', 'mirna_literature_references', 'mirna_mature' and 'literature_references' tables from Sanger Institute and piRNA's file from the piRNAdatabank. It was implemented locally in a high-speed DB2 architecture called BIRD (Biological Integration and Retrieval of Data), allowing to quickly address the whole gene set. For each bidirectional promoter, 10 randomly chosen unidirectional promoters of the same size were extracted from GeCo, assessing that a gene with no other gene in the vicinity of its TSS (1kbp) is a unidirectional promoter.

Sequence analysis and identification of potential SBS in promoters

Prediction of Staf binding sites in bidirectional and unidirectional promoters was performed with a position-specific scoring matrix (PSSM) built using 347 experimentally validated binding sites (11). Matrix generation and scan of sequences were done using the Genomatix suite, with MatDefine and MatInspector programs (17) respectively. Significant overrepresentation of SBS in bidirectional

promoters was evaluated using the khi² test for two proportions, with alpha=0.05 and 1 degree of freedom (df). The test was repeated 10 times on corresponding random sets (Supplemental Table 1) and assessed in 100% of the cases a significantly higher rate of predicted SBS in bidirectional promoters.

Reporter constructs

Thirteen human bidirectional promoters (extended intergenes EI1 to EI13) ranging from 229 to 751 bp were PCR amplified from human genomic DNA (list of primers and chromosomal localization available as Supplemental Table 2) and cloned into pGEMTeasy (Promega). BamHI and SacI sites were included in the forward and reverse primers, respectively, used for PCR amplification of EI11 and EI12. KpnI sites were included in the primers used in EI13 amplification. The dual luciferase reporters constructs containing EI1 to EI10 were obtained as follows. EcoRI fragments isolated from pGEMTeasy constructs harboring EI1 to EI10 were subcloned into EcoRI-cut pFRLN50 dual luciferase reporter vector generated by insertion of the NcoI linker CATGGGAGCTCGAATTCGAGCTC in the NcoI site of the pFRL vector (18). The luciferase reporter constructs containing the EI11 and EI12 fragments were generated by subcloning the BamHI-SacI and NotI-SacI fragments isolated from pGEMTeasy-EI11 and EI12 into BamHI-SacI and NotI-SacI cut pFlashI luciferase reporter vector (SynapSys), respectively. The KpnI fragment isolated from pGEMTeasy-EI13 was subcloned in KpnI-cleaved pU6/Hae/RA.2/EcoRV construct (19,9). The resulting BamHI-Sac1 fragment containing EI13, followed by a 137 bp spacer derived from the β-globin gene, was subcloned into BamHI-SacI cut pFlashI luciferase reporter vector. PCR primers were designed so that the PCR products containing the EI1 to EI11 and EI13 promoters do not contain the AUG initiation codon of the endogenous gene. Instead, translation initiation is ensured by the AUG of the luciferase reporter genes. For the luciferase construct containing the EI12 birectional promoter, the endogenous AUG of the C19orf6 gene was in frame with the AUG of the firefly luciferase. MaxiU6 genes were created by inserting the GACCTCGAGGCGGTTC sequence at position 87 of the wt U6. In all mutant versions of the SBS, the NNCCCR sequence (N standing for any nucleotide, R for A or G) at positions 1-6 of the SBS was replaced by GGTTTC. The dual luciferase vector with part of the BUB1B promoter containing two SBS and the associated ACTACAA motif was prepared as in

Myslinski et al. (13). The 271 bp DNA fragment (Ch15 : 38240250-38240520, in hg18) was amplified by PCR and inserted as an EcoRI fragment into the EcoRI-cut pFRLN50. All mutants were generated with the QuickChange II XL site-directed mutagenesis kit (Stratagene) and verified by DNA sequencing.

ChIP assay and semi-quantitative PCR analysis

The ChIP procedure and the PCR analysis was performed as described in Myslinski et al. (11) using a rabbit polyclonal antibody against a C-terminal epitope of Staf. For the negative control, we used the PP1 to PP4 primer pairs hybridizing to unique regions located at 2.4, 2.1, 6.5 and 2.5 kbp upstream of the tRNASec, U4 ATAC, GAPDH and BUB1B genes and generating PCR products of 235, 211, 174, 226 bp, respectively. The primer sequences used in this study are available on request.

Transfection and bidirectional promoter activity assay

HeLa cells (5 x 10^3 cells/well in 96 well plates) were cotransfected using Lipofectamine 2000 with 200 ng of each experimental luciferase construct (containing wt or mutated EI1 to EI13 promoters) and 100 ng of the pCH110 internal control plasmid. After 24h, cells lysates were prepared with passive lysis buffer (Promega) and assayed for β-galactosidase and luciferase activities. The firefly and *Renilla* luciferase assays (promoters EI1 to EI10) were performed using the Dual-Luciferase TM Reporter assay system (Promega) on a GloMax™ 96 Plate Luminometer (Promega). The firefly luciferase assay (promoters EI11 to EI13) was performed using the Luciferase assay system (Promega). The luciferase activities were normalized to the β-galactosidase activity. Each transfection experiment and luciferase assay was done at least in triplicate. For U6 snRNA and RPPH1 (H1RNA) gene expression analysis, HeLa cells (9 x 10^5 cells) were cotransfected using Lipofectamine 2000 with 8.4 μg of wt or mutant luciferase constructs containing the EI11 to EI13 promoters with 600 ng of maxi5S RNA as internal standard. Cells were collected after 48h and total RNA was extracted using TRIREAGENT (Euromedex). Total RNA was analyzed by primer extension of two labeled oligonucleotides, one complementary to positions +88/+104 of the maxiU6 (human wt U6 gene numbering) and the other to positions +112/+129 of the maxi5S. The extended products were separated on a 6% denaturing gel and quantitated with a Fuji Bioimage Analyzer. The yield of extended maxiU6 was normalized to that of extended maxi5S. Each transfection experiment and

reverse transcriptase assay was done at least in triplicate. H1 RNA gene expression from the EI13 promoter was monitored as described in Myslinski et al. (9).

Transfection of siRNA, reverse transcription and RT-qPCR

The hStaf/ZNF143-specific siRNAs GCUGGAAGAUGGUACCACAGCUUAU (siRNA1), GGGCAUUUGCCAGUGCAACAAAUUA (siRNA2), GGAACGCACUCUGUUGCUAUGGUUA (siRNA3) or control-siRNA (Invitrogen) were transfected into HeLa cells using Lipofectamine 2000. according to the maufacturer's instructions (Invitrogen). Total RNA was isolated using TRIREAGENT (Euromedex) from cells harvested 72h post transfection, treated by RNase-free DNase and reverse transcribed using oligo (N)9 primer. cDNA was amplified by qPCR with gene specific primers on a Stratagene Mx3005P PCR system (Agilent Technologies) using EvaGreen qPCR Mix Plus (Euromedex). We used Primer3 software (http://frodo.wi.mit.edu/primer3/) to design primers so that the final amplicon was 100-150 bp. The primer sequences used in this study are available on request. All reactions were carried out in triplicate. Relative gene expression was calculated using the $\Delta C\tau$ method following the manufacturer's instructions.

Results

Promoter size and genes associated in bidirectional promoters

We extracted from the Genomic Context database (GeCo) pairs of human genes that are the closest neighbors on opposite strands and with their transcription start sites (TSS) separated by less than 1 kbp (see Materials and Methods). In this search, we included only those genes whose transcripts were not predicted to overlap at the 5' ends. As defined by Trinklein et al. (2), we considered the region between the two TSS as a putative bidirectional promoter. Therefore, the constituted set contained 842 bidirectional promoters and 1684 genes (Supplemental Table 3). Different interesting features regarding the size of the bidirectional promoters, distribution of the sizes and the length of the genes in the divergent gene pairs arose from our study. First, considering the size of the bidirectional promoters, Figure 1 shows the number of bidirectional promoters with a definite length plotted versus their length. We found a bias toward a small size: 83% of the promoters harbor a size less than 500 bp.

Second, the distribution of the bidirectional promoter size is not uniform. Indeed differents pools are present with a preferential size of 120-140 bp and multiples (240-280 bp and 360-420 bp) of it. Third, a bias toward a small size of genes associated to divergent promoters in comparison to unidirectional promoters was observed: 40.1 kbp versus 112.8 kbp (Supplemental Table 3). The bidirectional promoter set is constituted at 95% (801 out of the 842) of protein coding gene pairs (c-c gene pairs). The remainder is distributed into 2 small groups of gene pairs: 29 combine a ncRNA gene with a protein coding partner (nc-c gene pairs), and 12 contain ncRNA genes only (nc-nc gene pairs). All members of the nc-nc subgroup are entirely constituted of two tRNA genes. tRNA genes were also identified in 17 of the nc-c pairs. The ncRNA genes present in the other 12 nc-c gene pairs are: let-7i, mir-34b, mir-320, MRP-RNA, H1-RNA, SRP-RNA, 7SK RNA, scaRNA17 (U91), U13 snoRNA, U6.2 snRNA, U6.9 snRNA and U12 snRNA (Supplemental Table 3). Thus the closely located divergent gene pairs and the associated bidirectional promoter harbor a biased small size suggesting a putative functional importance of this particular feature.

Sequence analysis and Staf Binding Site identification in bidirectional promoters

The hStaf/ZNF143 Staf binding site (SBS) consists of a 18 bp sequence with a highly conserved consensus sequence CCCR at positions 3 to 6, a highly conserved C at position 12 and a more degenerate sequence at positions 1-2, 7-11 and 13-18 (8,14,11). The strategy to identify SBS in bidirectional promoters was the following. We collected a pool of 347 binding sites experimentally validated by ChIP (Myslinski et al., 2006) and converted them into a position-specific scoring matrix (PSSM) using MatDefine programs of the Genomatix suite (17) (Supplemental Figure 1). We then used MatInspector to search matrix matches in the sequences of intergenes and in the first 400 bp of the transcribed sequences of two genes partners. The search was extended to the first 400 bp of the genes because it is well established that the first exon and 5'part of the first intron of genes contain transcription factors binding sites (TFBS) (20). With an optimal cutoff score of 0.82, the search identified 602 SBS above the cutoff and which are distributed in 394 potential bidirectional promoters (394 out of 842: 46.7% of the bidirectional promoters) (Figure 3B, Supplemental Tables 3 and 4). Among these 394 pairs, 13 combine a ncRNA gene with a protein coding partner (nc-c gene pairs) and two combine

two ncRNA genes (nc-nc gene pairs) (Supplemental Table 5). Among the 13 pairs, six of them contain a tRNA gene and 2 harbor a miRNA gene (let-7i, mir-320). In the other 5 gene pairs, the ncRNA gene partner is the RNase MRP RNA, RNase P RNA (H1RNA), SRP-RNA, snRNAU6.2 and snRNAU6.9. The two nc-nc gene pairs contain each two tRNA genes, $tRNA^{Gly\text{-}TCC}$-$tRNA^{Trp\text{-}CCA}$ and $tRNA^{Arg\text{-}TCG}$-$tRNA^{Arg\text{-}CCT}$ (Supplemental Table 5).

Among 126 (21%) of the 602 identified SBS, inspection of sequences adjacent to the identified SBS revealed the presence, immediately upstream of the SBS, of the 9-bp RRACTAYRN motif or a one bp variant (Figure 2B). Among the 394 promoters identified as containing an SBS, 204 (51.7%) harbored at least one SBS in the intergene, the percentage being 24.2% (204/842) when considering the whole set of bididirectional promoters. A control set was generated by random extraction from unidirectional promoters of a bp number equivalent to the total bp number in the intergenic region of the 842 bidirectional promoters. The search of SBS in this and nine other different control sets showed that 10% of unidirectional promoters contain at least one SBS. Together, these results indicate that hStaf/ZNF143 binds to at least 47% of the bidirectional promoters and that the SBS is overrepresented in bidirectional promoters with a 2.4 fold enrichment compared to unidirectional promoters.

In vivo occupancy of bidirectional promoters by hStaf/ZNF143

To determine whether hStaf/ZNF143 is indeed associated in vivo with the bidirectional promoters identified as containing putative SBS, a chromatin immunoprecipitation assay was performed in HeLa cells. We examined 176 putative SBS with scores ranging from 0.82 to 0.99 (listed in Supplemental Table S6, scores in ST4), contained in 92 (23%) of the 394 promoters identified as containing SBS. The 92 promoters tested (89 pairs of protein coding genes only and 3 of protein coding/noncoding genes) were extracted with a random choice from the bidirectional promoter set (Supplemental Table S4). After immunoprecipitation with an antibody specifically recognizing hStaf/ZNF143, enrichment of the promoters was monitored by semi-quantitative PCR amplification with primers amplifying the regions surrounding the putative SBS. The specificity of the ChIP reaction was monitored by PCR amplification of four promoters recognized by hStaf/ZNF143 (Figure 2C. $tRNA^{Sec}$, synaptobrevin-

like1, aldehyde reductase and t-complex polypeptide 1 promoters as positive controls) and of four DNA fragments which do not harbor any hStaf/ZNF143 binding site (Figure 2D, negative controls). Each DNA sequence was tested with each of the three templates obtained from anti-hStaf/ZNF143, control ChIP and input chromatin. We tested two dilutions of DNA isolated with (Figure 2, lanes 1 and 2; Supplemental Figure 2A) or without antibody (Figure 2, lanes 3 and 4; Supplemental Figure 2A). In addition, a serial dilution of the input material was analyzed to demonstrate that the PCR was quantitative within a linear range of amplification (Figure 2, lanes 5-7; Supplemental Figure 2A). As expected, the positive controls yielded a signal of intensity higher with anti-hStaf/ZNF143 than in the no-antibody control (Figure 2C, compare lanes 1,2 and 3,4). In contrast, no specific signal could be obtained with the primer pairs PP1 to PP4 amplifying DNA sequences lying several kbp upstream of the $tRNA^{Sec}$, U4 ATAC, GAPDH and BUB1B genes because these remote regions were not expected to interact with hStaf/ZNF143 (compare lanes 1,2 and 3,4 in Figure. 2D). Among the 92 PCR amplifications, 72 yielded interpretable results (Figure 2E, F and Supplemental Figure 2A). Two amplifications (I536 and I841, Figure 2F) were close to background level but the remaining 70 promoters provided clear positive signals (Figure 2E, Supplemental Figure 2A and Supplemental Table 6). These experiments shows that 97% (70 out of 72) of the bidirectional promoters tested did harbor genuine hStaf/ZNF143 binding sites. In an additional control, we amplified from ChIP samples a DNA region corresponding to bidirectional promoters identified in silico but lacking SBS. Of the ten promoters that were tested, none provided positive amplification (Supplemental Figure 2B), suggesting that the hStaf/ZNF143 sequence is necessary for protein binding in vivo. Taken together, these results demonstrate the robutness of the computational screens and reveal the high prevalence of bona fide direct targets of the hStaf/ZNF143 transcription factor in bidirectional promoters.

Functional activity of the hStaf/ZNF143 binding site in bidirectional expression of protein gene pairs

In a first attempt, the ability of ten of the bidirectional promoters (associated with hStaf/ZNF143) to initiate transcription from both directions was assessed. These were EI1 to EI10 and are constituted of pairs of protein coding genes only (Table 1 and Supplemental Table 6). DNA fragments containing the

full intergenic region lying between the transcription start sites (TSS) and parts of the first exon of the protein genes were inserted into a dual reporter vector (8) in which the transcription activities in the two opposite directions could be tested simultaneously via the readout of the firefly and *Renilla* luciferases (Table 1, Figure 3 and Supplemental Figure 3). HeLa cells were transiently transfected with the different constructs and the luciferase activities of the resulting cells extracts was measured. Analysis of the firefly and *Renilla* luciferase activities, normalized to that of the β-galactosidase control (Figure 3 and Supplemental Figure 3), showed that all the tested constructs bore bidirectional activity. The transcription activity depending on the promoter tested was 9.2 to 100-fold in the *Renilla* and 6 to 160-fold in the firefly luciferase direction that of the promoter-less empty vector (Figure 3 and Supplemental Figure 3). The in silico data found that the fragments inserted in the dual luciferase reporter vector contain 1-3 SBS, with 17 SBS in total (Supplemental Table 4) . However, visual inspection of the sequence revealed the presence of 7 additional putative SBS, in fragments EI1, EI3, E16, EI9 and E110, that were unlisted by the bioinformatic screen (SBS with a score below 0.82; SBS1 in EI1, SBS1 and SBS2 in EI3, SBS4 in EI6, SBS3 and SBS4 in EI9, and SBS2 in EI10; Supplemental Figure 3). To determine whether the 24 SBS (17 +7) play a role in bidirectional transcription, the effect of substitutions at positions 1-6 in the SBS core sequence NNCCCR was tested on expression of the reporter gene. Such substitutions are known to totally abolish formation of the DNA-protein complex (8,13). 22 of the 24 SBS identified in the ten promoters were mutated singly, and simultaneously mutations of both SBS1 and SBS2 were engineered in EI1, EI4 and EI7 where they reside in the same promoter (Supplemental Figure 3). All single mutations had an effect on transcription efficiency. Five substitutions altered transcription unidirectionally whereas it was affected bidirectionally for 17 of them (77%) (Supplemental Figure 3 and Table 1). In the ten promoters tested, at least one SBS was involved in bidirectional control of transcription. For 16 of the 17 mutants with bidirectionally-altered transcription, the promoter activity decreased simultaneouly in the two directions demonstrating that these SBS are involved in up-regulating both genes in the pair. Interestingly, in the case of SBS1 in EI7, the SBS mutation decreased promoter activity in the COL4A3BP direction and increased it in the POLK direction (Table 1 and Supplemental Figure 3). In the five substitutions leading to unidirectional alteration of transcription, three of them exhibited a

unidirectional decrease: SBS3 in EI3, AHSA1 direction; SBS1 and SBS2 in EI9, ENY2 direction; and two a unidirectional increase: SBS1 in EI5, STRADB direction SBS1 in EI6, ZMAT5 direction (Table 1 and Supplemental Figure 3). The transcriptional status of the mutants containing two simultaneously mutated SBS (SBS1 and SBS2 in EI1, EI3 and EI7) essentially recapitulates the effects observed wih the single mutants. Taken together, these results demonstrate the functional importance of the hStaf/ZNF143 binding sites in bidirectional transcription of divergent genes.

