Signatures nucléotidiques de l'activité des enhancers développementaux chez l'ascidie Ciona intestinalis

Marion Gueroult Bellone

▶ **To cite this version:**

Marion Gueroult Bellone. Signatures nucléotidiques de l'activité des enhancers développementaux chez l'ascidie Ciona intestinalis. Biologie cellulaire. Université Montpellier, 2016. Français. NNT : 2016MONTS029 . tel-03215429

HAL Id: tel-03215429
https://theses.hal.science/tel-03215429

Submitted on 3 May 2021

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

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

Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de Docteur

Délivrée par l'**Université de Montpellier**

Préparée au sein de l'école doctorale **CBS2**
Et du **Centre de Recherche de Biochimie Macromoléculaire**

Spécialité : **Biologie cellulaire**

Présentée par **Marion Guéroult-Bellone**

Signatures nucléotidiques de l'activité des enhancers développementaux chez l'ascidie *Ciona intestinalis*

Soutenue le 15 janvier 2016 devant le jury composé de

Dr. François SPITZ	EMBL, Heidelberg	Rapporteur
Dr. Hitoyoshi YASUO	LBDV, Villefranche s/ Mer	Rapporteur
Dr. François PAYRE	CBD, Toulouse	Examinateur
Pr. Stephen BAGHDIGUIAN	Université de Montpellier	Examinateur
Dr. Patrick LEMAIRE	CRBM, Montpellier	Directeur de thèse
Dr. Jacques PIETTE	CRBM, Montpellier	Co-directeur de thèse

REMERCIEMENTS

Je tiens à exprimer mes sincères remerciements à tous ceux et celles qui m'ont donné les moyens d'entreprendre une thèse et apporté le soutien pour la mener à bien.

Mes premiers remerciements vont à mes directeurs de thèse, dont les rôles complémentaires d'encadrement m'ont permis de me construire en tant que jeune chercheuse au quotidien et dans la durée. Patrick, merci d'avoir cru en moi il y a sept ans, et de m'avoir soutenue depuis, à chaque étape de mon parcours, dans la réussite comme dans la galère. Jacques, j'ai eu la chance que tu t'impliques totalement dans ma thèse, alors que ton encadrement n'était au départ qu'administratif, merci. Je suis fière d'avoir été ta dernière thésarde !

Je remercie Hitoyoshi Yasuo, François Spitz, François Payre et Stephen Baghdiguian, qui ont accepté sans hésitation, de faire partie de mon jury de thèse, avec une pensée spéciale pour Yasuo qui a déjà corrigé mes deux rapports de stage de M2.

Je remercie également les membres de mon comité de thèse, Hitoyoshi Yasuo, Robert Feil et Emmanuel Douzery pour leur bienveillance et les conseils qui ont permis d'encrer mon projet dans des réalités temporelles et matérielles.

Je tiens à remercier mes collègues actuels et passés, pour la bonne ambiance, et tous les bons moments passés ensemble. Ulla, Mathieu et Maggie, avec qui nous avons mis en marche le labo Lemaire2.0, Cyril, Edwin, Homa, Emilie, Delphine, Léo, Christelle, Matija, Alicia, mes géniaux stagiaires, Andrès et Carine, et les autres étudiants de passage. C'était un plaisir de travailler à vos côtés, vous allez me manquer !

Merci Pierre pour ta gentillesse et le suivi à distance de mes travaux qui sont la suite logique de ta thèse.

J'ai une pensée émue pour les nombreuses ciones mortes pour la science et pour le brin de fantaisie du labo : nos mascottes Séb le crabe et Patrick l'étoile de mer.

Ilaria, Cao, Sheng, Thierry, Séb, François, Khaled, Marie-Odile, Abdullah, Mathy, Ngoc, Nikola, Laetitia, habitants du CRBM et voisins, votre bonne humeur, vos blagues, votre gentillesse, font de ce labo un lieu accueillant à toute heure du jour ou de la nuit.