The simultaneous presence of two SBS in the promoters of the C19orf6-U6.2 and MED16-U6.9 gene pairs down-regulated expression of the U6.2 and U6.9 snRNA genes

The hStaf/ZNF143-sites containing activity of three bidirectional promoters associating a protein and a ncRNA gene was tested by transient transfection assays. The gene pairs tested were MED16-U6.9, C19orf6-U6.2 and PARP2-RPPH1. The U6.9 and U6.2 ncRNA genes encode U6 snRNA (21), RPPH1 coding for H1 RNA, the RNA component of the RNase P (EI11 to EI13 in Table 1). To monitor the expression of these promoters, constructs were engineered as follows. For MED16-U6.9 (EI11) and C19orf6-U6.2 (EI12), we transfected constructs containing part of the first exon of MED16 or C19orf6, the full intergenic region between the transcriptional start sites, and the U6.2 or U6.9 genes. In these constructs, parts of the protein coding genes are placed in front of the firefly luciferase reporter (EI12 and EI11 in Figure 3 and Supplemental Figure 3), and a 16 bp fragment was inserted into both U6 genes to distinguish the transiently expressed U6 snRNAs from the endogenous genes. For PARP2-RPPH1, the construct contained part of the first exon of PARP2 and the full intergenic region placed between the luciferase reporter and a chimeric gene consisting of a 137 bp spacer derived from the β-globin gene followed by an efficient RNA polymerase III termination site (19,9). Expression in the PARP2 direction was analyzed after transfection by the luciferase assay, that toward RPPH1 gene by RNase protection assay of an antisense RNA probe and normalization to the expression of an α-globin mRNA included as the internal standard. The results of these experiments clearly establish that the three promoters possess a bidirectional transcription activity (Figure 3 and Supplemental Figure 3). To further address whether the SBS identified in silico within these promoters (1-3 SBS in PARP2-RPPH1, C19orf6-U6.2 and MED16-U6.9) are indeed involved in the bidirectional

activity, we examined again the effects of SBS substitutions on transcription activity. Mutation of the single SBS in PARP2-RPPH1 decreased expression in both directions (Supplemental Figure 3). Substitution of either SBS1 or SBS2 in MED16-U6.9, C19orf6-U6.2 decreased expression in the direction of the protein coding genes but, surprisingly, increased expression in the U6 direction. As expected, however, the simultaneous mutations of both SBS dramatically decreased the promoter activity in the U6 snRNA direction (Figure 3 and Supplemental Figure 3). Taken together, our data indicate that the three PARP2-RPPH1, C19orf6-U6.2 and MED16-U6.9 intergenic regions act as bidirectional promoters and that the simultaneous presence of two SBS in C19orf6-U6.2 and MED16-U6.9 clearly down-regulated the expression of the U6.2 and U6.9 genes.

A DNA fragment containing an SBS, with or without the associated RRACTACAN motif, is sufficient for bidirectional promoter activity

Having shown that hStaf/ZNF143 binding and bidirectional transcription activity are correlated, we asked whether an SBS-containing DNA fragment isolated from a promoter driving unidirectional transcription is sufficient to lead to bidirectional transcription. To do this, we selected a 271 bp fragment from the BUB1B promoter that contains 2 functional SBS associated to the ACTACAA motif that was identified in our previous studies (13). HeLa cells were transiently transfected with constructs containing wt or mutant versions of the BUB1B promoter fragment inserted into a dual luciferase reporter vector. Figure 4 shows that the wt fragment actually possessed bidirectional promoter activity. However, the simultaneous mutations of the two ACTACA and SBS motifs disabled totally the promoter, demonstrating that the bidirectional activity is directly due to the presence of the SBS and/or associated ACTACA motifs (Figure 4). The bidirectional activity dropped 14-fold following the simultaneous mutation of the two SBS whereas that of both ACTACAA motifs led to a 2-fold reduction (Figure 4). Lastly, we tested the transcription capacity of a construct containing only one SBS without the associated ACTACA motif. Figure 4 shows that the presence of the sole SBS motif is sufficient for bidirectional activity. These results establish without ambiguity that the SBS is an element with sufficient intrinsic ability to direct efficient transcription activity in the direct and reverse orientations.

Knock-down of hStaf/ZNF143 leads to impaired expression from bidirectional gene pairs

Finally, we wished to examine the functional importance of the hStaf/ZNF143 transcription factor in controling the expression of gene pairs in vivo. To this end, the abundance of the corresponding mRNAs was measured in HeLa cells cells where hStaf/ZNF143 was knocked-down by RNAi. Cells were treated with a mixture of siRNAs and the actual reduction of the hStaf/ZNF143 protein level was monitored by Western blot. A decrease of more than 70% of hStaf/ZNF143 was obtained 72h post-transfection; this effect was specific because no change was observed with siRNA-control treated cells (Figure 5A). The hStaf/ZNF143 mRNA level also dropped specifically to more than 70 % in siRNA-treated cells (Figure 5B). The steady-state level of mRNAs produced by expression of the eleven gene pairs C11orf10-FEN1, KNTC1-RSRC2, C14orf133-AHSA1, TMEM186-PMM2, TRAK2-STRADB, ZMAT5-UCRC COL4A3BP-POLK, DOCK4-ZNF277, NUDCD1-ENY2, ZNF189-MRPL50 and RPPH1-PARP2 was measured by RT-qPCR in 72h treated cells (Figure 5C). The bidirectional promoters in these gene pairs are contacted by hStaf/ZNF143 in vivo and gene expression was dependent on SBS integrity in transient transfection assays. Four independent RNAi experiments indicated that the mRNA levels obtained from the 11 gene pairs decreased from 63 to 79 % in hStaf/ZNF143-depleted cells versus non-depleted cells (Figure 5C).

We conclude from these experiments that (i) variation in the mRNA levels from all the analyzed genes, following hStaf/ZNF143 depletion, is consistent with the data obtained from in vivo functional assays (ChIP and luciferase assays); (ii) hStaf/ZNF143 positively controls expression of divergent gene pairs.

Discussion

We have performed a human genome scale analysis to evaluate the binding of transcription factor hStaf/ZNF143 to bidirectional promoters and to define the functional significance of its recrutment in the expression of divergent gene pairs. Our findings that hStaf/ZNF143 binds to at least 47% of the bidirectional promoters in the human genome, and that Staf Binding Sites (SBS) are 2.4-fold overrepresented in these promoters compared to unidirectional ones, constitute strong evidence that

hStaf/ZNF143 is an essential regulator of bidirectional transcription. Direct experimental validation by ChIP on 78 promoters yielded 90% success, indicating that very few of the identified sites were false positives. Importantly, the high rate of success demonstrates the robustness of the bioinformatic screen using the PSSM established from a pool of 347 SBS that were previously experimentally validated (11). Furthermore, our bioinformatic screen showed that 21% of the identified SBS were associated with the 9bp functional RRACTACAN motif. The transcription level of constructs harboring only the RRACTACAN motif is very low. Paradoxically, mutation of RRACTACAN motif in the presence of a wt SBS yielded a 50% drop of transcription efficiency, suggesting an important role of this motif, but only in the context of the SBS. It may well be that this motif serves as a binding site for an unknown transcription factor. It might also induce a particular DNA local structure enhancing the affinity of hStaf/ZNF143 to its cognate sequence. The RRACTACAN sequence is particulary interesting in the light of the work describing the discovery of motifs overrepresented in human bidirectional promoters (6). This analysis yielded five motifs, including the NRF-1, GABPA, YY1 and NFY transcription factor binding sites, and the fifth one of sequence ACTACANNTCCC with no known identified transcription factor. It was predicted that the ACTACANNTCCC sequence and the cognate transcription factor would play an important role in regulating bidirectional transcription. Our work established clearly that ACTACAN indeed corresponds to the 3' part of the RRACTACAN motif that we identified and that NTCCC represents SBS residues 1-5. Lin et al. (6) did not report residues 6-18 of the SBS as overrepresented in bidirectional promoters. This is very likely because of the lower sequence conservation in this part of the SBS. A similar motif with the ACTAYRNNNCCCR consensus sequence was previously reported by Xie et al. (22) who ranked it fourth among 174 motifs in terms of conservation across several mammalian transcription factors. In the present work, we showed that a significant part (20%) of the identified SBS is linked to the RRACTACAN motif. If this motif is the target for an unknown transcription factor, we propose that this factor is also an essential actor working in conjunction with hStaf/ZNF143 to regulate bidirectional transcription.

We have found that a few number of gene pairs associate a protein coding gene with a noncoding RNA gene. In two of the pairs, the ncRNA partner is a U6 snRNA gene. In the human genome, five

functional U6 snRNA genes have been identified (21) and we showed here that the U6.2 and U6.9 genes are associated to a protein coding gene in a bidirectional promoter containing two functional SBS. We had previously demonstrated that RNA polymerase III transcription of the human U6.1 gene requires Staf, and that the transcriptional activation is performed via a single SBS (8). In the case of the bidirectional promoters, the mutation of either of the two SBS caused a reduction of the transcription activity in the direction of the protein gene. This result was expected as is the simultaneous mutation of each SBS which reduced transcription activity in both directions. However, and much to our surprise, the mutation of one single SBS, whatever it was, led to enhanced U6 expression. This surprising result must be examined in the context of the snRNA gene promoter architecture. In fact, all the snRNA gene promoters characterized so far contain only one single SBS (23) and our report is the first describing the presence of two SBS in the promoter of an snRNA gene. In the light of our data, it looks as if the presence of two SBS is incompatible with the optimal formation of transcription complexes and is detrimental to U6 expression.

In conclusion, we have shown that hStaf/ZNF143 binds and governs expression from a majority of bidirectional promoters. The presence of a single SBS is sufficient to direct bidirectional activity and the intrinsic ability of hStaf/ZNF143 makes it an ideal factor to govern bidirectional promoters. In an earlier work, we had demonstrated that hStaf/ZNF143 binds and controls expression from a large number of unidirectional promoters (11). The picture emerging from our current and previous studies strengthens the role of hStaf/ZNF143 as an essential factor for controling gene expression in humans.

Acknowledgments

We are grateful to H.T. Jacobs (University of Tampere, Finland) for the gift of the pFRLN50 reporter vector and M. Frugier for anti-AspRS. We also thank C. Graber and T. Strub, two undergraduate students, for their involvement in the preparation of several constructs, S. Baudrey and A. Schweigert for valuable technical assistance. The Ligue Inter-Régionale Contre le Cancer Grand-Est is thanked for

a funding support. Y.N.A is a recipient of an Allocation de Recherche from the Ministère de l'Enseignement et de la Recherche.

References

1. Adachi, N. and Lieber, M.R. (2002) Bidirectional gene organization: a common architectural feature of the human genome. *Cell*, **109**, 807-809.
2. Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P. and Myers, R.M. (2004) An abundance of bidirectional promoters in the human genome. *Genome Res*, **14**, 62-66.
3. Koyanagi, K.O., Hagiwara, M., Itoh, T., Gojobori, T. and Imanishi, T. (2005) Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene*, **353**, 169-176.
4. Yang, M.Q. and Elnitski, L.L. (2008) Diversity of core promoter elements comprising human bidirectional promoters. *BMC Genomics*, **9 Suppl 2**, S3.
5. Engstrom, P.G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S.L., Yang, L. *et al.* (2006) Complex Loci in human and mouse genomes. *PLoS Genet*, **2**, e47.
6. Lin, J.M., Collins, P.J., Trinklein, N.D., Fu, Y., Xi, H., Myers, R.M. and Weng, Z. (2007) Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res*, **17**, 818-827.
7. Schuster, C., Myslinski, E., Krol, A. and Carbon, P. (1995) Staf, a novel zinc finger protein that activates the RNA polymerase III promoter of the selenocysteine tRNA gene. *Embo J*, **14**, 3777-3787.
8. Schaub, M., Myslinski, E., Schuster, C., Krol, A. and Carbon, P. (1997) Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III. *Embo J*, **16**, 173-181.
9. Myslinski, E., Ame, J.C., Krol, A. and Carbon, P. (2001) An unusually compact external promoter for RNA polymerase III transcription of the human H1RNA gene. *Nucleic Acids Res*, **29**, 2502-2509.
10. Gerard, M.A., Myslinski, E., Chylak, N., Baudrey, S., Krol, A. and Carbon, P. The scaRNA2 is produced by an independent transcription unit and its processing is directed by the encoding region. *Nucleic Acids Res*, **38**, 370-381.
11. Myslinski, E., Gerard, M.A., Krol, A. and Carbon, P. (2006) A genome scale location analysis of human Staf/ZNF143-binding sites suggests a widespread role for human Staf/ZNF143 in mammalian promoters. *J Biol Chem*, **281**, 39953-39962.
12. Gerard, M.A., Krol, A. and Carbon, P. (2007) Transcription factor hStaf/ZNF143 is required for expression of the human TFAM gene. *Gene*, **401**, 145-153.
13. Myslinski, E., Gerard, M.A., Krol, A. and Carbon, P. (2007) Transcription of the human cell cycle regulated BUB1B gene requires hStaf/ZNF143. *Nucleic Acids Res*, **35**, 3453-3464.
14. Schaub, M., Krol, A. and Carbon, P. (1999) Flexible zinc finger requirement for binding of the transcriptional activator staf to U6 small nuclear RNA and tRNA(Sec) promoters. *J Biol Chem*, **274**, 24241-24249.
15. Schaub, M., Myslinski, E., Krol, A. and Carbon, P. (1999) Maximization of selenocysteine tRNA and U6 small nuclear RNA transcriptional activation achieved by flexible utilization of a Staf zinc finger. *J Biol Chem*, **274**, 25042-25050.

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

17. Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. and Werner, T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933-2942.

18. Zanotto, E., Shah, Z.H. and Jacobs, H.T. (2007) The bidirectional promoter of two genes for the mitochondrial translational apparatus in mouse is regulated by an array of CCAAT boxes interacting with the transcription factor NF-Y. *Nucleic Acids Res*, **35**, 664-677.

19. 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.

20. Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet*, **8**, 424-436.

21. Domitrovich, A.M. and Kunkel, G.R. (2003) Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. *Nucleic Acids Res*, **31**, 2344-2352.

22. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338-345.

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

Figure legends

Figure 1. Features of bidirectional promoter sizes in the human genome. The length of the bidirectional promoter is plotted versus the frequency of its occurrence.

Figure 2. Characterization of bidirectional promoters that are target for hStaf/ZNF143 and detection of promoter occupancy in vivo by chromatin immunoprecipitation (ChiP). **A**. Outline showing identification of the target promoters. The results of the ChIP and PCR assays performed on 92 promoters (89 protein coding pairs and 3 protein coding-noncoding RNA pairs) are indicated. **B.** SBS motif analysis with sequence logos depicting nucleotide distribution in the SBS (left panel) and in the SBS with the 5' associated submotif (right panel) **C-F**. Binding of the endogenous hStaf/ZNF143 to genomic sites in HeLa cells was analyzed by ChIP. Genomic DNA fragments, recovered from input material or immunoprecipitated with hStaf/ZNF143 antibody or no-antibody, were subjected to semi-quantitative PCR amplification in the presence of (α^{32}P)-dCTP with specific primer pairs. Lanes 1,2 and 3,4: serial dilutions of DNA immunoprecipitated with anti-hStaf/ZNF143 or no-antibody (No Ab), respectively. Lanes 5-7: serial dilutions of input material. Lane 8 (NC, negative control): PCR lacking the chromatin DNA template. Positive (C) and negative (D) controls for ChIP assays. Promoters used are indicated on the right of panel C. PCR products generated with the PP1-PP4 primer pairs originate from unique regions 2.4, 2.1, 6.5 and 2.5 kbp upstream of the $tRNA^{Sec}$, U4 ATAC, GAPDH and BUB1B genes, respectively. Typical positive and the two negative results obtained from bidirectional promoter regions are depicted in panels E and F, respectively. Bidirectional promoters are indicated by gene names.

Figure 3. The human TMEM186-PMM2 and C19orf6-snRNAU6.2 intergenic regions act as bidirectional promoters. Mutational analysis of the promoter identified the functional importance of the SBS in bidirectional transcription. **A.** Characteristics of the extended intergenes EI4 and EI12 (length and name of genes in pair), intergenes (name and length) and SBS (position, orientation and score) are indicated. **B.** Sequences of the EI4 (TMEM186-PMM2) and EI12 (C19orf6-snRNAU6.2)

regions cloned into the dual and firefly luciferase reporters, respectively. The sequence of the intergene is underlined, the SBS are highlighted in gray and the U6 gene is in italics. **C**. Schematic diagrams of the reporter genes used in the transfection assays. Mutations (-) are indicated; the relative transcription activities of the different constructs are indicated, representing either the luciferase activity/empty vector ratio or the relative activity in the reverse transcription assay for the U6 snRNA expression. The yield of extended maxiU6 was normalized to that of the extended maxi5S added as the internal standard. Values are mean +/-SE (error bars), n = 3 independent experiments with two replicates/group.

Figure 4. The Staf Binding Site (SBS) bears the capacity to direct bidirectional transcription.

Schematic diagrams of the reporter genes used in transfection assays and dual luciferase activities. The BUB1B promoter fragment -464/-194 (relative to the translation initiation codon) containing two SBS (SBS1 and SBS2) and two ACTACCA motifs was inserted into the dual luciferase vector. Boxed are ACTACAN (light gray) and SBS (solid box) elements. mut ACTACAA (open box) and mut SBS (open box) constructs contain ACTAC to TATGG and CCCA to TTTC substitutions, respectively. The relative transcription activities of the different constructs are indicated, representing the firefly and *Renilla* luciferase activity/empty vector ratios. Values are mean +/- SE (error bars), n = 3 independent experiments with two replicates/group.

Figure 5. Effects of the hStaf/ZNF143 knock-down on the mRNA level produced by divergent gene pairs. **A**. hStaf/ZNF143 protein was knocked-down in HeLa cells by RNAi. Cells, either treated with RNAi control or treated with hStaf/ZNF143 siRNAs were harvested 72h after transfection. Total cellular protein were subjected to Western blotting with polyclonal antibodies against a C-terminal epitope of hStaf/ZNF143 or anti-AspRS as the internal control. **B.** Total RNA was extracted from 72h treated and control cells, and relative quantification of hStaf/ZNF143 mRNA was carried out by quantitative RT-PCR. The bar indicated the relative content of mRNA normalized to U3 snoRNA (endogenous control) in treated with respect to control cells, fixed as 1. Values are expressed as a ratio and results are mean +/- SD (n=8). Statistical analysis was performed using paired two tailed Student's

t test. **C.** Relative mRNA quantification by quantitative RT-PCR in eleven divergent gene pairs. The measures were performed for each partner in eight pairs, and for one of the partners only in the three gene pairs C14orf133-AHSA1, DOCK4-ZNF277 and PARP2-RPH1. The undetermined partner is marked as nd, not determined. Values are expressed as in **B.**

Legend to Table 1

Characteristics of the 13 bidirectional promoters analyzed and effects of the SBS substitutions on the direction of transcription. EIn: name of the extended intergene. The intergene name refers to Supplemental Table 3. The numbers of SBS identified and SBS mutated in the intergene are mentioned. The effect of SBS substitution on the direction of transcription is indicated.

b: bidirectional effect with decrease in both directions

u^d (XXXX): unidirectional effect with decrease in the direction of gene in ( )

u^i (XXXX): unidirectional effect with increase in the direction of gene in ( )

b*: decreased and increased efficiencies in COL4A3BP and POLK directions, respectively

b**: decreased and increased efficiencies in protein gene and U6 directions, respectively

Supplemental data

Supplemental Figure 1

A. SBS probability matrix, as displayed by MatDefine. Each base occurrence of a 347 experimentally-validated training set is shown (11) and probabilities are calculated. Nucleotides are named using the International Union of Pure and Applied Chemistry (IUPAC) nomenclature. Ci: conservation index vector as defined in Cartharius et al. (17). This matrix serves as the basis to calculate a position specific scoring matrix. **B.** Profile of the related Staf binding site and its IUPAC consensus. The nucleotides in capital letters denote the core sequence.

Supplementary Figure 2

In vivo detection of promoter occupancy by hStaf/ZNF143 using ChIP and PCR analysis on bidirectional promoters. Figure S2A shows 72 promoters identified in silico to contain SBS with

positive and negative PCR results (results listed in Supplemental Table 6). Figure S2B shows negative PCR results on 10 bidirectional promoters identified in silico as devoid of SBS. Lanes 1,2 and 3,4: serial dilutions of immunoprecipitated DNA with anti-hStaf/ZNF143 or no-antibody, respectively. Lanes 5-7: serial dilutions of input material were analyzed to demonstrate that conditions were within the linear range of PCR amplification. Lane 8 (NC, negative control): PCR lacking chromatin DNA template. The accession numbers of the bidirectional promoters examined are indicated on the left.

Supplemental Figure 3

Functional analysis of 13 intergenic regions containing SBS. **A.** Characteristics of the extended intergenes EI1 to EI13 (length and names of gene in pair), intergenes (name and length) and SBS (position, orientation and score) are indicated. **B.** Sequences of the regions cloned into the dual (EI1 to EI10) or firefly luciferase reporters (EI11 to EI13). The sequence of the intergene is underlined, the SBS are highlighted in gray, the ACTACAN motif in direct or reverse orientation is in bold. The sequence of the U6 gene in EI11 and EI12 is indicated in italics. **C**. Schematic diagrams of the reporter genes used in the transfection assays. Mutations (-) are indicated. The relative transcription activities of the different constructs are indicated, representing the luciferase activity/empty vector ratio. The relative expression of U6 snRNA (EI11 and EI12) was measured by reverse transcription, that of H1RNA (EI13) by an RNase protection assay. The yields of extended maxiU6 and β-globin were normalized to those of the extended maxi5S and α-globin added as the internal standards, respectively. Values are mean +/- SE (error bars), n = 3 independent experiments with two replicates/group.

Supplemental Table 1

Control set of undirectional promoters

Supplemental Table 2

Characteristics of the 13 bidirectional promoters analyzed. EIn: name of the extended intergene. Name of intergene (In): extracted from Supplemental Table 3. PCR primers pair: primers used in EIn

amplification from genomic DNA. The length and chromosomal localization (hg18) of the cloned fragment are indicated.

Supplemental Table 3

The 842 bidirectional promoters identified in this study with the features of the intergene: chromosome localization, size, gene type and partners in the gene pairs. The presence of SBS (cutoff above 0,82) is indicated with chromosome localization, number, strand, SBS sequence, score. The sequences -9 to -2 upstream, 50 bp upstream and 50 bp downstream of the SBS are indicated.

Supplemental Table 4

The 603 SBS identified in the bidirectional promoters with the intergene, gene partners and SBS features.

Supplemental Table 5

The 23 SBS identified in the bidirectional promoters linking noncoding RNA-protein coding genes and two noncoding RNA genes. The intergene, gene partners and SBS features are indicated.

Supplemental Table 6

Characteristics of the 92 bidirectional promoters tested in ChIP and results of the assay (+, - positive and negative results. pp: no results in PCR assay). The 13 bidirectional promoters used in functional assays are indicated.

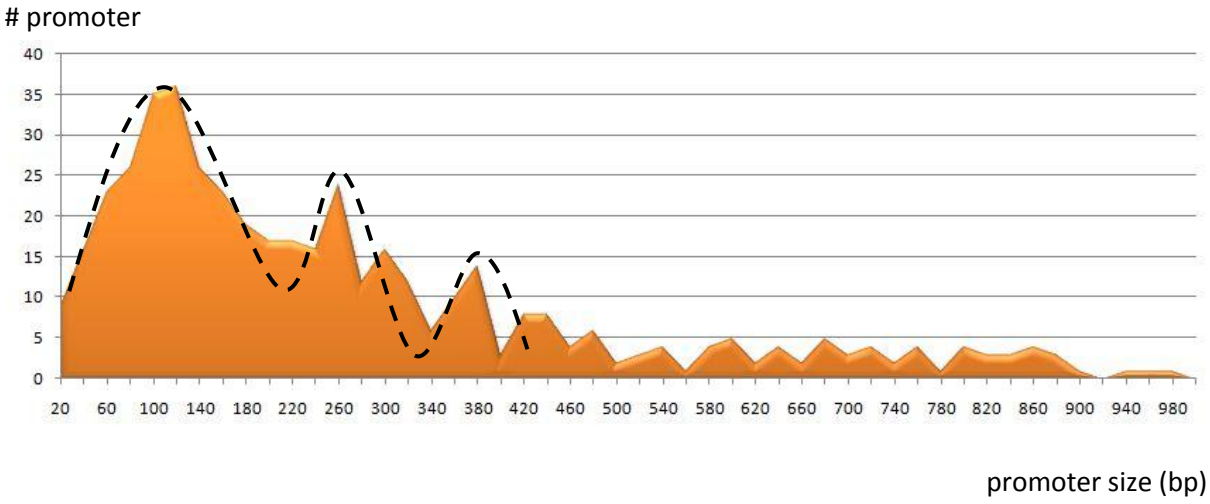
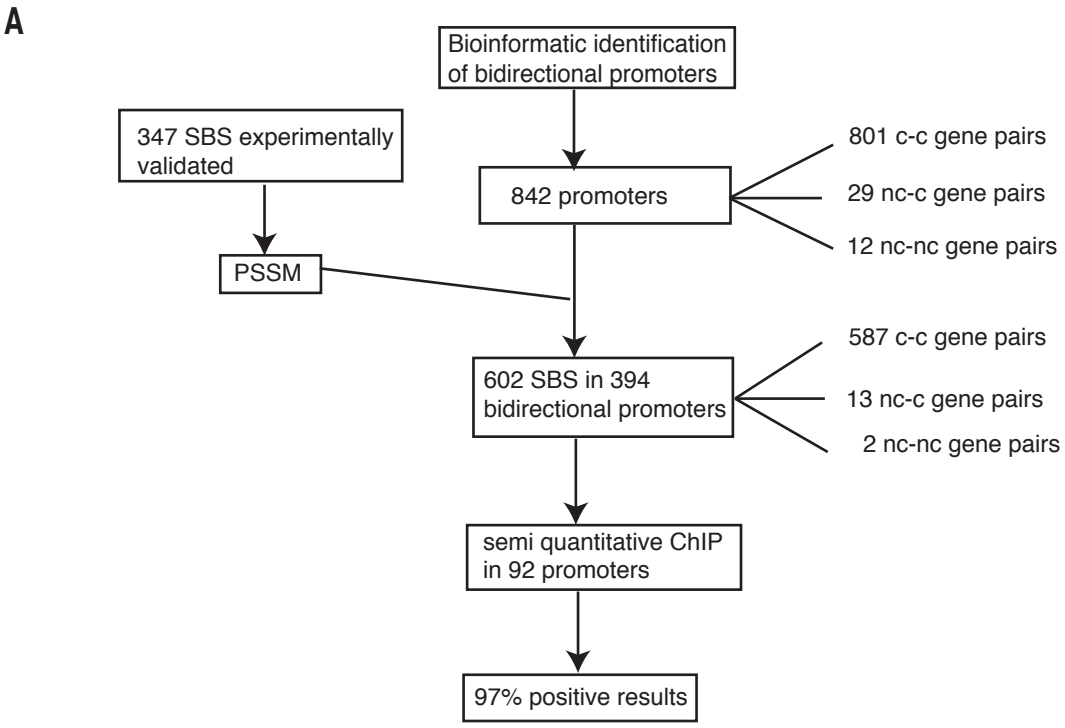
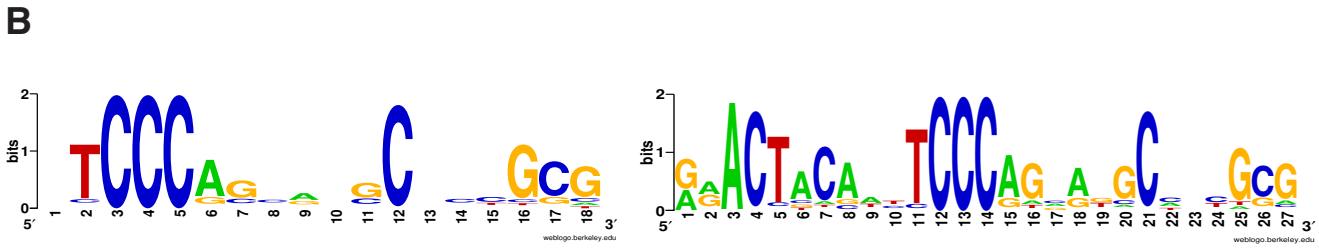
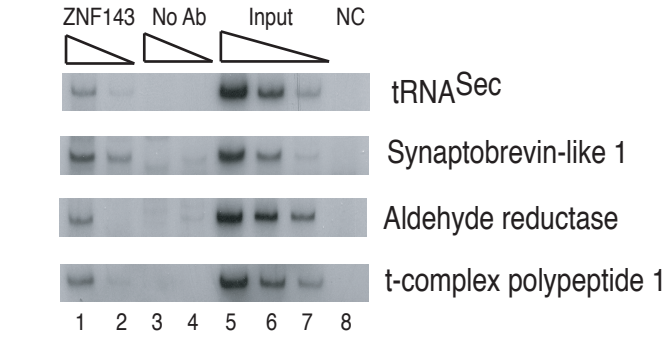
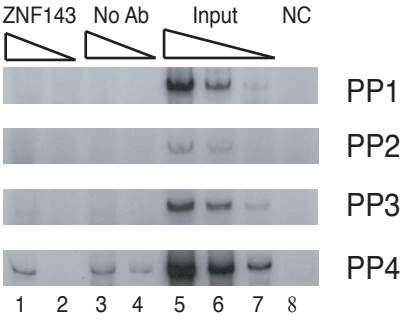

promoter
40
35
30
25
20
15
10
5
0
20 60 100 140 180 220 260 300 340 380 420 460 500 540 580 620 660 700 740 780 820 860 900 940 980
promoter size (bp)

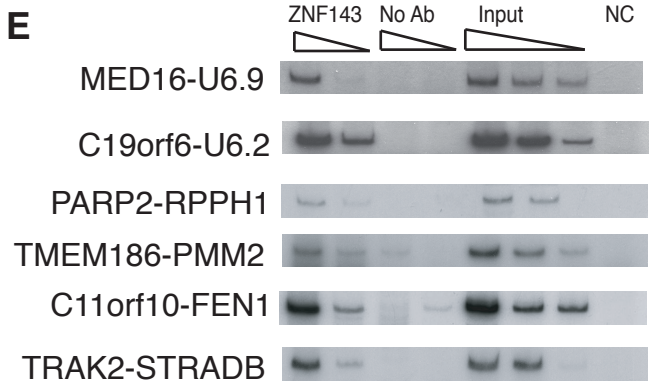
Figure 1


A
Bioinformatic identification of bidirectional promoters
347 SBS experimentally validated
842 promoters
801 c-c gene pairs
29 nc-c gene pairs
12 nc-nc gene pairs
PSSM
602 SBS in 394 bidirectional promoters
587 c-c gene pairs
13 nc-c gene pairs
2 nc-nc gene pairs
semi quantitative ChIP in 92 promoters
97% positive results


B
bits


C
ZNF143
No Ab
Input
NC
tRNASec
Synaptobrevin-like 1
Aldehyde reductase
t-complex polypeptide 1
1 2 3 4 5 6 7 8


D
ZNF143
No Ab
Input
NC
PP1
PP2
PP3
PP4
1 2 3 4 5 6 7 8


E
ZNF143
No Ab
Input
NC
MED16-U6.9
C19orf6-U6.2
PARP2-RPPH1
TMEM186-PMM2
C11orf10-FEN1
TRAK2-STRADB

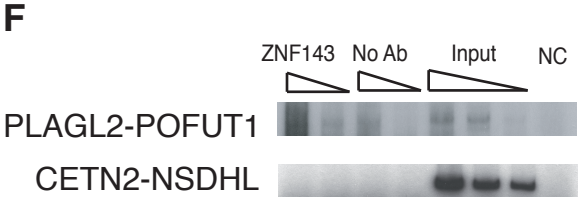

F
ZNF143
No Ab
Input
NC
PLAGL2-POFUT1
CETN2-NSDHL

Figure 2

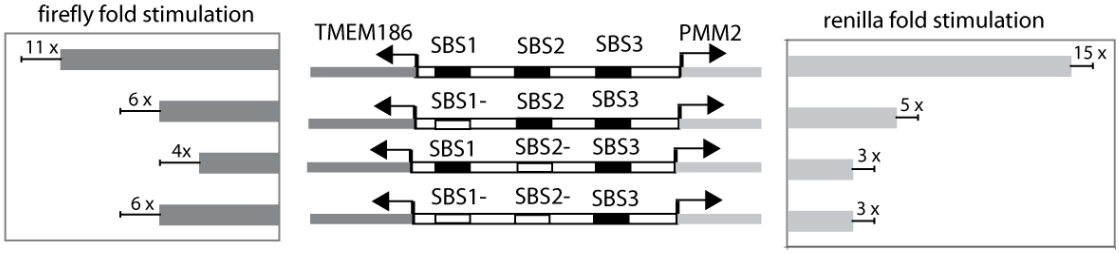
A

Extended intergene	Genes in pair	intergene	SBS	position	strand	score
EI4	**TMEM186-PMM2**	I298	SBS1	89-106	-	0.909
268 bp		201 bp	SBS2	137-154	+	0.846
			SBS3	171-187	+	0.871

B

1-GGCCCCACCACCTGCTGCCGGAAGTAAATCCCACCGGCCCCCGGAACCCCTAAGAGCGCATGC
TCGAATGTACAAGGCGGGCGTGATCTGCGTTGCACCCTGGGAGTTGCGGTCCAGGAATCGTGGCTG
CCGCTACTCCCAGGCGTTATGGGAACGGAGTCCCCTCCTCTTCCCGACGTGCCCTGCGACTCAGCG
GCCGAACCCGGAAGTTCCGGGCCGAGTTCCTCGTGCCAACGTGTCTTGTAAGGTGCGGCTAGAAAC
TGGGGAC-268

C



A

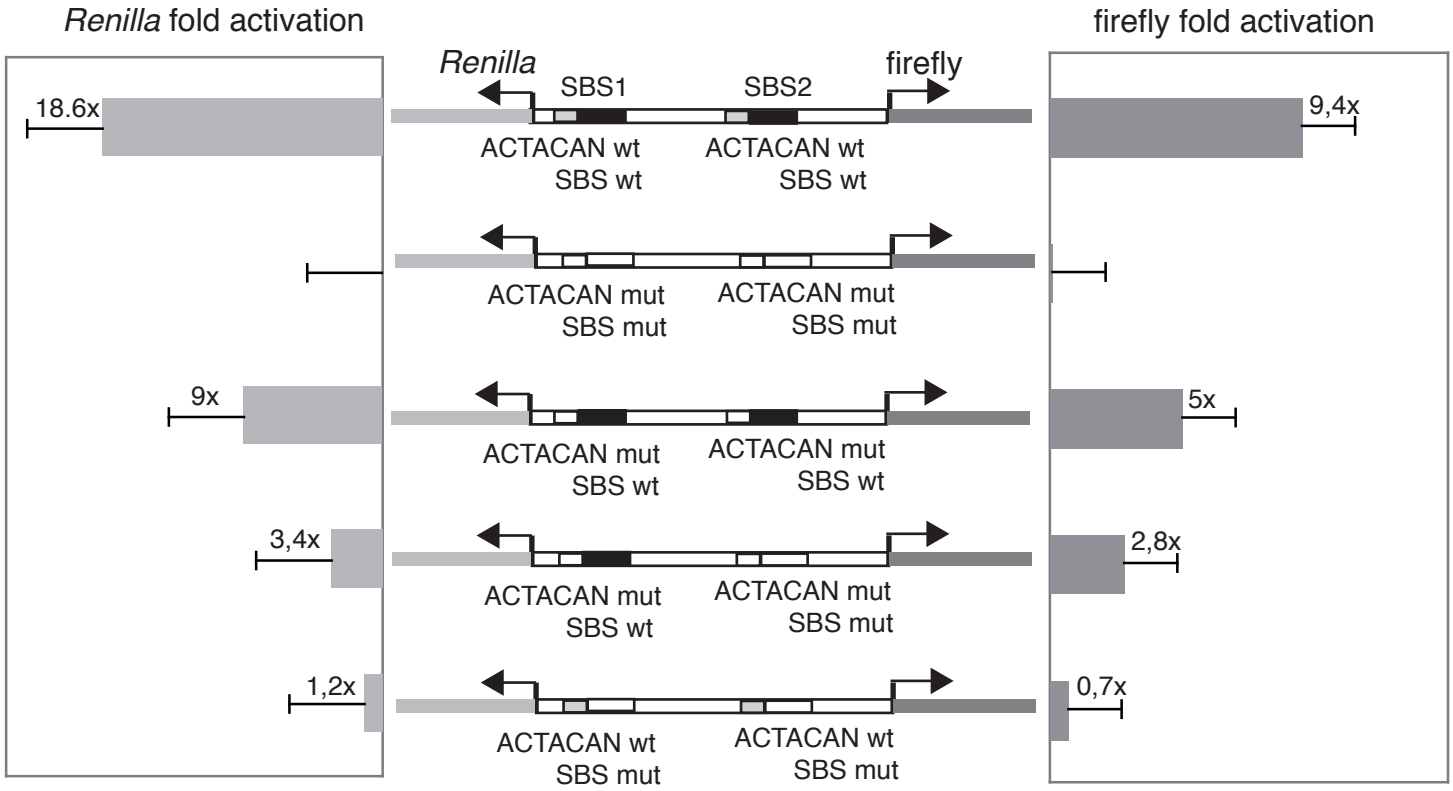
Extended intergene	Genes in pair	intergene	SBS	position	strand	score
EI12	**C19orf6-snRNA U6.2**	I419	SBS1	81-98	-	0.912
523 bp		379 bp	SBS2	288-305	+	0.855

B

1-ATCTTCCGCTTTCTCGTCCGGCTGCGGCGCTGCTGACGCTAGCGAGTCGCCACGCCGGGCAAG
AGCGGCCCCCCTGCGCCCGCAGAGAACGCTGGGATGCCAGCGGCGCCCGCGGAGGCCTCACCCCCT
ACCTCGGCCGCTCCAGGGGGCGGGCCTGCATCTGGGCCACCTCTTTTGCATATTGGCACCCACAAT
CCACCGCGGCTATGAGGCCAGTATAAGGCGGTAAAATTACGATAAGATATGGGATTTTACGTGATC
GAAGACATCAAAGTAAGCGTAAGCACGAAAGTTGTTCTGCAACATACCACTGTAGGAAATTATGCT
AAATATGAAACCGACCATAAGTTATCCTAACCAAAAGATGATTTGATTGAAGGGCTTAAAATAGGT
GTGACAGTAACCCTTGAGTC*GTGCTCGCTTCGGCAGCACATATACTAAAATTGGAACGATACAGAG*
*AAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAATTCGTGAAGCGTTCCATATTTTT*GCT-
523