Merci à Sandrine Urvoy, Géraldine Cubéro, Monique Miras, Caty Morel et Michel Désarménien pour la gestion administrative de ma chaotique réinscription en 4ème année.

J'ai eu le plaisir d'enseigner à l'université pendant trois ans, j'en remercie les équipes de TP des UE biocel et microbio de L1 qui m'ont tant appris, et les étudiants qui donnent envie de continuer dans cette voie.

Je tiens à remercier du fond du cœur ma famille qui a toujours su respecter mes choix et m'encourager. Je remercie mes parents qui m'ont donné la liberté et les moyens de faire les études que je voulais dans de bonnes conditions, Anouk, Titou, mes grands-parents, parrain, marraine, oncle et tante, cousins qui se sont toujours intéressés à mes travaux sur ces « drôles d'éponges bizarres ».

Je remercie mes colocataires, voisine, sous-locataires, squatteurs du salon, voyageurs de passage, pour la bonne humeur, le réconfort, les bons repas et les discussions passionnées sans lesquels je ne me serais jamais sentie chez moi à Montpellier. Je tiens également à remercier mes amis, qui sont restés présents quand je l'étais beaucoup moins.

Enfin, je suis très reconnaissante au Docteur Bouix dont la clairvoyance m'a permis de ne pas passer tout à fait à côté de ma thèse.

RÉSUMÉ

Les enhancers sont des régulateurs cruciaux de l'expression des gènes pendant le développement embryonnaire. L'ascidie *Ciona intestinalis* est un organisme-modèle qui se prête à l'étude de ces séquences *cis*-régulatrices car ses enhancers sont généralement petits et compacts, et le lignage invariant des cellules chez l'embryon permet de visualiser leur activité avec une résolution cellulaire.

Deux signatures indépendantes associées à l'activité d'un enhancer avaient été identifiées : la présence de sites de fixation pour des facteurs de transcription spécifiques, et une signature dinucléotidique globale à l'échelle des enhancers. (Khoueiry 2010). Cependant, si ces signatures corrèlent avec l'activité des enhancers, elles ne permettent pas d'identifier de nouveaux enhancers grâce à leur séquence.

Pendant ma thèse, j'ai utilisé un enhancer neural précoce de *Ciona*, le très bien caractérisé élément-a du gène *Otx*, comme enhancer-modèle. Ce petit enhancer (55pb), est lié par les facteurs de transcription Gata4/5/6 et Ets1/2 et activé par la voie de signalisation FGF. Afin de mieux comprendre les déterminants de l'activité neurale précoce d'un enhancer, j'ai testé l'impact de mutations ponctuelles affectant l'affinité de sites de fixation de l'élément-a pour les facteurs de transcription. J'ai également randomisé les spacers, séquences situées entre les sites de fixation pour ETS et GATA dans quatre clusters de ces sites.

Nos résultats suggèrent au moins deux niveaux de contrôle de la régulation en *cis* : i) la spécificité spatiotemporelle de l'activité d'un enhancer est définie par l'identité des sites de fixation des facteurs de transcription, et ii) son niveau d'activité dépend à la fois de l'affinité des facteurs de transcription pour leurs sites de fixation et la composition des spacers. La majorité des variants randomisés de l'élément-a sont actifs dans les mêmes lignées cellulaires que le sauvage et leurs niveaux d'activité sont très divers. Le même résultat est obtenu en randomisant les séquences intercalantes d'un autre cluster ETS/GATA actif. La randomisation de ces séquences a même conféré de l'activité enhancer à de nombreux variants de clusters inactifs. En accord avec leur activité neurale précoce et la présence de sites de fixations pour ETS et GATA, ces variants, comme l'élément-a, répondent à l'induction neurale de FGF.

Nous n'avons pas réussi à expliquer l'action des séquences intercalantes sur l'activité des enhancers par des caractéristiques simples de leurs séquences (nucléotidique ou dinucléotidique), et l'on ne comprend pas pourquoi il est si simple de créer un enhancer synthétique quand la majorité des clusters génomiques de sites de fixations putatifs pour ETS et GATA sont inactifs. En utilisant une approche de fixation *in vitro* des facteurs de transcription, nous avons montré que la randomisation des séquences intercalantes peut affecter la fixation d'un facteur de transcription sur l'élément a, sans changer la séquence primaire du site de fixation, mais que la fixation sur l'élément entier ne peut pas toujours être expliquée par la fixation sur les sites isolées. Ces résultats suggèrent que la structure physique de l'hélice d'ADN des enhancers peut jouer un rôle important dans le contrôle de l'activité d'un gène.

ABSTRACT

Enhancers are crucial elements for the control of gene expression during embryonic development. The ascidian *Ciona intestinalis* offers unique experimental features to study these *cis*-regulatory sequences: enhancers are generally small and compact and their activity can be tracked at the single cell level thanks to the invariant cell lineage of ascidian embryos.

Previous work identified two independent signatures associated with enhancer activity: the presence of specific transcription factors binding sites (TFBS) and a global dinucleotide signature along enhancers (Khoueiry, 2010). Although they correlate with enhancer activity, these signatures are insufficient to identify enhancer sequences from their sole sequence.

During my thesis, I used a well-characterized early neural *Ciona* enhancer, the a-element of the *Otx* gene, as a model enhancer. This small (55pb) enhancer, is bound by Gata4/5/6 and Ets1/2 and is activated by the FGF pathway. To better understand the determinants of early neural enhancer activity, I tested the impact of point mutations affecting the affinity of the a-element TFBS for their binding TF and of the randomization of the spacer sequences that separate the TFBS in four ETS and GATA binding site clusters.

Our results suggest at least two levels of *cis*-regulatory control: spatiotemporal specificity of enhancer activity is encoded in the identity of TF-binding sites, while the level of enhancer activity is set both by the affinity of TFs for their binding sites and by the composition of the spacer sequences. A surprisingly high number of variants of the a-element with randomized spacers are active, always in the same cell lineages as the WT. These variants, however, display a wide range of activity levels. This effect is also observed when the spacers in another active ETS/GATA cluster are randomized. Randomization of the spacers can even confer enhancer activity to a large fraction of inactive cluster variants. Consistent with their early neural activity and with the presence of ETS- and GATA-binding sites, these variants are, like the a-element, responsive to the FGF neural inducer.

We could not link the action of the spacers on enhancer activity to any simple nucleotide or dinucleotide sequence features and it currently remains unclear why it is so easy to create a synthetic enhancer while most putative genomic ETS/GATA clusters are inactive. Using in vitro transcription factor binding assays, we showed that randomization of spacer sequences can affect TF binding to the a-element without changing the primary sequence of the binding site, and that extended minimal TFBS do not always recapitulate binding to the whole element. These results suggest that the physical structure of the DNA helix around the binding sites may play an important role in the control of enhancer activity.

RÉSUMÉS GRAND PUBLIC

Les enhancers sont des séquences d'ADN qui contrôlent l'expression des gènes. Leur rôle est très important au cours du développement : toutes les cellules d'un embryon contiennent le même matériel génétique, mais, selon les gènes qui y sont exprimés, elles vont former différents tissus, organes… Pendant ma thèse, j'ai essayé de comprendre comment est codée l'information dans ces séquences, chez l'ascidie *Ciona intestinalis*, un petit invertébré marin. Les embryons de cette espèce se développent rapidement, ont un petit nombre de cellules qui se divisent toujours de la même façon, et l'on peut facilement y introduire de l'ADN contenant la séquence d'intérêt et visualiser dans quelle(s) cellule(s) elle a la capacité d'activer un gène.

On sait que les enhancers sont activés par des protéines qui se fixent sur des séquences qu'elles reconnaissent dans ceux-ci. J'ai montré que les séquences situées entre ces sites de fixation jouent également un rôle important pour l'activité des enhancers.