C

Figure 3



ACTACAN wt- SBS wt

ACTACAN mut- SBSwt

ACTACAN wt-SBS mut

ACTACAN mut-SBS mut

Figure 4

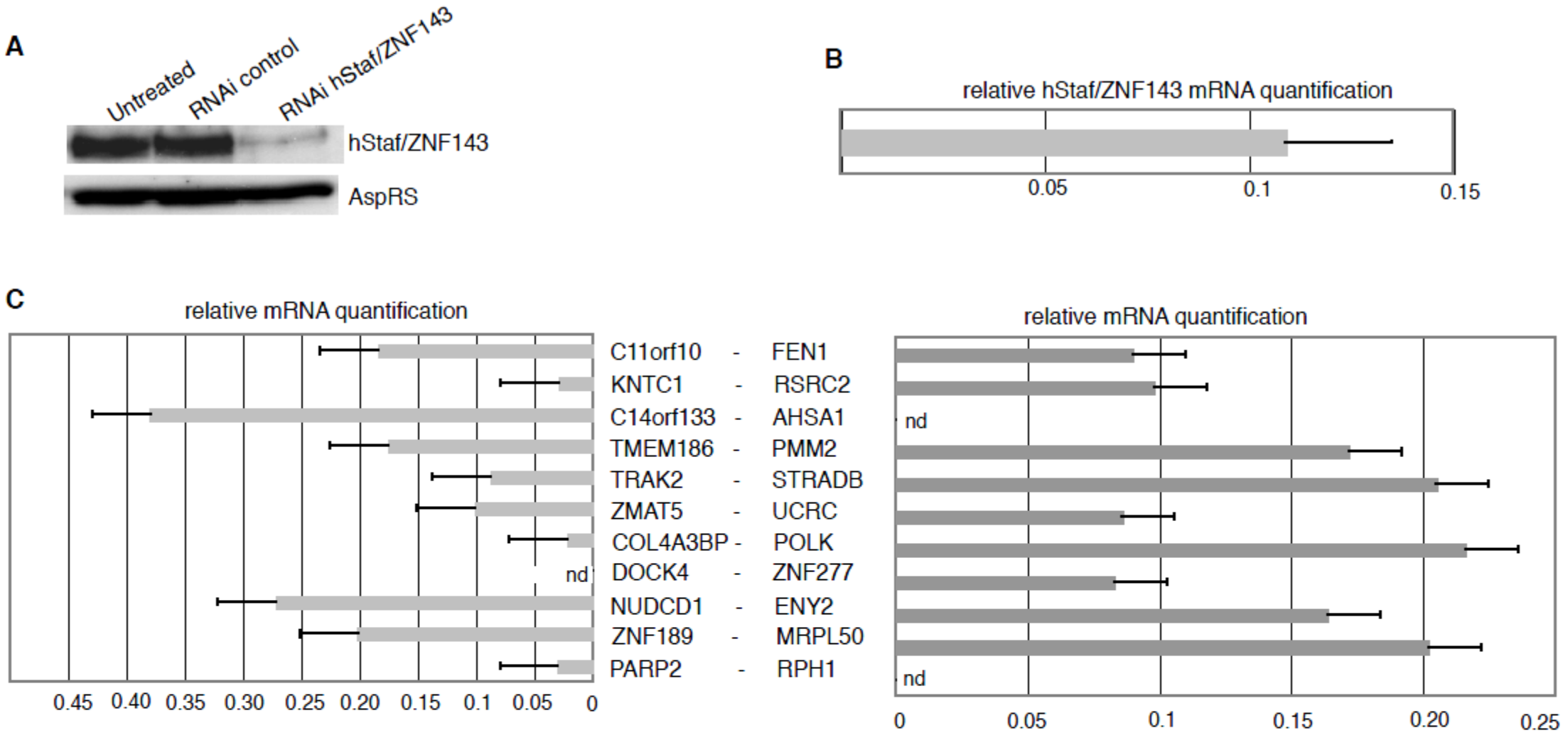

A
Untreated
RNAi control
RNAi hStaf/ZNF143
hStaf/ZNF143
AspRS
B
relative hStaf/ZNF143 mRNA quantification
0.05
0.1
0.15
C
relative mRNA quantification
C11orf10 - FEN1
KNTC1 - RSRC2
C14orf133 - AHSA1
TMEM186 - PMM2
TRAK2 - STRADB
ZMAT5 - UCRC
COL4A3BP - POLK
DOCK4 - ZNF277
NUDCD1 - ENY2
ZNF189 - MRPL50
PARP2 - RPH1
nd
0.45 0.40 0.35 0.30 0.25 0.20 0.15 0.1 0.05 0
relative mRNA quantification
nd
nd
0 0.05 0.1 0.15 0.20 0.25

Figure 5

3. Détermination expérimentale du répertoire des sites de fixation de hStaf à l'échelle du génome humain

3.1. Premier ChIP-Seq

Afin de déterminer le répertoire des sites fixant effectivement hStaf in vivo, nous avons décidé de réaliser une expérience de ChIP-Seq en utilisant un anticorps dirigé contre la partie C-terminale de hStaf. Après immunoprécipitation de l'ADN fragmenté de cellules HeLa-s3 associé à hStaf, une banque d'ADN a été construite au laboratoire puis le séquençage de l'ADN de la banque fut réalisé par la société GATC. La technique étant alors encore dans ses premiers stades de développement et beaucoup de données ont été perdues du fait du manque d'optimisation. Ainsi, seulement 4.7 10^6 de tags ont été obtenus et parmi eux seulement 0,94 10^6 purent être replacés à des positions uniques dans le génome humain. Nous avons tout de même choisi de tenter d'exploiter ces résultats en considérant, de manière discutable, que la présence de 5 tags pouvait être suffisante pour former un pic. Nous avons développé un pipeline de traitements permettant de détecter ces "pics", de récupérer la séquence associée à chaque pic, d'éventuellement étendre cette séquence et finalement de confronter les localisations des pics avec celles des 165000 sites SBS que nous avions préalablement identifié in silico. Il fût ainsi possible de déceler 293 pics, parmi lesquelles 32% contenaient en moyenne 1,8 SBS par pic avec un excellent score moyen de 0.9. Par ailleurs, les pics correspondant renfermaient un nombre moyen de 11 tags. Le faible nombre de pics obtenus ne semblait pas dû à notre méthode de traitement des données mais bien au manque d'optimisation de la technique de ChIP-Seq puisque parmi les 400 SBS expérimentalement validés par ChIP et PCR en 2006, seulement 53 (13.5%) ont pu être retrouvés par cette expérience de ChIP-Seq. Les résultats s'avéraient tout de même prometteurs et le fort score des SBS identifiés validait la robustesse de la matrice.

3.2. Un second ChIP-Seq qui génére 4088 pics d'étiquettes avec un expect de 10-7

De nouveaux échantillons d'ADN ont ensuite été préparés par ChIP au laboratoire par Patryk Ngondo Mbongo puis séquencés sur la plate forme de séquençage à haut débit Illumina Solexa de l'IGBMC (Illkirch). Les parties 5' de 25.10^6 fragments ont été séquencés sur 32 nucléotides et 78 % d'entre eux ont été placés sur des sites uniques dans le génome humain. Disposant dès lors de suffisamment de données biologiques, nous avons utilisé le programme MACS (Zhang *et al.*, 2008) permettant de détecter, sur le génome, les enrichissements locaux en tags, qui s'organisent en pics de tags par rapport au bruit de fond de l'expérience. Nous avons ainsi pu

mettre en évidence 4088 pics avec une p valeur forte de 10^{-7}. Nous allons décrire maintenant les divers traitements réalisés et les caractéristiques des séquences qui nous ont permis de mettre en évidence la présence de sites SBS dans 3060 de ces pics. La stratégie développée est présentée dans la Figure 72.

3.2.1. La moitié des pics renferme des SBS prédits par notre crible bioinformatique

Nous avons développé un programme permettant d'identifier les SBS qui, parmi les 165000 SBS issus de nos prédictions bioinformatiques (paragraphe 1.2) se retrouvaient dans les 4088 pics de tags de l'expérience de ChIP-Seq. Il est apparu que seulement 2074 (51%) des pics recelaient au moins un des 165000 SBS prédits et la moyenne du nombre de SBS dans ces pics est proche de 2. Le score de l'ensemble des sites s'étale de 0.82 à 0.99 avec une valeur moyenne de 0,88 selon MatInspector. Dans la suite du texte les 2074 pics renfermant des SBS issus des prédictions bioinformatiques seront dénommés pics SBS1. Les 2014 pics qui ne renfermaient apparemment pas de SBS seront qualifiés de pics orphelins.

L'analyse par GeCo des pics orphelins, a mis en évidence la présence d'un élément répété de type Alu dans ou a proximité de ceux-ci dans 50% des cas (1096/2014). Le mapping des étiquettes sur le génome a été réalisé sur une version non masquée du génome. Par contre, nos prédictions sur la localisation des SBS ont quant à elles été réalisées sur une version masquée du génome. Il ne nous a donc pas été possible d'obtenir des informations en croisant les données concernant les SBS prédits avec les pics renfermant des éléments répétés.

3.2.2. Les SBS validés antérieurement sont présents dans les séquences associées aux pics SBS1

Nous avons en premier lieu examiné, si les SBS validés expérimentalement par ChIP-PCR lors de l'étude précédente (Myslinski *et al.*, 2006) étaient présents dans les pics SBS1. Ces 2074 pics renferment 387 des 400 sites validés expérimentalement (97%). Ce pourcentage élevé atteste de la qualité des données générées lors de ce ChIP-Seq.

3.2.3. Les pics orphelins et les pics SBS1 ont des caractéristiques contextuelles propres

3.2.3.1. Paramètres bioinformatiques

Nous nous sommes tout d'abord demandés si la forte proportion de pics orphelins (2014 pics) n'était pas liée au score utilisé par notre méthode de prédiction des sites. Le fait d'abaisser le seuil autorisé du score des SBS prédits par MatInspector de 0,82 à 0,81 n'affecte quasiment pas le nombre de pics orphelins (5% d'orphelins en moins), mais augmente seulement le nombre de sites prédits dans les pics en renfermant déjà. Cette augmentation dans les régions renfermant déjà des SBS pourrait résulter soit d'un biais dans la composition ou soit de l'existence d'un pool de sites plus ou moins dégénérés, faisant émerger des faux positifs. L'algorithme de prédiction ne semble donc pas en cause pour rendre compte de la présence les pics orphelins.

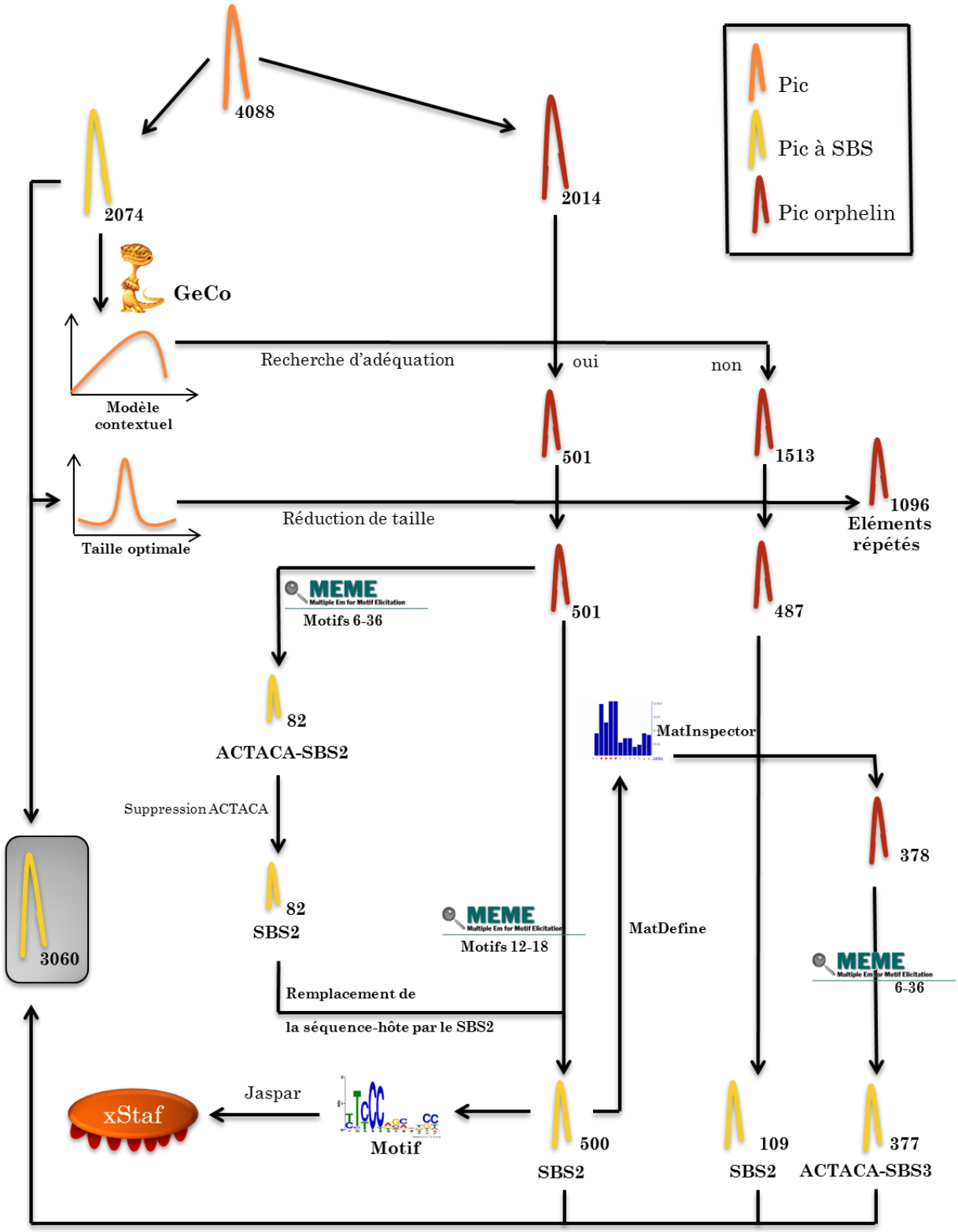


Figure 72 : Analyse des pics orphelins ayant permit d'identifier 3060 pics possédant en réalité des sites SBS2 et SBS3 dévoyés

3.2.3.2. Les SBS ont une position préférentielle au sein des pics

Ne pouvant identifier le candidat à l'origine du pic dans les pics a multiples SBS prédits, nous avons évalué la position relative des SBS au sein des 1070 pics à SBS1 ne possédant qu'un site. Il apparaît que les SBS sont essentiellement placés dans la partie médiane du pic et retrouvés dans les 400 pb encadrant la partie centrale des pics (Figure 73).

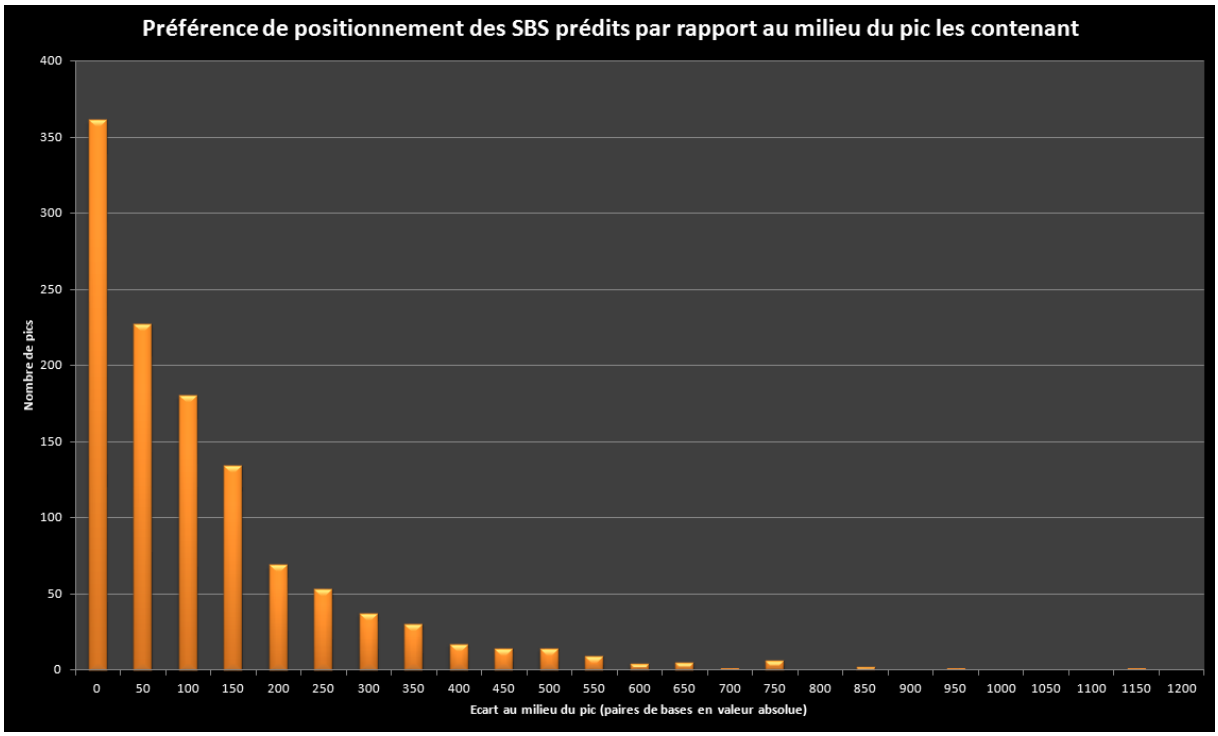


Figure 73 : Préférence de positionnement des SBS prédits par rapport au milieu du pic les contenant.

3.2.3.3. Le nombre de tags dans les pics orphelins et plus faible que dans les pics SBS1

Nous avons évalué comment se distribuaient les pics SBS1 et les pics orphelins en fonction du nombre de tags qu'ils renferment. Pour éviter le biais introduit par les pics larges, nous avons normalisé le nombre de pics par la largeur de la base du pic. Les résultats sont présentés dans la Figure 74. Globalement, 11% des pics SBS1 ont moins de 50 tags, 21% ont de 50 à 100 tags et 69% ont plus de 100 tags, contre respectivement 51%, 32% et 17% pour les pics orphelins (Tableau 4). Il apparaît donc qu'un nombre important de tags est une caractéristique des pics SBS1 qui peut refléter une forte affinité du facteur hStaf pour ces sites. Il est donc fort probable que les pics orphelins doivent contenir d'autres types de sites avec une affinité moindre pour la protéine qu'il sera difficile d'identifier par le seul modèle mathématique du site de fixation que nous avons établi. Une autre éventualité serait que les pics orphelins constituent des faux positifs de l'expérience de ChIP-Seq. Il est difficile pour l'instant d'appréhender les faux-

positifs expérimentaux car nous ne disposons pas de résultats de séquence obtenus soit avec de l'ADN nu (ADN input) soit avec de l'ADN obtenu par immunoprécipitation de la chromatine avec un anticorps dirigé contre un autre facteur de transcription. Néanmoins la présence d'un nombre conséquent de tags dans une part non-négligeable de pics orphelins rend très peu probable qu'ils représentent des faux positifs expérimentaux.