* * * * *

Enhancers are DNA sequences, which control the expression of neighboring genes. They are important during development, as all the cells of an embryo contain the same genetic information, but, according to the genes they express, will specialize in different tissues, organs... During my PhD, I aimed at understanding how is spatio-temporal activity encoded in these sequences, using a small marine invertebrate, the sea squirt *Ciona intestinalis* as a model organism. Embryos of this species are very suitable to study gene regulation, as they develop quickly, have a small number of cells that divide in an invariant and stereotyped manner, and one can easily introduce DNA inside and visualize in which cell(s) it is able to activate gene expression.

Enhancers are activated by the fixation of proteins on specific sequences. I showed that sequences located between these binding sites are also very important for enhancers, as they set their activity levels.

TABLE DES MATIÈRES

GLOSSAIRE

3C/ CCC : Capture de la conformation des chromosomes (Chromosome Conformation Capture) et ses variants : **4C** (Circularized Chromosome Conformation Capture), **5C** (Carbon Copy Chromosome Conformation Capture), **Hi-C** (High Chromosome Contact Capture)

ADN : Acide DésoxyriboNucléique

A : Adénine

C : Cytosine

G : Guanine

T : Thymine

ADP : Adénosine DiPhosphate

ARN : Acide RiboNucléique

ARN-pol II : ARN polymérase II

ARNe : ARN « enhancer »

ARNm : ARN messager

ATAC-seq : Identification de la chromatine accessible aux transposases et séquençage haut-débit (Assay for Transposase-Accessible Chromatin with highthroughput SEQuencing)

BMP : Bone Morphogenetic Protein

C2H2 : Doigt de zinc contenant 2 Cystéines et 2 Histidines

ChIP-on-chip : Immunoprécipitation de la chromatine sur puce à ADN (Chromatin ImmunoPrecipitation and DNA microarray)

ChIP-seq : Immunoprecipitation de la chromatine et séquençage (Chromatin Immuno-Precipitation and sequencing)

CNS : Système nerveux central (Central Nervous System)

CpG : Cytosine-phosphate-Guanine

CRISPR : Courtes répétitions palindromiques groupées et régulièrement espacées (Clustered Regularly Interspaced Short Palindromic Repeats),

crx : Cone-rod homeobox protein

CTCF : Facteur liant CCCTC (CCCTC-binding Factor)

CTD : Domaine C-Terminal (C-Terminal Domain)

DBD : Domaine de liaison à l'ADN (DNA-Binding Domain)

DPE : Element en aval du promoteur (Downstream Promoter Element)

dsx : doublesex

eFS : enhancer FACS-seq

EMSA : Retard sur gel (Electrophoretic Mobility Shift Assay)

ENCODE : Encyclopédie des éléments de l'ADN (Encyclopedie of DNA elements) chez l'humain : **modENCODE** pour les données de nématode et drosophile ; **mouseENCODE** pour la souris

eve : even skipped

FACS : Tri de cellules par fluorescence (Fluorescent Activated Cell Sorter)

FAIRE : Isolation d'éléments régulateurs assistée au formaldéhyde (Formaldehyde Assisted Isolation of Regulatory Elements)

FGF : Facteur de croissance des fibroblastes (Fibroblast Growth Factor)

H3K4me3, H3K4me1, H3K27ac (**H3** : histone 3, **K** : Lysine, **ac** : acétylation, **me1** : monométhylation, **me3** : triméthylation)

HOT : Régions fixant un très grand nombre de facteurs de transcription (Highly Occupied Target)

LCR : Région de contrôle d'un locus (Locus control Region)

Mb, kb, pb : Méga base, kilobase, paire de base

NCE : Élément non-codant (Non Coding Element)

PBM : Protein Binding microarrays

PCR : Réactions de polymérisation en chaîne (Polymerase Chain Reaction)