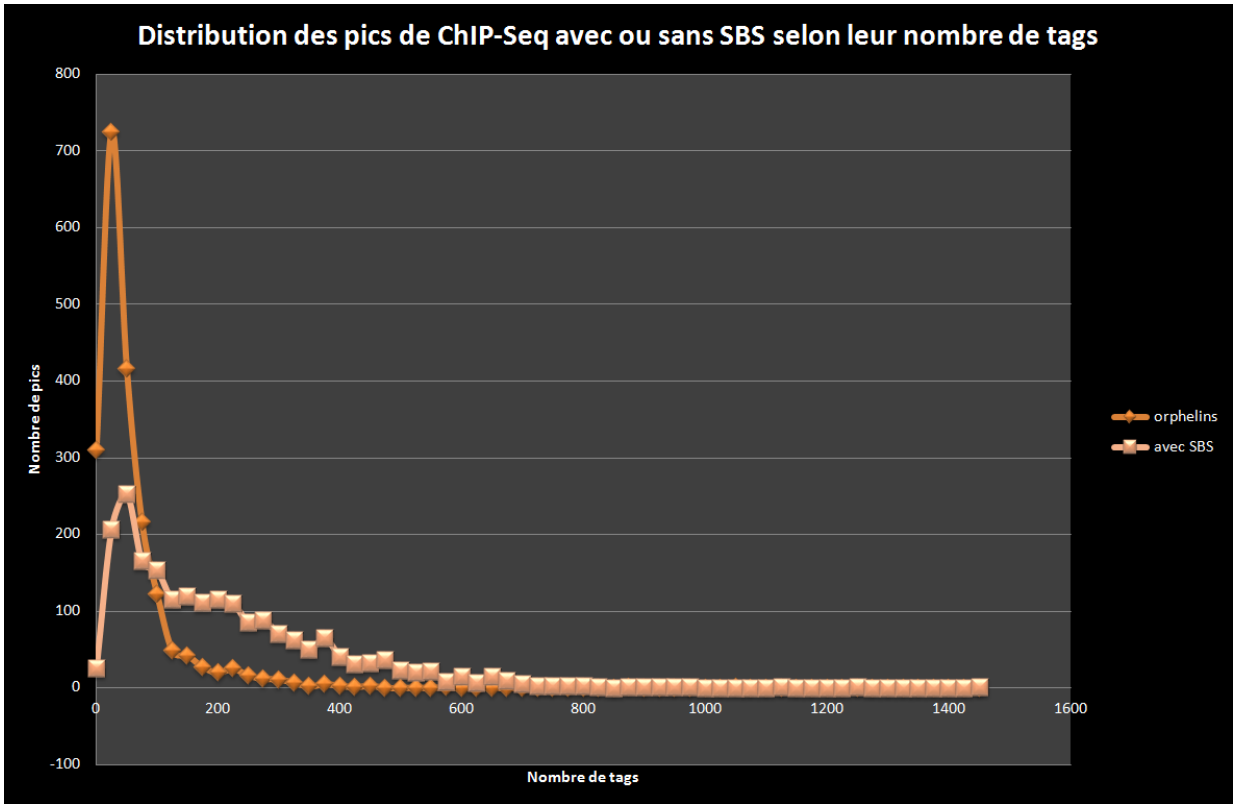


Figure 74 : Distribution des pics de ChIP-Seq avec ou sans SBS selon leur nombre de tags.

On voit ici que les pics possédant peu de tags sont bien plus nombreux parmi les pics orphelins que parmi ceux possédant un SBS prédit.

3.2.3.4. Le contexte génomique est différent pour les pics SBS1 et les pics orphelins

Nous avons exploité GeGo pour renforcer le critère discriminant du nombre de tags par une signature contextuelle propre aux pics SBS1. GeCo permet d'utiliser des données disponibles, sur la localisation de la polymérase II, des îlots CpG et des zones d'euchromatine, pour la même lignée cellulaire que celle utilisée pour l'expérience de ChIP-Seq, la lignée HeLa-s3. L'idée sous-jacente étant qu'un certain nombre de pics orphelins contiendrait des SBS dévoyés avec une empreinte contextuelle similaire aux pics SBS1. Nous avons évalué et comparé, pour les pics orphelins et les pics SBS1, la présence proche de la polymérase II, d'îlots CpG et de

zones d'euchromatine. Il apparaît que, 81% des pics SBS1 sont présents dans des îlots CpG, 82% co-localisent avec les sites de la polymérase II et 95% avec les zones d'euchromatine. Ces caractéristiques sont beaucoup plus lâches pour les pics orphelins pour lesquels la co-localisation avec les îlots CpG, la Pol II et l'euchromatine ne représentent plus respectivement que 36%, 39% et 56% (Tableau 4).

Nous avons également intégré la distribution des deux types de pics par rapport aux gènes en distinguant: l'intergène (jusque 2kbp en amont du TSS), Exon1, Intron1, autres exons et introns, l'intergène en 3' du gène (jusque 2kbp en aval du gène). Il s'avère que les pics à SBS1 sont résolument attachés aux gènes et ont plus particulièrement une localisation de type «promoteur» étendu avec comme marqueur fort une région couvrant environ 4000pb autour du TSS (promoteur plus le 1er exon et le 1er intron) (Tableau 4).

	Orphelins	SBS
Tags <50	51%	11%
Tags [50-100]	32%	20%
Tags > 100	17%	69%
Desert génique	46%	1%
<2kb	36%	31%
Exon1	14%	8%
Intron1	10%	2%
Ailleurs dans le gène	14%	1%
CpG	36%	81%
Pol II	39%	82%
Chromatine	56%	95%

Tableau 4 : Récapitulatif des fréquences d'occurrence des critères contextuels caractérisant les pics à SBS et orphelins

On voit ici que le pic à SBS est résolument orienté "promoteur" avec une localisation préférentielle à proximité du TSS, des présences d'euchromatine, îlots CpG et polymérase II plus importantes et un nombre d'étiquettes plus grand.

En étudiant la combinatoire des éléments susmentionnés, nous avons défini un modèle de contexte des pics SBS1. Si le pic ne vérifiait pas un de ces cinq critères (nombre de tags élevé, proximité d'un ilôt CpG, d'un gène, de la polymérase, de l'euchromatine), son score pour ce critère serait de zéro. S'il le vérifiait, son score ne serait pas de 1 mais de « p » où p est la

probabilité pour un pic SBS1 donné, de valider ce critère (0.81 pour « être dans un îlot CpG »). Ceci permet indirectement de limiter la pénalisation d'un pic ne validant pas le critère.

Nous avons établi le score des pics SBS1 et des pics orphelins via cette fonction de score contextuel. Les scores se répartissent en 5 classes allant de 0 à 4 et nous avons observé l'appartenance globale des deux types de pics à chacune des classes. Il ressort que les pics SBS1 se répartissent massivement dans les classes 3 et 4 alors que les pics orphelins se situent majoritairement dans les classes précédentes (Figure 75).

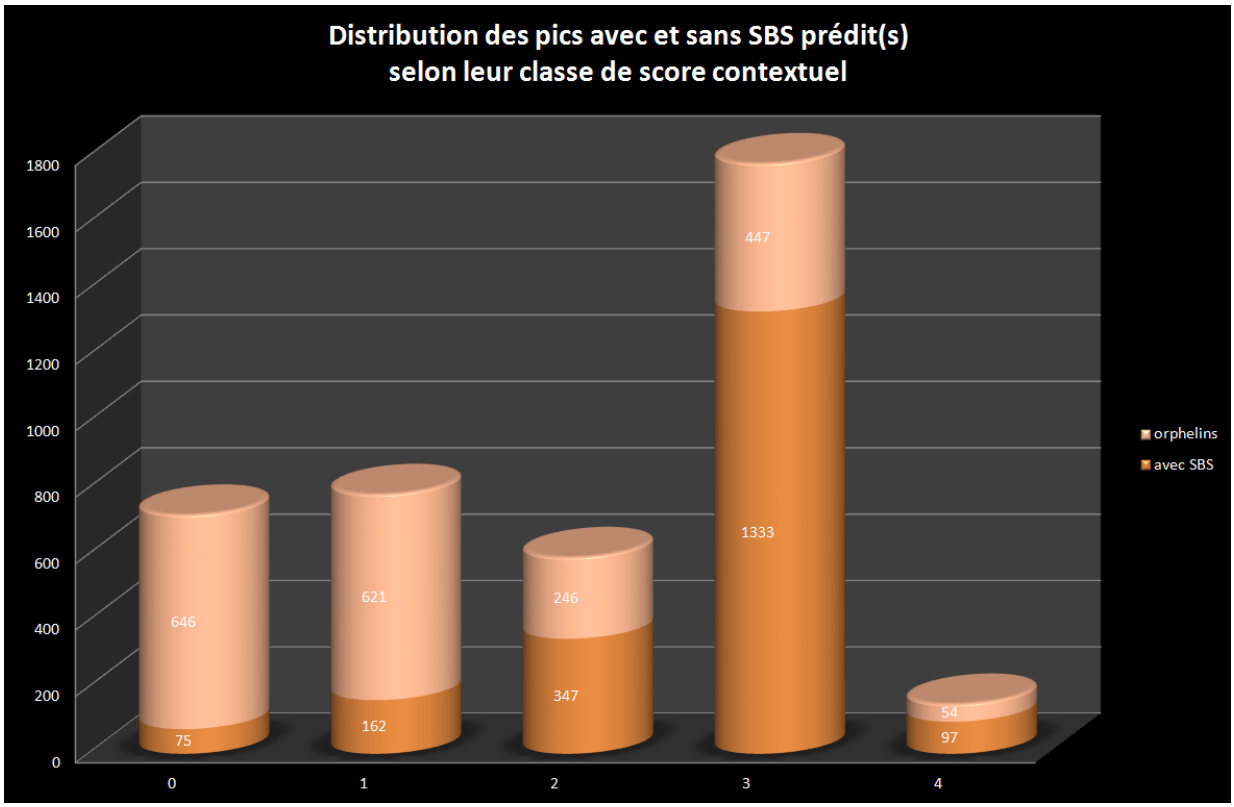


Figure 75 : Distribution du nombre de pics avec et sans SBS selon leur score contextuel.

Les pics à SBS se répartissent majoritairement dans les classes de score 3 et 4, alors que les pics orphelins sont préférentiellement dans les classes de 0 à 2.

L'analyse du contexte génomique permet donc de différencier les populations alors que le modèle mathématique ne le permettait plus. De manière attendue, un certain nombre de pics orphelins présentent un comportement de type "pics à SBS1". Ces 501 pics (447 + 54) au score supérieur ou égal à 3 ont donc une forte probabilité d'être de réels pics contenant un SBS dévoyé.

3.2.3.5. Mise en évidence d'un second type de SBS, le SBS2

Nous avons alors cherché, pour les séquences correspondant aux 501 pics orphelins (447 + 54) dont le score est supérieur ou égal à 3, à faire émerger un motif en utilisant le logiciel MEME. Ce programme basé sur des modèles de Markov cachés, permet de rechercher des mots de longueur variable et d'en étudier la pertinence par rapport à leur séquence flanquantes. Il est apparu que 82 pics orphelins contiennent une séquence de type RRACTACA-SBS2, où le SBS2 est un motif beaucoup plus dévoyé que le SBS1 obtenu par nos prédictions bioinformatiques. En effet, une flexibilité importante apparaît sur les positions C_{12}, C_{14} et C_{21} du motif de la Figure 76.

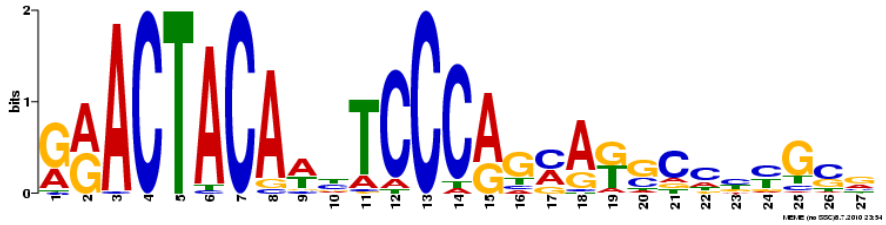


Figure 76 : Weblogo de l'ACTACA-SBS2 mit en évidence dans les séquences de 82 pics orphelins au contexte proche d'un pic contenant un SBS classique

Afin de mettre en évidence d'autres séquences parmi les 501 pics orphelins à fort contexte SBS, nous avons essayé d'aider encore plus le programme à détecter les modèles sous-jacents. Pour cela, nous avons privé les 82 ACTACA-SBS2 de leur ACTACA, avons remplacé l'intégralité de la séquence du pic les contenant par ce seul SBS2 et réanalysé les 501 séquences recentrées sur 400pb. La fonction de « tuteur » de ces SBS2 eût les résultats attendus puisqu'elle permit de mettre en évidence un motif de la même longueur que le SBS2 aux allures proches. Ce motif était présent une fois dans toutes les séquences à l'exception d'une seule et ressortait avec une *p*-valeur forte de 2.10^{-15}. La séquence de ce motif est très dévoyée et ne renferme pas la partie 3' du site SBS (Figure 77).

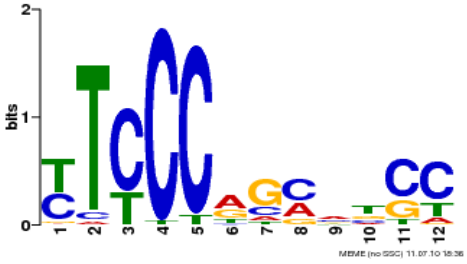


Figure 77 : Web logo du SBS2 seul identifiés dans 500 pics « orphelins »

Nous avons par la suite construit, via MatDefine, une PSSM à partir de ce motif SBS2 afin d'explorer les cœurs de séquences des 1513 pics orphelins restants, de score contextuel compris entre 0 et 2. GeCo nous ayant alertés sur la présence d'éléments Alu dans 1026 des séquences associées à ces pics nous les avons exclu de notre analyse afin de ne pas induire les HMM en erreur. En scannant ainsi les 487 pics orphelins restant nous avons pu mettre en évidence 109 pics possédant ce site SBS2. De manière intéressante, par examen manuel des séquences, on retrouve l'ACTACA non-lié directement au SBS mais quelques nucléotides plus loin, ou inversé. Il se pourrait donc que cette séquence puisse exister en tant que telle en dehors de tout SBS et suggèrerait un autre facteur impliqué dans sa reconnaissance.

Enfin, notre analyse se centra sur les 378 pics orphelins restants (487–109) dans l'espoir d'identifier un autre signal. De manière surprenante, l'analyse par MEME mit en évidence un troisième type de SBS (SBS3) dans 377 de ces séquences. Le SBS3 est toujours associé au motif RRACTACA, a complètement perdu la cytosine en position 12 du SBS originel (qui serait en position 21 sur la figure Figure 78) et voit la cytosine 5 très malmenée. Le seul signal fort demeurant est la cytosine 4 et le couplage systématique au motif RRACTACA qui laisse supposer que le maintien fonctionnel d'un tel site est indissociable de ce motif.

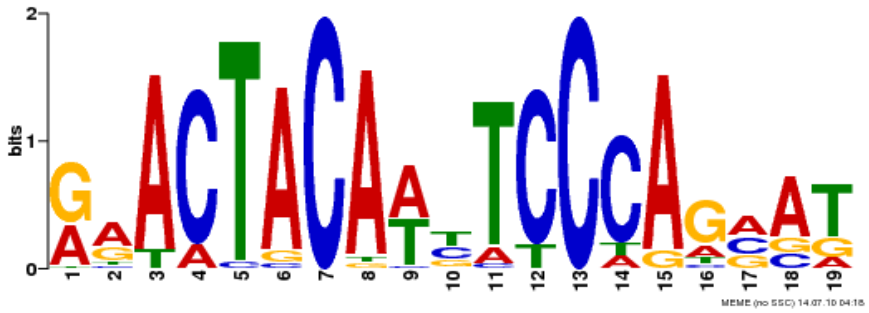


Figure 78 : Web logo de l'ACTACA-SBS3 identifié parmi 377 séquences de pics « orphelins »

Cette analyse, soulignant l'intérêt et la puissance de la caractérisation par le contexte génomique est résumée à la Figure 72.

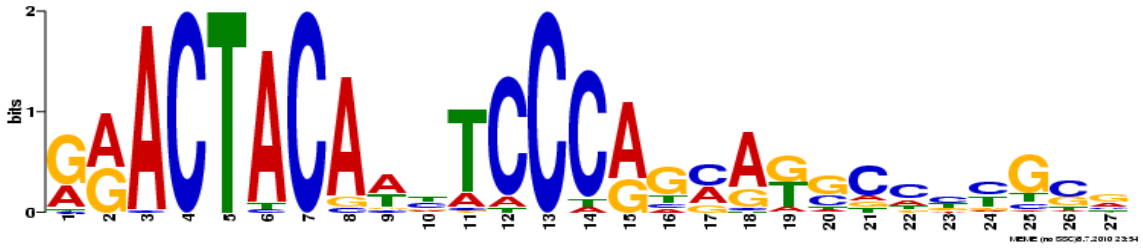
3.2.3.6. Un motif de type RRACTACA est fréquemment trouvé en 5' des SBS1, 2 et 3

35% des SBS présents dans les pics SBS1 sont associés à une séquence de type RRACTACA en 5' du SBS. Il en est autrement pour les SBS2, pour lesquels le motif RRACTACA n'est retrouvé associé que dans 16% des cas alors que pour les SBS3 la présence du motif RRACTACA est systématique.

L'examen des séquences révèle un jeu subtil entre assouplissement et pression sur les différentes bases d'un motif bipartite liant toujours hStaf et tenant *a priori* les modèles mathématiques de recherche en échec. Par ailleurs, en dehors du motif RRACTACA, aucun autre motif n'émerge en contact direct avec le SBS, laissant supposer que, quel que soit son mode d'action, celui-ci est le seul à venir renforcer le SBS (Figure 79).

A

B



C

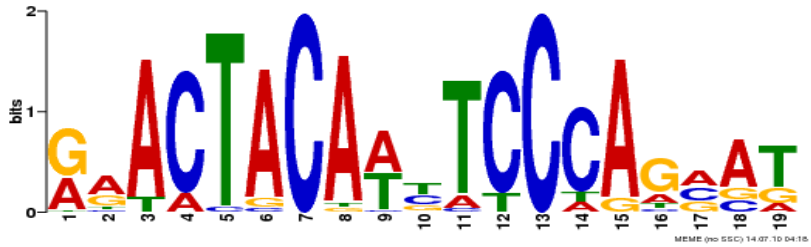


Figure 79 : Profils des SBS1 (A), SBS2 (B) et SBS3 (C) précédé de leur ACTACA lorsqu'il leur est associé.

Selon le SBS, la présence de ce motif aura un effet plus ou moins important sur le relâchement de pression du SBS. Au sein même de l'ACTACA pourtant plus conservé de la paire, la pression varie en fonction des SBS.

De manière surprenante, 1073 SBS issus de nos prédictions informatiques et associés à la séquence ACTAYR ne sont pas retrouvés dans les pics identifiés par le ChIP-Seq. La longueur de la séquence résultant de l'association entre ACTACA et SBS est de 25 nucléotides dont 12 avec une probabilité d'occurrence de plus de 50%. La probabilité d'identifier ce motif par hasard est donc très faible. Il est possible que l'expérience n'ait pas mis à jour l'ensemble des sites. Ceci peut éventuellement être du à la lignée cellulaire utilisée ou simplement à la non accessibilité de certains complexes dans les conditions de l'expérience ou encore à un possible rôle d'un second facteur reconnaissant spécifiquement tout ou partie de la séquence ACTAYR-SBS.

3.2.3.7. L'analyse des gènes-cibles fait de hStaf un régulateur de haut-niveau

Nous nous sommes servi des localisations géniques fournies par GeCo pour identifier les gènes cibles associés à des SBS contenus dans les pics situés à proximité de gènes. Au total, 2193 gènes protéiques et 121 gènes d'ARN non-codants sont à proximité de SBS identifiés dans ce travail. hStaf apparaît donc comme un acteur majeur de l'expression du génome et bien qu'initialement identifié dans le contrôle de l'expression de gènes d'ARN non-codants, hStaf serait principalement impliqué dans celle des gènes protéiques. hStaf est retrouvé associé aux gènes cibles déjà connus comme les gènes d'ARNsn et de l'ARNt^{Sec} mais également à proximité de gènes d'ARNmi, d'ARNr ou d'autres ARNt, sans qu'un quelconque mécanisme ait été identifié pour ces derniers (Figure 80).

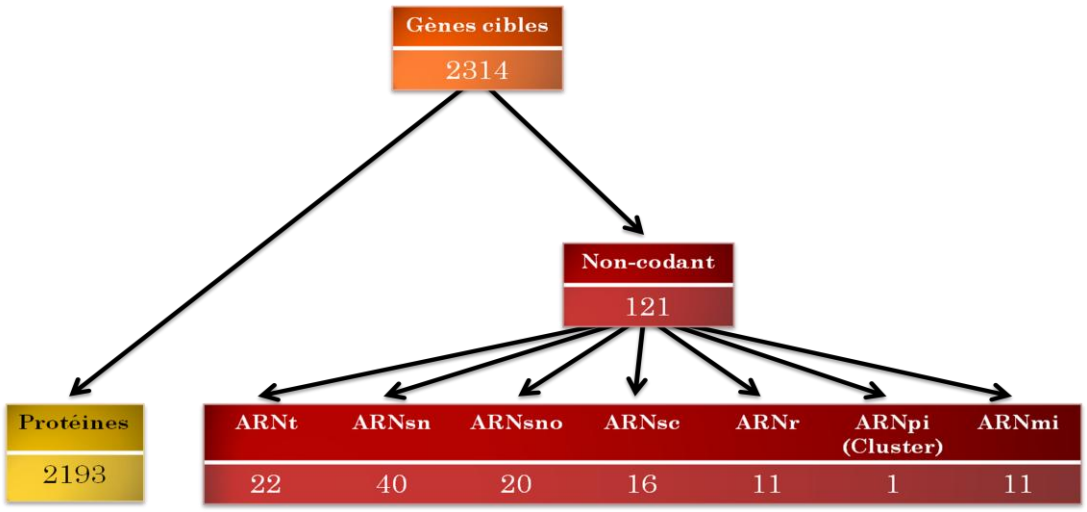


Figure 80 : Répartition des gènes-cibles de hStaf selon leur type

Une prévalence est observée pour les gènes codants.Parmi les ARN non-codants, on retrouve la prévalence des ARNsn et ARNsno et des types plus atypiques comme des ARNt classiques.

Nous avons étudié la fonction et la localisation cellulaire des protéines exprimées à partir des gènes identifiés, à l'aide du portail DAVID. Nous avons séparé ces gènes selon qu'ils soient associés à l'un ou l'autre des 3 types de SBS. Globalement il s'agit toujours de gènes de maintenance de la cellule, ce qui est cohérent avec l'abondance des sites de hStaf dans les îlots CpG. Toutefois, parmi les enrichissements très significatifs (*p*-valeur < 10^{-7}), certaines spécificités apparaissent pour chacun des types de SBS.

Les gènes sous la dépendance du SBS1 présentent un très fort enrichissement en protéines à doigts à zinc qui représentent 15% de ses cibles. Plus précisément, hStaf régulerait près d'un quart des protéines KRAB du génome. Ces protéines sont en majorité des répresseurs de la transcription, avec un nombre très variable de doigts à zinc de type C_2H_2. Si les premières protéines KRAB sont apparues avec les tétrapodes, leur colonisation massive du génome est associée aux mammifères et leur localisation génomique se fait généralement au sein de clusters de gènes aux profils d'expression tissulaire, supposément du fait de la présence d'éléments régulateurs similaires. En particulier, le plus important cluster est situé dans la région 19p12-p13.1 (Bellefroid *et al.*, 1993) et représente à lui seul la moitié des protéines KRAB (Rousseau-Merck *et al.*, 2002). Nous avons en conséquence étudié la distribution chromosomique des pics totaux de notre expérience et constaté en enrichissement majeur sur le chromosome 19. Le chromosome 19 étant le plus dense en gènes, nous avons voulu vérifier que l'enrichissement observé est plus fort que la linéarité globale des sites par rapport aux gènes. Nous avons ainsi comparé la densité en pics et en gènes de chaque chromosome en rapportant le nombre de ces entités à la taille du dit chromosome (Figure 81). Si pour la quasi totalité des pics, la densité suit celle des gènes, pour le chromosome 19 et pour le chromosome 5 l'enrichissement est indiscutable. Ainsi, hStaf serait un acteur majeur de la régulation des protéines de la famille KRAB et viendrait confirmer leur co-régulation supposée. Pour le chromosome 5, les gènes concernés codent pour deux familles de protéines impliquées dans l'adhésion cellulaire les cadhérines et les thrombospondines qui seraient sous le même schéma régulateur par hStaf que les protéines à KRAB domain. Toutefois l'enrichissement par rapport aux autres familles de gènes régulés par hStaf, comme les protéines KRAB.

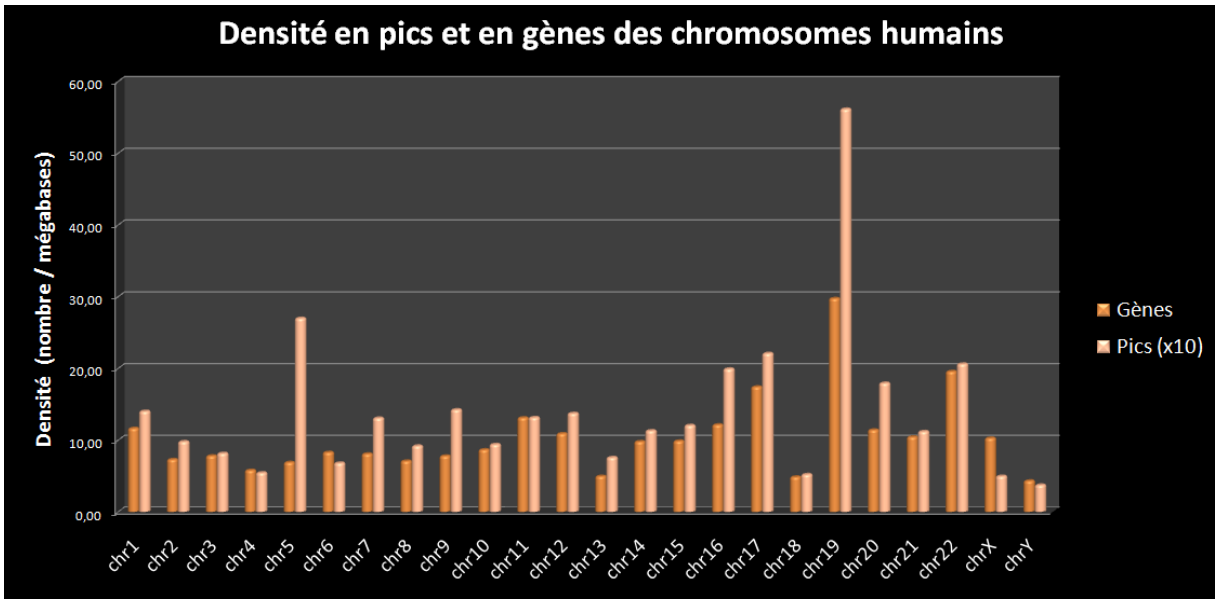


Figure 81 : Densité en pics et gènes des chromosomes humains

Un enrichissement en pics à SBS est observé sur le chromosome 19 et le chromosome 5. Ces enrichissements ne peuvent être totalement expliqués par la densité en gènes de ces chromosomes.

Par ailleurs les cibles de hStaf présentent globalement un enrichissement dans 7 autres fonctions. De manière surprenante, de nombreuses cibles sont des gènes codant pour des protéines mitochondriales et plus particulièrement de la membrane mitochondriale et ces gènes sont associés à des SBS1 et 2. Par ailleurs, on note un enrichissement marqué pour les protéines liées au cycle cellulaire et plus modestement à la réparation (précédemment identifiées dans les promoteurs bidirectionnels). L'ensemble de ces dernières observations, couplé au rôle possible de hStaf dans l'expression de régulateurs de la transcription comme les protéines KRAB, laisse entrevoir un rôle de régulateur de très haut niveau pour hStaf qui se placerait en amont des grandes cascades de régulation et coordonnerait le cycle cellulaire, la réparation et la transcription par un jeu d'activation et de répression. Enfin, une part importante des gènes cibles concerne les gènes de complexes ribonucléoprotéiques et les gènes du *RNA processing* des ARNm et des ARNnc. Enfin, si des gènes protéiques sont associés principalement avec des SBS1 comme des gènes de la réparation ou du cytosquelette, le seul enrichissement significatif des cibles des SBS2 est trouvé dans des gènes du catabolisme des protéines. Ces résultats sont résumés à la Figure 82 où la fraction du total des cibles d'un type de SBS est reportée pour chaque fonction ou localisation, à condition que l'enrichissement soit significatif, sans quoi cette fraction est rapportée à 0 pour éviter toute mésinterprétation (Figure 82). Enfin, de manière

intéressante, un pic contenant un SBS2 est localisé dans le premier intron du gène de hStaf, suggérant un possible rôle de hStaf dans le contrôle de sa propre expression.

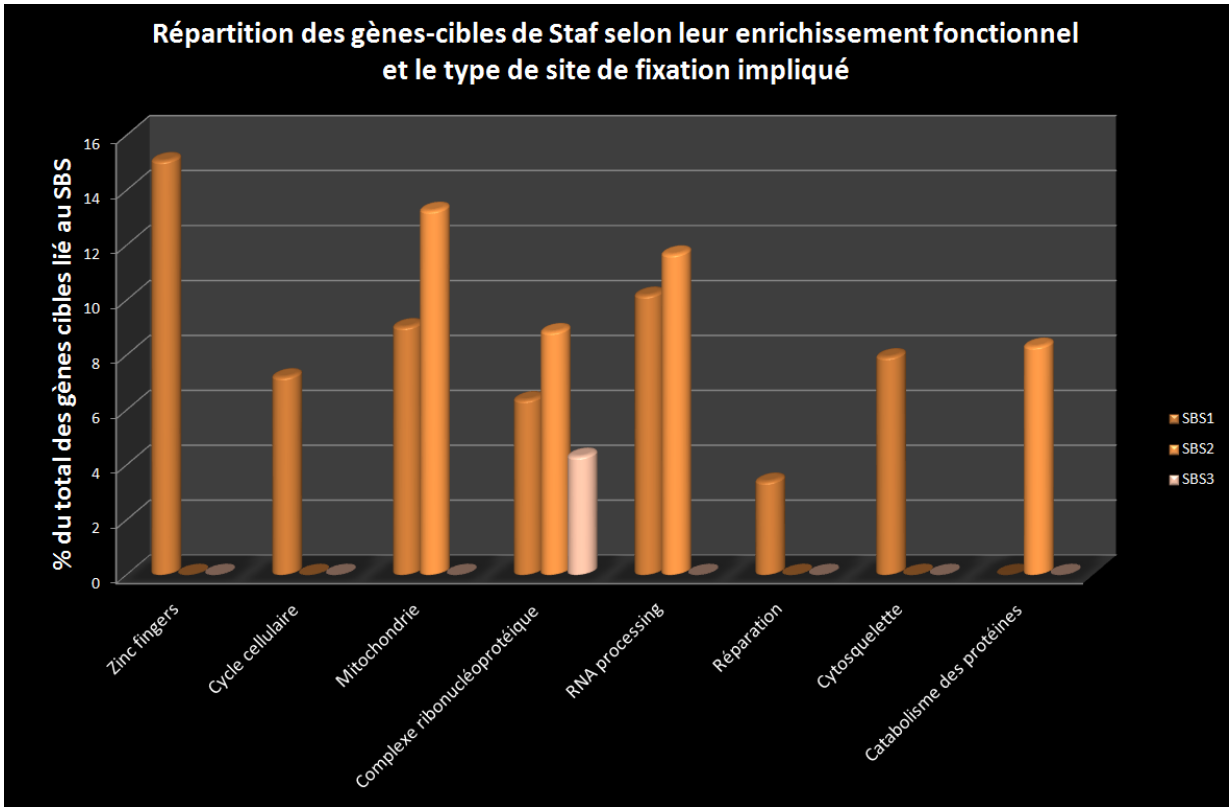


Figure 82 : Répartition des gènes cibles de hStaf selon leur enrichissement fonctionnel et leur type de SBS

Les fonctions ne présentant pas d'enrichissement ont été ramenées à 0 pour éviter toute mésinterprétation. On voit ici que le SBS1 régulerait majoritairement des protéines à doigts-à-zinc, des protéines impliquées dans le cycle cellulaire, la réparation ou le cytosquelette. Un enrichissement en catabolisme des protéines semble être propre au SBS3 et d'autres localisations majeures comme la mitochondrie et les complexes ribonucléoprotéiques sont également surreprésentés.

A la lumière de cette analyse, le facteur hStaf s'impose comme un élément clé de la régulation du génome humain avec, potentiellement, 2314 gènes cibles (proches de 10 % des gènes totaux), dont 121 ARNnc et 2193 gènes protéiques. Ces derniers semblent eux-mêmes des acteurs des grandes fonctions de la vie de la cellule telles que le cycle cellulaire, la transcription ou la réparation, plaçant Staf très en amont des cascades de régulations. Son rôle dans l'activation de familles de protéines co-localisées comme les protéines à domaine KRAB du chromosome 19 ou les cadhérines et les thrombospondines du chromosome 5 semble avéré mais demande à être plus largement étudié. En outre, notre analyse a permis de retrouver la quasi-totalité des sites suggérés par l'expérience de ChIP-Seq, que nous avons pu catégoriser en trois classes et à qui certaines fonctions des gènes cibles semblent échoir de manière spécifique. Enfin, nous avons démontré l'importance de Staf, validée expérimentalement, dans les promoteurs bidirectionnels.

Nous avons abordé l'étude de ce facteur de manière originale, en partant des connaissances acquises par la biologie (l'expérimental) et construit, en premier lieu, un modèle exploratoire permettant d'identifier les grands traits des SBS à étudier et avons anticipé les résultats expérimentaux à haut-débit. De plus, en couplant un modèle mathématique, aux connaissances, expériences biologiques et à l'analyses du contexte informationnel des gènes, nous avons pu passer outre les faiblesses de chaque domaine. La biologie nous a apporté à la fois le modèle et la validation, mais a profité de l'analyse contextuelle et des prédictions mathématiques pour expliquer les résultats. A l'inverse, notre analyse contextuelle n'aurait pu être si solide sans les données expérimentales sur le facteur Staf et sur le contexte informationnel de la transcription sur lequel repose notre architecture.

Les résultats expérimentaux (ChIP-Seq) concernant le nombre de régions génomiques potentiellement reconnues par le facteur Staf ont été obtenus dans les dernières semaines de la rédaction de ce manuscrit. Les analyses GeCo que nous avons menées ont permis de vérifier ou de conforter certaines hypothèses, mais plus encore, ces analyses ont permis d'ouvrir de nombreuses questions et pistes de recherche qu'il faudra valider à l'avenir par de nouvelles expériences à haut débit.

Discussion et perspectives

Chapitre 9 : discussion et perspectives

1. Le facteur hStaf/ZNF143

1.1. Un nombre important de cibles... sous-estimé ?

De par le fonctionnement de hStaf avec (au moins) deux machineries de transcription, et au regard de notre étude démontrant sa présence dans les promoteurs de 2314 gènes protéiques et d'ARNnc, ce facteur s'impose comme un des facteurs de transcription les plus importants du génome en terme de cibles. Il est, à cet effet, possible que le nombre de cibles ait été sous évalué. En effet, notre objectif était d'être le plus stricte possible. Aussi avons-nous choisi une valeur d'expect de 10^{-7} pour la détection expérimentale des pics, quand les analyses de routine se cantonnent à 10^{-5} voire 10^{-4}. Cependant, en se bornant à 10^{-5}, c'eût été 8000 pics qui seraient apparus et il pourrait être intéressant de vérifier la puissance de notre méthode sur de si faibles signaux.

1.2. Un facteur universel ?

Grace à sa base d'annotation génique riche, GeCo permit d'assigner de nombreux types de gènes aux SBS. Au delà des 2193 gènes protéiques, les 121 gènes d'ARNnc se décantent en 7 types. Parmi eux, étaient connus l'ARNtSec, certains ARNsn et ARNsno. Mais l'apparition d'une vingtaine d'ARNt (ou de précurseurs d'ARNt) ouvre les portes vers un univers à part car il s'agit d'un tout autre type de gènes transcrits par la PolIII dont le mécanisme ne fait pas intervenir la PSE possédant le site SBS. Plus surprenant est le fait qu'une dizaine d'ARNr apparaissent dans cette étude. Pour l'heure, gardons nous de conclure quoique ce soit mais dans l'éventualité d'un rôle de hStaf dans l'expression de certains ARN ribosomaux transcrits par PolI, ce sont potentiellement les 3 machineries de transcription qui fonctionneraient avec ce facteur, ce qui serait exceptionnel.

1.3. Rôle de l'ACTACA

Nous avons vu que 35% des SB1, 16% des SBS2 et 100% des SBS 3 possédaient une séquence ACTACA. il a été démontré au laboratoire que cette séquence permet dans la grande majorité des cas d'améliorer la transcription *in vivo*, à l'exception d'un seul cas où il la réprimait. Toutefois, dans le cadre de l'ARNtSec, le SBS sans ACTACA est associé au pic à l'intensité la plus forte de l'expérience ChIP-Seq. De plus, un millier de nos prédictions bioinformatiques

assurait un ACTAYRN-SBS en dehors de tout pic de ChIP-Seq. Enfin, la position de l'ACTACA en -7 souffre d'exceptions et nous avons observé certains cas où celui-ci était plus éloigné du SBS, voire en position inverse. Tous ces faits plaident en la faveur d'un second facteur avec lequel hStaf pourrait soit coopérer, soit être en compétition mais les règles présidant à ces mécanismes s'annoncent complexes.