PIC : Complexe de pré-initiation (Pre-Initiation Complex)

QTL : Locus d'un trait quantitatif (Quantitative Trait Loci)

SECOMOD : Recherche de modules conservés au cours de l'évolution (SEarch for Evolutionary COnserved MODules)

SELEX : Evolution systématique des ligands par enrichissements exponentiels (Systematic Evolution of Ligands by EXponential enrichments)

SNP : Polymorphisme d'un seul nucléotide (Single Nucléotide Polymorphism)

STARR-seq : Séquençage de régions régulatrices activant leur propre transcription (Self-Transcribing Active Regulatory Region Sequencing)

TAD : Domaine d'association topologique (Topology Association Domain)

TALEN : Transcription Activator-Like Effector Nucleases

TF : Facteur de transcription (Transcription Factor)

TFBS : Site de fixation de facteur de transcription (Transcription Factor Binding Site)

TSS : Site d'initiation de la transcription (Transcription Start Site)

UCE : Eléments ultra-conservés (Ultra-Conserved Element)

WRPW : Tryptophane-Arginine-Proline-Tryptophane

Y1H : Technique du simple hybride (Yeast one Hybride)

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Le code génétique qui permet de traduire une séquence d'acide désoxyribonucléique (ADN) en un enchaînement d'acides aminés formant une protéine, a été décrypté au début des années 60. Si ce code permet de prédire la protéine produite par un gène codant, il n'explique pas où, quand et en quelle quantité cette protéine est produite au sein de l'organisme. La synthèse d'ARN à partir de l'ADN génomique - la transcription - est la première étape du processus qui conduit de l'ADN à la présence de la protéine. Ce processus hautement régulé permet une interprétation différentielle par chaque type de cellule de la même information génétique. L'information contrôlant la transcription est encodée dans des séquences, dites *cis*-régulatrices, localisées à plus ou moins grande distance du gène qu'elles contrôlent. Ces séquences restent mal comprises.

Déchiffrer le fonctionnement des séquences régulatrices aurait d'importantes retombées sur notre compréhension des processus génétiques de développement et d'évolution, qu'ils soient normaux ou pathologiques. Toutefois, cette tâche est rendue particulièrement ardue par la complexité des différents niveaux de contrôle de l'expression des gènes. Par exemple, la seule initiation de la transcription implique des interactions de dizaines de protéines entre elles et avec l'ADN, mais aussi des changements épigénétiques de la structure locale de la chromatine.

La première partie de mon introduction traitera de la transcription et de ses mécanismes de contrôle ainsi que de certaines caractéristiques des séquences *cis*-régulatrices. Je présenterai ensuite l'ascidie *Ciona intestinalis*, l'organisme modèle sur lequel j'ai travaillé pendant ma thèse et qui se révèle être particulièrement adapté aux études transcriptionnelles.

Plan de l'introduction

I. Le contrôle de la transcription

Comment peut-on produire au sein d'un organisme des cellules différenciées, exprimant des jeux de protéines très différents, alors que ces cellules contiennent toutes, à quelques rares exceptions près, le même ADN ? Selon le « dogme central de la biologie moléculaire » (Crick, 1970), les protéines ne sont pas produites directement à partir de l'ADN. L'ADN est d'abord copié sous forme d'une molécule intermédiaire, l'ARN, dans un processus nommé transcription. Cet ARN est alors traduit en protéine(s) (Figure I.1A). Des percées majeures en biologie moléculaire, biochimie et génétique ont montré que la production d'un grand nombre de types de cellules différentes à partir d'une même molécule d'ADN résulte principalement du contrôle du nombre de copies ARN de chaque gène produites au sein de chaque type de cellule, c'est à dire du contrôle de l'expression des gènes (Britten & Davidson, 1969), même si le contrôle de la traduction de l'ARN en protéine, les modifications post-traductionnelles, la localisation et la dégradation des protéines contribuent également à la construction du phénotype (Figure I.1B). De plus, pendant le développement embryonnaire, l'orchestration précise de l'expression des gènes dans le temps et dans l'espace est indispensable au bon déroulement des événements morphogénétiques qui déterminent la forme finale de l'organisme en devenir et positionnent chaque cellule différenciée au sein de son organe (Peter & Davidson, 2015). Le contrôle de l'expression des gènes joue enfin un rôle majeur au cours de l'évolution (Davidson & Erwin, 2006), et la complexification des mécanismes de contrôle de l'expression des gènes pourrait être à l'origine, ou avoir accompagné, l'émergence des organismes pluricellulaires complexes (Michael Levine & Tjian, 2003). Ces travaux placent donc le contrôle de l'expression des gènes au cœur des processus de différentiation, de morphogénèse et d'évolution de l'embryon.