1.4. Localisations des SBS et modes d'action

Au regard de l'analyse contextuelle fournie par GeCo, les localisations des SBS sont profondément attachées au promoteur. Parmi elles, l'enrichissement marqué en tout début de premier intron lance le débat quant à d'autres mécanismes de transcription dans lesquels hStaf serait impliqué. En effet, nous avons démontré que hStaf était un acteur important des promoteurs bidirectionnels et le SBS n'affiche aucun critère d'orientation pour être fonctionnel. Ceci pose la question de l'implication de hStaf dans l'activation d'une transcription bidirectionnelle hors-intergène, donc une transcription inverse dans le cadre de ce début de premier intron. On l'a vu, cette localisation au sein du premier intron est un point chaud de la transcription inverse (Finocchiaro *et al.*, 2007) donc tout concourt pour que hStaf soit impliqué dans ce phénomène. Cette hypothèse ouvre une nouvelle voie de recherches expérimentales à explorer, notamment par la mise en place d'une expérience de RNA-Seq. En effet, si la fixation de hStaf à proximité de tant de gènes, possiblement bien plus, est irréfutable et sans précédent, elle n'atteste pas d'une activation effective et ne permet pas non plus d'observer une quelconque répression. Accéder aux transcrits réels, codants et non codants, comme c'est désormais possible via le *deep-sequencing* permettra de lever le voile sur ces multiples implications de hStaf. De même, la localisation de certains sites dans le dernier exon pourrait être expliquée par cette expérience. Nous avons limité notre étude des promoteurs bidirectionnels à des gènes en orientation inverse, mais certains gènes proches, dans la même orientation, possèdent un SBS dans leur dernier exon. Une activation transgénique de ceux-ci pourrait également être possible. Nous avons repéré deux sites très proches avec un ACTACA chacun. Au regard de cette séquence, ces deux ACTACAN-SBS pourraient même s'arranger en tige-boucle et donc avoir un rôle à jouer dans la stabilité de l'ARN. Si cela sort complètement de notre champ d'investigation, cela promet toutefois au facteur hStaf de longues années d'étude.

1.5. SBS et promoteurs bidirectionnels.

L'analyse par GeCo a permis de mettre en évidence un rôle privilégié de hStaf dans 20% des promoteurs bidirectionnels. Ce système très particulier permit de mettre en évidence une certaine fréquence dans la distribution de tailles de promoteurs et de la lier possiblement à

l'épigénétique. Il s'agit donc d'un "écosystème" modèle où l'on pourrait déceler d'autres pistes pour comprendre les réseaux de régulation impliquant hStaf ou, à l'inverse, ce qui fait que hStaf n'intervient pas dans une régulation. Mieux connaitre la biologie du facteur ou des réseaux de régulation dans lesquels il se place permettrait d'enrichir nos modèles et ainsi peut-être de répondre à certaines questions qu'adresse cette discussion. Par ailleurs, GeCo nous a permis d'identifier des caractéristiques propres aux promoteurs bidirectionnels en général. Notre analyse des gènes impliqués demande à être complétée par d'autres possibilités qu'offrent GeCo, comme l'analyse des régions génomiques de ces promoteurs ou de la conservation de leur séquence. Par une étude contextuelle, on pourrait même estimer si les 4 cas présentés dans la Figure 65 possèdent une signature contextuelle propre ou si les analyses à venir peuvent être menées sur l'ensemble de ces promoteurs.

1.6. ZNF143 vs ZNF76

Notre étude a permis de lever le voile sur beaucoup de spécificités de l'orthologue humain de hStaf, ZNF143, mais il ne peut être étudié qu'en regard de son paralogue sur lequel bien peu est connu. En effet, bien souvent, le paralogue d'une protéine à doigt-à-zinc acquiert une spécialisation supplémentaire ou de remplacement et assure une complexification des mécanismes. ZNF76 et ZNF143 peuvent tous deux reconnaitre le SBS1 mais rien ne dit que leur différence ne se marque pas sur le SBS2 très dévoyé, malgré leur DBD proche. Au niveau de l'expression, ZNF143 est surexprimé dans l'ovaire alors que ZNF76 l'est dans le testicule, mais ils sont globalement retrouvés dans tous les tissus, ce qui est cohérent avec le fort nombre de cibles de ZNF143. Une dynamique particulière semble donc à l'œuvre dans les gonades qui constitueront de fait un bon système pour comprendre le mode d'action de ces deux paralogues.

2. Le contexte informationnel dans les analyses de routine

Au travers du développement de GeCo, nous avons cherché à poser les bases d'un vaste champ d'investigation, celui du contexte informationnel, qui s'est avéré particulièrement robuste dans le cadre de hStaf. De plus, lors de l'étude réalisée par GeCo sur les cibles du récepteur à l'acide rétinoïque, cet outil a pu proposer avec succès les expériences à mener en priorité par le laboratoire de Cécile Rochette-Egly. Par ailleurs, dans le cadre d'une utilisation ayant comme périmètre les gènes, elle put retrouver des résultats biologiques intuitifs lors de l'analyse des

gènes à variants de transcrits tissu-spécifiques. En plus des études présentées dans ce manuscrit, GeCo a été largement utilisée dans diverses collaborations. D'une part, dans l'analyse fine de promoteurs, comme celui de RARβ dans le cadre d'une collaboration avec Natacha Rochel à l'IGBMC où notre architecture permit, via ses aspects de localisation et phylogénétique, de trouver un régulateur inconnu (manuscrit en préparation). Une autre analyse de ce type a été menée pour le promoteur de CNGA3 étudié en collaboration avec Ronald Carpio à Tübingen où elle permit de mettre en évidence un exon shuffling et une régulation par YY1 validée expérimentalement (manuscrit en préparation). Enfin, l'analyse du promoteur de CYP26A1 conduit GeCo à proposer 9 candidats à sa régulation, actuellement testés au laboratoire de Pascal Dollé à l'IGBMC. D'autre part, GeCo fut impliquée dans des analyses plus massives comme la recherche des cibles du récepteur nucléaire à la vitamine D à l'échelle du génome en collaboration avec Dino Moras à l'IGBMC où elle permit de proposer 300 candidats massivement impliqués dans le métabolisme impliquant le calcium et le développement. Enfin, son score de divergence est actuellement intégré dans diverses études allant de l'analyse fondamentale des kinases du génome humain à l'étude des gènes responsables de maladies mono-géniques (site SM2PH) .

L'implication importante de GeCo dans des projets aussi nombreux et divers rend compte de l'importance du contexte informationnel dans l'analyse de routine car elle apporte un autre regard sur une problématique et permet d'identifier rapidement ce qui échappe à l'œil et à la mémoire. Que ce soit en se concentrant sur la part informationnelle d'une région génomique ou d'un ensemble de gènes, l'analyse informationnelle, malgré ses données demandant à être complétées, permet déjà à notre architecture de se poser en complément des méthodes tenues en échec. Les projets de séquençages se multipliant, nos données vont se faire avec le temps de plus en plus complètes et nos analyses de plus en plus précises. Nous sommes pour cela déjà informatiquement prêts à intégrer les données complémentaires disponibles sur nos vecteurs informationnels, que ce soit en termes de charge de base de données qu'en terme de protocoles d'intégration. Une nouvelle modification d'histone dans une lignée cellulaire ne constitue que deux mots à ajouter au script et une paire d'heures pour être intégrée définitivement. De plus, la base s'optimise en fonction des données et se répare quotidiennement grâce à un ensemble de protocoles automatisés. Nous avons hâte de voir ce que seront les analyses de demain quand la charge informationnelle rendra négligeables les doutes face aux certitudes. Nous cherchons bien sûr à étendre le type de nos vecteurs informationnels et, en tête de liste, se trouve la méthylation de l'ADN qui viendra grossir les rangs de ces vecteurs.

Par ailleurs, certains descripteurs CLON ne sont pour l'heure pas étudiés pour l'ensemble des vecteurs informationnels, comme par exemple l'occupation génique des éléments répétés ou le nombre d'exons par taille de gènes. Nous prévoyons en conséquence d'étendre ces descripteurs aux vecteurs qu'ils ne caractériseraient pas encore.

Enfin, par souci d'ergonomie, nous voulions maximiser la simplicité d'utilisation de cette version pilote et avons sciemment omis certaines options qui devront être présentes dans le futur, comme le choix de la lignée cellulaire dans laquelle les données épigénétiques ont été mesurées. Les statistiques par lignées cellulaires et par modifications d'histones sont calculées depuis le départ, attendant que le portail les exploite finement. Il se posera alors le problème du poids de ces nombreux axes potentiels (80 dormant à ce jour). En effet, dans le cadre notamment du calcul du score de divergence, faut-il considérer de la même manière 80 vecteurs informationnels dont 20 sont des modifications d'histones ? Le choix peut-il être laissé à l'utilisateur ? Comment pré-calculer sans lourdeur des statistiques sur toutes les combinaisons possibles de lignées cellulaires (ou de modifications) que pourrait choisir un utilisateur ?

3. De l'avenir des bases de données biologiques et de la nécessité de s'y préparer

A l'heure prochaine du "séquenceur massif de paillasse", on entrevoit déjà le dépassement total de nos capacités de stockage de telles données. En effet, l'équilibre permanent entre la technique de production, le système de stockage et la technologie d'accès s'y verra grandement déstabilisé. Le stockage, en l'absence de technologie de compression innovante pourra toujours se reposer sur la multiplication des supports de stockage. En revanche le système d'accès aux (bases de) données sera quant à lui le réel challenge de cette future ère *post-deep-sequençing*.

Entre le début et la fin de ce projet de thèse, les règles du jeu ont déjà grandement changé et GeCo tel que nous la proposions serait dépassée d'ici un an ou deux d'un point de vue performance, sans anticipation de notre part. Rien que la mise à jour est passée de 8 heures à 32 par la seule intégration des SNP. Heureusement, nous avions anticipé cette dérive et investi dès le départ dans deux actions majeures. La première a été de nous tourner progressivement vers le système BIRD qui s'impose désormais, avec des standards de tables aux millions

d'enregistrements, comme la seule planche de salut pour maintenir des outils d'accès aux données génomiques. La seconde a été de proposer ce que beaucoup proposent, à savoir un découplage de l'interactivité du site, les calculs s'effectuant en tache de fond et l'utilisateur étant prévenu à la fin de l'exécution. C'est là aussi une chose que nous voulions éviter de prime abord puisqu'elle ne se justifiait pas et demande beaucoup de contrôles, tant sur l'exécution mais aussi sur les actions de l'utilisateur ou encore en terme de suivi des jobs, de charge serveur, de délais de validité des données et de leur espace réservé. Nous avons sur la fin du projet posé les bases d'une telle architecture semi-interactive en découplant le gène informationnel de l'architecture native pour qu'il puisse intégrer des centaines de milliers de coordonnées. Le fait qu'il se contente d'utiliser le moteur de localisation sur les gènes nous permit de traiter les 165000 SBS en 45 minutes. A terme, c'est tout GPS qui devrait fonctionner ainsi. De plus, l'affichage granulaire devra également être revu car, les volumes et le niveau de détail augmentant, ce sont les navigateurs qui deviennent insuffisants si l'on poursuit dans notre voie du "tout affiché mais caché". Un investissement dans la technologie AJAX permettant des requêtes asynchrones réalisées côté serveur devraient aisément pallier ce problème.

4. Le résultat biologique à l'ère du *deep-sequencing*

En développant GeCo, nous avons posé la première pierre d'une analyse guidée par l'information. Elle répond au besoin croissant de replacer les choses dans leur contexte informationnel attestant de leurs possibles aspects atypiques. A cette étape de maturité, elle présuppose que l'utilisateur sache encore où regarder et plus globalement connaisse la question qu'il se pose. Ce présupposé risque de devenir de moins en moins vrai à mesure que les questions biologiques se complexifient et que les données de systèmes très différents viennent à être liées d'une manière qu'on ne pourrait imaginer de prime abord avec les connaissances actuelles. La seconde phase de GeCo sera donc d'intégrer des moteurs de *data mining* permettant d'extraire les messages cachés des résultats afin que l'utilisateur ne se retrouve pas avec des masses de données, certes conceptualisées mais totalement inexploitables. De nombreux travaux sont menés en ce sens, notamment au laboratoire, comme AlexSys se basant sur le *machine learning* pour optimiser les alignements multiples ou sur l'*inductive logic programming* et des probabilités de bayésiens naïfs pour souligner d'improbables corrélations génotype-phénotype. Les outils de ce type sont encore peu

nombreux à être embarqués dans un site web mais nul doute qu'il s'agira sous peu d'un standard bioinformatique dont GeCo ne saurait se passer.

La biologie du futur, qui s'annonce exclusivement massive, force à changer les méthodes d'analyse. En extrapolant, les entités mesurées ne sont plus des entités biologiques mais statistiques. Ces entités n'auront plus rien d'appréhendable unitairement et auront des dimensions d'analyse de plus en plus nombreuses. A la lumière de notre étude, les tests statistiques classiques sont déjà dépassés par de tels volumes et l'analyse de demain se "résumera" a des comparaisons de "massif à massif", tentant de faire émerger des sous-populations. On a commencé à l'entrevoir avec le facteur hStaf où l'on dispose d'un ensemble de règles qui souffrent toutes d'exceptions et que l'on ne peut écarter. Ces exceptions deviennet donc des modalités de variables: "AVEC ACTACA" , "SANS ACTACA", "AVEC ACTACA REPRESSEUR" pour le seul ACTACA, mais qui trouve également écho dans la localisation, dans la nature des pics, etc. Les méthodes permettant de segmenter les distributions font encore aujourd'hui l'objet de travaux intensifs mais après ? Typifier est en soi une part de réponse à une problématique biologique mais ne laisse pas la même satisfaction que par le passé, d'analyser en profondeur. Cette dérive n'est pas cantonnée au monde de la recherche. Les données de tous les domaines sont désormais ingérables tant dans l'analyse que dans le stockage. Des solutions comme le *cloud computing* permettent désormais de faire transiter et analyser l'information de manière délocalisée sans qu'on sache où elle se trouve ni quelle partie est analysée et signe le début d'une nouvelle ère de l'information "*pick-up*" où le *data mining* a le vent en poupe et où les décideurs de demain sont des ordinateurs se basant sur une information incomplète avec des outils statistiques non-préparés à cela. Le réel enjeu de l'analyse informationnelle repose donc plus dans les enjeux des futures statistiques que sur la thématique (biologie...) à laquelle on l'applique.

Conclusion

Conclusion

Répondant au besoin de replacer les résultats génomiques dans leur contexte informationnel, nous avons développé la base de données "GEnomic COntext" (GeCo). Celle-ci, par une structuration originale et une collection de moteurs compatibles avec une analyse statistique rigoureuse, permet de replacer un ensemble de gènes ou de régions génomiques dans leur contexte en utilisant leur contenu et leur contenant informationnels pour souligner les possibles spécificités de ces entrées. Couplée a son portail d'interrogation convivial, elle intègre des visualisations innovantes permettant un accès personnalisé aux données et des visualisations simplifiées d'une importante quantité d'informations. Les possibles applications de GeCo sont multiples, de l'analyse de régions ci-régulatrices à la détection de signatures contextuelles d'un ensemble de gènes, en passant par le filtrage phylogénétique comme elle permit de le faire dans l'analyse des éléments de réponse à l'acide rétinoïque où elle assura de retrouver des élément connus et identifia de nouvelles cibles qui furent validées expérimentalement.

L'analyse des sites de fixation du facteur hStaf/ZNF143 dans le génome humain tira pleinement partie de la puissance de GeCo. Son étude bioinformatique en amont, reposant sur les connaissances biologiques accumulées sur ce facteur, nous a permis d'identifier les pistes de recherche. Ainsi, en s'attachant aux promoteurs bidirectionnels, nous avons pu mettre en évidence le rôle important joué par hStaf dans cet agencement de gènes dont il régulerait 1/5ème des cas. L'analyse expérimentale par ChIP et par gènes rapporteur a permis de confirmer ce rôle et 94% des sites testés par ces deux méthodes ont répondu positivement, attestant de la qualité de nos prédictions. Une étude à haut-débit par ChIP-Seq, a alors été menée et mit en évidence un nombre important de 4088 pics. Toutefois, près de la moitié ne contenait pas une de nos prédictions pourtant nombreuses. C'est cette fois en aval que GeCo démontra sa puissance, en dressant un profil contextuel des pics contenant un site et en le propageant à ceux n'en contenant de prime abord pas. Par un jeu de purification des populations de sites, elle put aider les modèles mathématiques à retrouver leur robustesse et faire émerger des sites dans la quasi-totalité des cas. L'analyse des 2314 gènes cibles donc 2193 protéines identifia des enrichissement fonctionnels en gènes impliqués dans la transcription, le cycle

cellulaire et la réparation, plaçant hStaf très en amont des cascades de régulation et régulant parfois d'importantes sous-familles de gènes co-localisés.

A la lumière de cette étude, l'analyse par le contexte informationnel arrive donc en complément des méthodes mathématiques et de l'expérimentale. Cette approche multidisciplinaire rendue possible dans le cadre d'une collaboration entre l'unité de biologie moléculaire ARN de l'IBMC et l'équipe de bioinformatique LBGI de l'IGBMC permit de dépasser les limites de chaque discipline et d'identifier puis valider les pistes biologiques pour accéder au message sous-jacent.

Bibliographie

Bibliographie

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control *19*, 716–723.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T.*, et al.* (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203-207.

Aravin, A.A., Hannon, G.J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science *318*, 761-764.

Arya, G., and Schlick, T. (2009). A tale of tails: how histone tails mediate chromatin compaction in different salt and linker histone environments. J Phys Chem A *113*, 4045-4059.

Babbitt, G.A., Tolstorukov, M.Y., and Kim, Y. (2010). The molecular evolution of nucleosome positioning through sequence-dependent deformation of the DNA polymer. J Biomol Struct Dyn *27*, 765-780.

Babushok, D.V., Ostertag, E.M., and Kazazian, H.H., Jr. (2007). Current topics in genome evolution: molecular mechanisms of new gene formation. Cell Mol Life Sci *64*, 542-554.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol *2*, 28-36.

Bailey, T.L., and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics *14*, 48-54.

Balmer, J.E., and Blomhoff, R. (2009). Evolution of transcription factor binding sites in mammalian gene regulatory regions: handling counterintuitive results. J Mol Evol *68*, 654-664.

Bard, N., Bolze, R., Caron, E., Desprez, F., Heymann, M., Friedrich, A., Moulinier, L., Nguyen, N.H., Poch, O., and Toursel, T. (2010). Decrypthon grid - grid resources dedicated to neuromuscular disorders. Stud Health Technol Inform *159*, 124-133.

Barnes, P.J. (2009). Targeting the epigenome in the treatment of asthma and chronic obstructive pulmonary disease. Proc Am Thorac Soc *6*, 693-696.

Barski, O.A., Papusha, V.Z., Kunkel, G.R., and Gabbay, K.H. (2004). Regulation of aldehyde reductase expression by Staf and CHOP. Genomics *83*, 119-129.