Pendant ma thèse, je me suis intéressée à cette étape clef du contrôle de l'activité des gènes.
Seule une toute petite partie d'un génome code pour des protéines. La majorité des génomes animaux est donc dite non-codante. Pourtant, c'est précisément au sein de ces séquences que se trouve la majorité de l'information génétique qui contrôle l'expression des gènes. Décoder le mode d'action des séquences *cis*-régulatrices est donc indispensable pour comprendre le développement et l'évolution des espèces. Pourtant, plus de 35 ans après l'identification de la première séquence régulatrice active dans des cellules eucaryotes (M. L. Goldberg, PhD thesis, Stanford University, 1979), les règles qui confèrent des propriétés *cis*-régulatrices à un

fragment d'ADN restent très mal comprises. Ainsi, contrairement aux séquences codantes, personne n'est actuellement capable de prédire l'activité *cis*-régulatrice d'une région d'ADN sur la seule base de sa séquence. Il n'existe donc, pour l'instant, aucun logiciel capable de scanner efficacement un génome à la recherche de séquences régulatrices comme c'est le cas pour les éléments transposables (par exemple RepeatMasker, (Tempel, 2012)) ou les modèles de gènes codants (pour revue, voir Sleator, 2010).

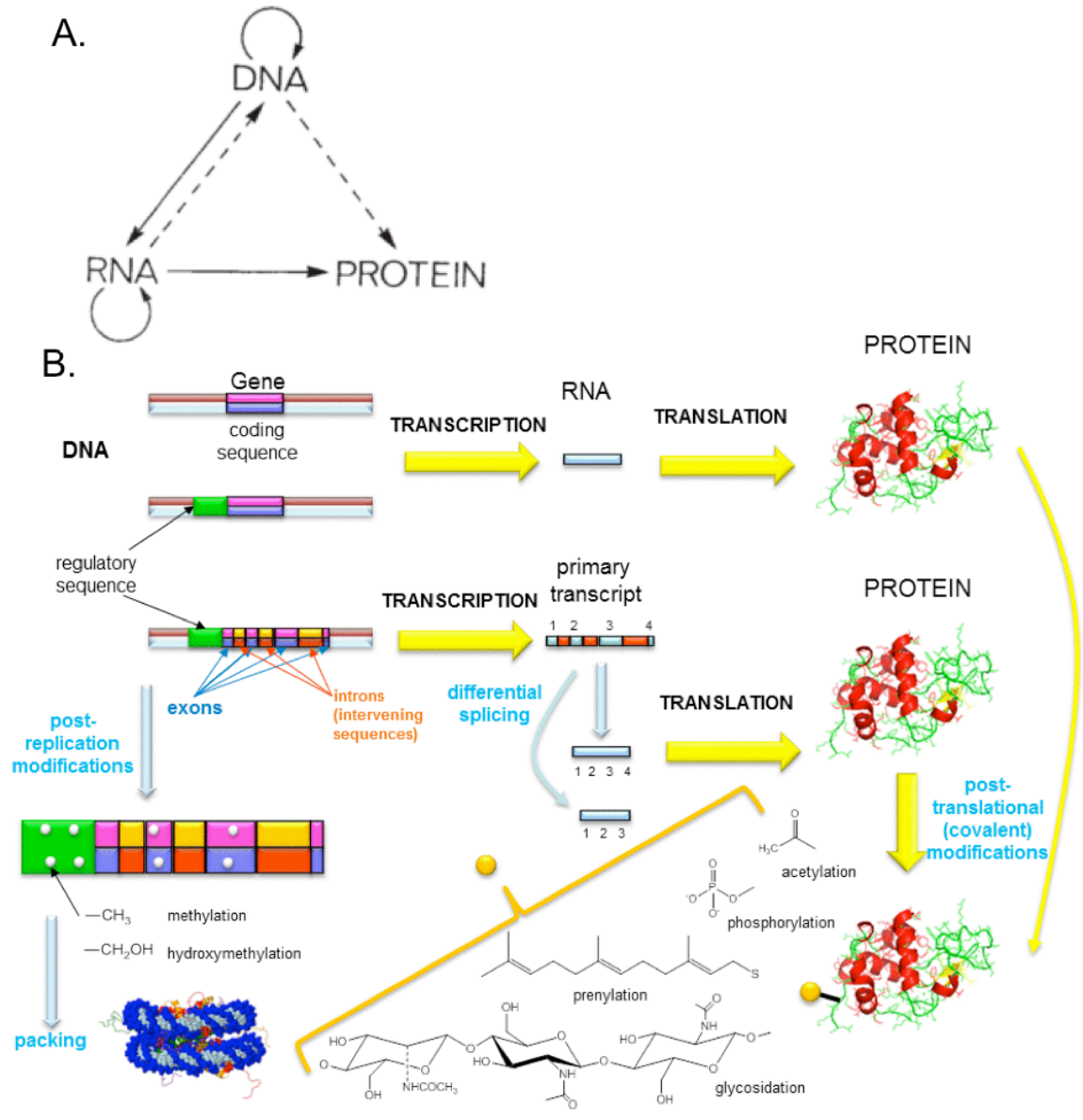


Figure I.1: Complexification du Dogme central de la biologie moléculaire. A. Le dogme central de la biologie moléculaire présenté par Crick (Crick, 1970). Les flèches décrivent les connaissances en 1958, les lignes pleines représentent des transferts probables, les pointillés des transferts possibles. L'absence de flèche représente les transferts impossibles postulés par le dogme central. **B**. Le dogme central (1ère ligne) remis au goût du jour (Figure empruntée au bioWIKI de l'UC Davis).