Bellefroid, E.J., Marine, J.C., Ried, T., Lecocq, P.J., Riviere, M., Amemiya, C., Poncelet, D.A., Coulie, P.G., de Jong, P., Szpirer, C.*, et al.* (1993). Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymphoid cells. EMBO J *12*, 1363-1374.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K.*, et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bertone, P., Gerstein, M., and Snyder, M. (2005). Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Res *13*, 259-274.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E.*, et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pîlot project. Nature *447*, 799-816.

Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D.*, et al.* (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res *16*, 656-668.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D*., et al.* (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res *14*, 708-715.

Blencowe, B.J. (2006). Alternative splicing: new insights from global analyses. Cell *126*, 37-47.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science *299*, 1391-1394.

Bolden, J.E., Peart, M.J., and Johnstone, R.W. (2006). Anticancer activities of histone deacetylase inhibitors. Nat Rev Drug Discov *5*, 769-784.

Bono, H., Kasukawa, T., Furuno, M., Hayashizaki, Y., and Okazaki, Y. (2002). FANTOM DB/ database of Functional Annotation of RIKEN Mouse cDNA Clones. Nucleic Acids Res *30*, 116-118.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311-322.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M*., et al.* (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol *18*, 630-634.

Brunner, A.L., Johnson, D.S., Kim, S.W., Valouev, A., Reddy, T.E., Neff, N.F., Anton, E., Medina, C., Nguyen, L., Chiao, E*., et al.* (2009). Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res *19*, 1044-1056.

Bulyk, M.L., Johnson, P.L., and Church, G.M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res *30*, 1255-1261.

Bush, E.W., and McKinsey, T.A. (2010). Protein acetylation in the cardiorenal axis: the promise of histone deacetylase inhibitors. Circ Res *106*, 272-284.

Carbon, P., and Krol, A. (1991). Transcription of the Xenopus laevis selenocysteine tRNA(Ser)Sec gene: a system that combines an internal B box and upstream elements also found in U6 snRNA genes. EMBO J *10*, 599-606.

Carbon, P., Murgo, S., Ebel, J.P., Krol, A., Tebb, G., and Mattaj, L.W. (1987). A common octamer motif binding protein is involved in the transcription of U6 snRNA by RNA polymerase III and U2 snRNA by RNA polymerase II. Cell *51*, 71-79.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics *21*, 2933-2942.

Cedar, H., Solage, A., Glaser, G., and Razin, A. (1979). Direct detection of methylated cytosine in DNA by use of the restriction enzyme MspI. Nucleic Acids Res *6*, 2125-2132.

Chang, L.W., Nagarajan, R., Magee, J.A., Milbrandt, J., and Stormo, G.D. (2006). A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. Genome Res *16*, 405-413.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res *13*, 3021-3030.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Das, D., Banerjee, N., and Zhang, M.Q. (2004). Interacting models of cooperative gene regulation. Proc Natl Acad Sci U S A *101*, 16234-16239.

de la Grange, P., Gratadou, L., Delord, M., Dutertre, M., and Auboeuf, D. (2010). Splicing factor and exon profiling across human tissues. Nucleic Acids Res *38*, 2825-2838.

Di Leva, F., Ferrante, M.I., Demarchi, F., Caravelli, A., Matarazzo, M.R., Giacca, M., D'Urso, M., D'Esposito, M., and Franze, A. (2004). Human synaptobrevin-like 1 gene basal transcription is regulated

through the interaction of selenocysteine tRNA gene transcription activating factor-zinc finger 143 factors with evolutionary conserved cis-elements. J Biol Chem *279*, 7734-7739.

Elnitski, L., Hardison, R.C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M.J., Schwartz, S., Miller, W., and Chiaromonte, F. (2003). Distinguishing regulatory DNA from neutral sites. Genome Res *13*, 64-72.

Elnitski, L.L., Shah, P., Moreland, R.T., Umayam, L., Wolfsberg, T.G., and Baxevanis, A.D. (2007). The ENCODEdb portal: simplified access to ENCODE Consortium data. Genome Res *17*, 954-959.

Faulkner, G.J., and Carninci, P. (2009). Altruistic functions for selfish DNA. Cell Cycle *8*, 2895-2900.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T.*, et al.* (2009). The regulated retrotransposon transcriptome of mammalian cells. Nat Genet *41*, 563-571.

Fedorova, E., and Zink, D. (2008). Nuclear architecture and gene regulation. Biochim Biophys Acta *1783*, 2174-2184.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. Nature *421*, 448-453.

Finocchiaro, G., Carro, M.S., Francois, S., Parise, P., DiNinni, V., and Muller, H. (2007). Localizing hotspots of antisense transcription. Nucleic Acids Res *35*, 1488-1500.

Gabdank, I., Barash, D., and Trifonov, E.N. (2009). Nucleosome DNA bendability matrix (C. elegans). J Biomol Struct Dyn *26*, 403-411.

Gerard, M.A., Krol, A., and Carbon, P. (2007). Transcription factor hStaf/ZNF143 is required for expression of the human TFAM gene. Gene *401*, 145-153.

Gerard, M.A., Myslinski, E., Chylak, N., Baudrey, S., Krol, A., and Carbon, P. (2010). The scaRNA2 is produced by an independent transcription unit and its processing is directed by the encoding region. Nucleic Acids Res *38*, 370-381.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. Genome Res *17*, 669-681.

Goldstein, B.A., Hubbard, A.E., Cutler, A., and Barcellos, L.F. (2010). An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. BMC Genet *11*, 49.

Gordan, R., and Hartemink, A.J. (2008). Using DNA duplex stability information for transcription factor binding site discovery. Pac Symp Biocomput, 453-464.

Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P.M.*, et al.* (2010). Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. Nat Methods *7*, 47-49.

Gross, D.S., and Garrard, W.T. (1988). Nuclease hypersensitive sites in chromatin. Annu Rev Biochem *57*, 159-197.

Grossman, C.E., Qian, Y., Banki, K., and Perl, A. (2004). ZNF143 mediates basal and tissue-specific expression of human transaldolase. J Biol Chem *279*, 12190-12205.

Grunstein, M. (1992). Histones as regulators of genes. Sci Am *267*, 68-74B.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol *8*, R24.

Hasler, J., Samuelsson, T., and Strub, K. (2007). Useful 'junk': Alu RNAs in the human transcriptome. Cell Mol Life Sci *64*, 1793-1800.

Hochschild, A., and Ptashne, M. (1986). Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix. Cell *44*, 681-687.

Holliday, R., and Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science *187*, 226-232.

Ishiguchi, H., Izumi, H., Torigoe, T., Yoshida, Y., Kubota, H., Tsuji, S., and Kohno, K. (2004). ZNF143 activates gene expression in response to DNA damage and binds to cisplatin-modified DNA. Int J Cancer *111*, 900-909.

Jaenisch, R. (1997). DNA methylation and imprinting: why bother? Trends Genet *13*, 323-329.
Johnson, A.D. (2010). An extended IUPAC nomenclature code for polymorphic nucleic acids. Bioinformatics *26*, 1386-1389.
Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.
Jurka, J. (2004). Evolutionary impact of human Alu repetitive elements. Curr Opin Genet Dev *14*, 603-608.

Kazazian, H.H., Jr. (2004). Mobile elements: drivers of genome evolution. Science *303*, 1626-1632.
Keene, M.A., Corces, V., Lowenhaupt, K., and Elgin, S.C. (1981). DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. Proc Natl Acad Sci U S A *78*, 143-146.
Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res *31*, 3576-3579.
Kim, Y., McLaughlin, N., Lindstrom, K., Tsukiyama, T., and Clark, D.J. (2006). Activation of Saccharomyces cerevisiae HIS3 results in Gcn4p-dependent, SWI/SNF-dependent mobilization of nucleosomes over the entire gene. Mol Cell Biol *26*, 8607-8622.
Koshibu, K., Graff, J., Beullens, M., Heitz, F.D., Berchtold, D., Russig, H., Farinelli, M., Bollen, M., and Mansuy, I.M. (2009). Protein phosphatase 1 regulates the histone code for long-term memory. J Neurosci *29*, 13079-13089.
Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods *6*, 291-295.
Krogh, A., and Mitchison, G. (1995). Maximum entropy weighting of aligned sequences of proteins or DNA. Proc Int Conf Intell Syst Mol Biol *3*, 215-221.
Kubota, H., Yokota, S., Yanagi, H., and Yura, T. (2000). Transcriptional regulation of the mouse cytosolic chaperonin subunit gene Ccta/t-complex polypeptide 1 by selenocysteine tRNA gene transcription activating factor family zinc finger proteins. J Biol Chem *275*, 28641-28648.

Laudet, V., Hanni, C., Coll, J., Catzeflis, F., and Stehelin, D. (1992). Evolution of the nuclear receptor gene superfamily. EMBO J *11*, 1003-1013.
Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science *262*, 208-214.
Lerat, E., and Semon, M. (2007). Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. Gene *396*, 303-311.
Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G.*, et al.* (2007). The diploid genome sequence of an individual human. PLoS Biol *5*, e254.
Liang, Y., Vogel, J.L., Narayanan, A., Peng, H., and Kristie, T.M. (2009). Inhibition of the histone demethylase LSD1 blocks alpha-herpesvirus lytic replication and reactivation from latency. Nat Med *15*, 1312-1317.
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, E.T., Pott, S., and Huss, M. (2010). Q&A/ ChIP-seq technologies and the study of gene regulation. BMC Biol *8*, 56.
Liu, X., Brutlag, D.L., and Liu, J.S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput, 127-138.

Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., *et al.* (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet *38*, 431-440.
Lomvardas, S., and Thanos, D. (2001). Nucleosome sliding via TBP DNA binding in vivo. Cell *106*, 685-696.
Lusk, R.W., and Eisen, M.B. (2008). Use of an evolutionary model to provide evidence for a wide heterogeneity of required affinities between transcription factors and their binding sites in yeast. Pac Symp Biocomput, 489-500.

Mach, C.M., Hargrove, B.W., and Kunkel, G.R. (2002). The Small RNA gene activator protein, SphI postoctamer homology-binding factor/selenocysteine tRNA gene transcription activating factor, stimulates transcription of the human interferon regulatory factor-3 gene. J Biol Chem *277*, 4853-4858.
Marguerat, S., and Bahler, J. (2010). RNA-seq: from technology to biology. Cell Mol Life Sci *67*, 569-579.
Margulies, E.H., Maduro, V.V., Thomas, P.J., Tomkins, J.P., Amemiya, C.T., Luo, M., and Green, E.D. (2005). Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. Proc Natl Acad Sci U S A *102*, 3354-3359.
Marstrand, T.T., Frellsen, J., Moltke, I., Thiim, M., Valen, E., Retelska, D., and Krogh, A. (2008). Asap: a framework for over-representation statistics for transcription factor binding sites. PLoS One *3*, e1623.
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.
McCoy, M.W., Allen, A.P., and Gillooly, J.F. (2009). The random nature of genome architecture: predicting open reading frame distributions. PLoS One *4*, e6456.
McGhee, J.D., Wood, W.I., Dolan, M., Engel, J.D., and Felsenfeld, G. (1981). A 200 base pair region at the 5' end of the chicken adult beta-globin gene is accessible to nuclease digestion. Cell *27*, 45-55.
Mendez-Acuna, L., Di Tomaso, M.V., Palitti, F., and Martinez-Lopez, W. (2010). Histone post-translational modifications in DNA damage response. Cytogenet Genome Res *128*, 28-36.
Miller, J., McLachlan, A.D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J *4*, 1609-1614.
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.
Murphy, J.T., Skuzeski, J.T., Lund, E., Steinberg, T.H., Burgess, R.R., and Dahlberg, J.E. (1987). Functional elements of the human U1 RNA promoter. Identification of five separate regions required for efficient transcription and template competition. J Biol Chem *262*, 1795-1803.
Myslinski, E., Gerard, M.A., Krol, A., and Carbon, P. (2006). A genome scale location analysis of human Staf/ZNF143-binding sites suggests a widespread role for human Staf/ZNF143 in mammalian promoters. J Biol Chem *281*, 39953-39962.
Myslinski, E., Gerard, M.A., Krol, A., and Carbon, P. (2007). Transcription of the human cell cycle regulated BUB1B gene requires hStaf/ZNF143. Nucleic Acids Res *35*, 3453-3464.
Myslinski, E., Krol, A., and Carbon, P. (1998). ZNF76 and ZNF143 are two human homologs of the transcriptional activator Staf. J Biol Chem *273*, 21998-22006.
Myslinski, E., Schuster, C., Huet, J., Sentenac, A., Krol, A., and Carbon, P. (1993). Point mutations 5' to the tRNA selenocysteine TATA box alter RNA polymerase III transcription by affecting the binding of TBP. Nucleic Acids Res *21*, 5852-5858.

Nishida, K., Frith, M.C., and Nakai, K. (2009). Pseudocounts for transcription factor binding sites. Nucleic Acids Res *37*, 939-944.
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.

Pan, W., Wei, P., and Khodursky, A. (2008). A parametric joint model of DNA-protein binding, gene expression and DNA sequence data to detect target genes of a transcription factor. Pac Symp Biocomput, 465-476.

Peaston, A.E., Evsikov, A.V., Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D., and Knowles, B.B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell *7*, 597-606.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res *38*, D105-110.

Prakash, A., and Tompa, M. (2005). Discovery of regulatory elements in vertebrates through comparative genomics. Nat Biotechnol *23*, 1249-1256.

Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V*., et al.* (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science *317*, 86-94.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res *23*, 4878-4884.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N*., et al.* (2010). An atlas of combinatorial transcriptional regulation in mouse and man. Cell *140*, 744-752.

Roh, T.Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A *103*, 15782-15787.

Rousseau-Merck, M.F., Koczan, D., Legrand, I., Moller, S., Autran, S., and Thiesen, H.J. (2002). The KOX zinc finger genes: genome wide mapping of 368 ZNF PAC clones with zinc finger gene clusters predominantly in 23 chromosomal loci are confirmed by human sequences annotated in EnsEMBL. Cytogenet Genome Res *98*, 147-153.

Salih, F., Salih, B., Kogan, S., and Trifonov, E.N. (2008a). Epigenetic nucleosomes: Alu sequences and CG as nucleosome positioning element. J Biomol Struct Dyn *26*, 9-16.

Salih, F., Salih, B., and Trifonov, E.N. (2008b). Sequence structure of hidden 10.4-base repeat in the nucleosomes of C. elegans. J Biomol Struct Dyn *26*, 273-282.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet *8*, 424-436.

Saur, D., Seidler, B., Paehge, H., Schusdziarra, V., and Allescher, H.D. (2002). Complex regulation of human neuronal nitric-oxide synthase exon 1c gene transcription. Essential role of Sp and ZNF family members of transcription factors. J Biol Chem *277*, 25798-25814.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A *103*, 1412-1417.

Schaub, M., Krol, A., and Carbon, P. (1999a). Flexible zinc finger requirement for binding of the transcriptional activator staf to U6 small nuclear RNA and tRNA(Sec) promoters. J Biol Chem *274*, 24241-24249.

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

Schaub, M., Myslinski, E., Krol, A., and Carbon, P. (1999b). Maximization of selenocysteine tRNA and U6 small nuclear RNA transcriptional activation achieved by flexible utilization of a Staf zinc finger. J Biol Chem *274*, 25042-25050.
Schaub, M., Myslinski, E., Schuster, C., Krol, A., and Carbon, P. (1997). Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III. EMBO J *16*, 173-181.
Schuster, C., Krol, A., and Carbon, P. (1998). Two distinct domains in Staf to selectively activate small nuclear RNA-type and mRNA promoters. Mol Cell Biol *18*, 2650-2658.
Schuster, C., Myslinski, E., Krol, A., and Carbon, P. (1995). Staf, a novel zinc finger protein that activates the RNA polymerase III promoter of the selenocysteine tRNA gene. EMBO J *14*, 3777-3787.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res *13*, 103-107.
Shalon, D., Smith, S.J., and Brown, P.O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res *6*, 639-645.
Siomi, M.C., Mannen, T., and Siomi, H. (2010). How does the royal family of Tudor rule the PIWI-interacting RNA pathway? Genes Dev *24*, 636-646.
Song, L., and Crawford, G.E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc *2010*, pdb prot5384.
Spannhoff, A., Hauser, A.T., Heinke, R., Sippl, W., and Jung, M. (2009). The emerging therapeutic potential of histone methyltransferase and demethylase inhibitors. ChemMedChem *4*, 1568-1582.
Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol *21*, 1973-1985.
Stormo, G.D., and Tan, K. (2002). Mining genome databases to identify and understand new gene regulatory systems. Curr Opin Microbiol *5*, 149-153.

Taft, R.J., Glazov, E.A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G.J., Lassmann, T., Forrest, A.R., Grimmond, S.M., Schroder, K.*, et al.* (2009). Tiny RNAs associated with transcription start sites in animals. Nat Genet *41*, 572-578.
Tanaka, Y., Yamashita, R., Suzuki, Y., and Nakai, K. (2010). Effects of Alu elements on global nucleosome positioning in the human genome. BMC Genomics *11*, 309.
Tang, F. (2010). Small RNAs in mammalian germline: Tiny for immortal. Differentiation *79*, 141-146.
Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics *17*, 1113-1122.
Tsankov, A.M., Thompson, D.A., Socha, A., Regev, A., and Rando, O.J. (2010). The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol *8*, e1000414.
Tycowski, K.T., Aab, A., and Steitz, J.A. (2004). Guide RNAs with 5' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. Curr Biol *14*, 1985-1995.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Wakasugi, T., Izumi, H., Uchiumi, T., Suzuki, H., Arao, T., Nishio, K., and Kohno, K. (2007). ZNF143 interacts with p73 and is involved in cisplatin resistance through the transcriptional regulation of DNA repair genes. Oncogene *26*, 5194-5203.
Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57-63.
Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z.*, et al.* (2006). A global map of p53 transcription-factor binding sites in the human genome. Cell *124*, 207-219.
Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res *24*, 238-241.

Wold, B., and Myers, R.M. (2008). Sequence census methods for functional genomics. Nat Methods *5*, 19-21.
Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B.*, et al.* (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725-737.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.
Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E.S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci U S A *104*, 7145-7150.
Xu, D., Liu, H.J., and Wang, Y.F. (2005). BSS-HMM3s: an improved HMM method for identifying transcription factor binding sites. DNA Seq *16*, 403-411.
Yagi, S., Hirabayashi, K., Sato, S., Li, W., Takahashi, Y., Hirakawa, T., Wu, G., Hattori, N., Ohgane, J., Tanaka, S.*, et al.* (2008). DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) in mouse promoter regions demonstrating tissue-specific gene expression. Genome Res *18*, 1969-1978.

Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K., and Sugano, S. (2006). DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. Nucleic Acids Res *34*, D86-89.
Yamashita, R., Wakaguri, H., Sugano, S., Suzuki, Y., and Nakai, K. (2010). DBTSS provides a tissue specific dynamic view of Transcription Start Sites. Nucleic Acids Res *38*, D98-104.
Yu, X., Lin, J., Zack, D.J., and Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. Nucleic Acids Res *34*, 4925-4936.

Zhang, J. (2000). Protein-length distributions for the three domains of life. Trends Genet *16*, 107-109.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.
Zhang, Z., and Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol *2*, 11.
Zhou, Q., and Liu, J.S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics *20*, 909-916.