Je présenterai d'abord un état des lieux des connaissances sur la transcription, sur les séquences régulatrices et les protéines qui en s'y fixant contrôlent leur activité. Je détaillerai ensuite les signatures et règles de grammaire qui peuvent être extraites des séquences *cis*-régulatrices, et qui pourraient un jour, si elles sont mieux définies, servir à prédire l'activité *cis*-régulatrices de toute séquence d'ADN.

A. La transcription eucaryote et son contrôle

1. Le mécanisme de la transcription

La transcription peut être découpée en trois étapes : l'initiation, l'élongation et la terminaison. Chez les eucaryotes, la majorité des gènes incluent des introns et la production d'un ARN messager mature (ARNm) nécessite une étape supplémentaire de maturation : l'épissage des introns (Shefer, Sperling, & Sperling, 2014; Will & Lührmann, 2011). Ces processus sont généralement couplés (Jonkers & Lis, 2015). Je concentrerai ici mon attention sur l'initiation et l'élongation de la transcription. La terminaison ne sera pas abordée, mais de nombreux travaux existent également sur ce sujet (Porrua & Libri, 2015).

L'étape d'initiation de la transcription implique la formation à proximité du site de démarrage de la transcription (TSS) du complexe de pré-initiation (PIC), qui comprend les facteurs de transcription généraux TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH et l'ARN polymérase II (figure I.2) (Orphanides, Lagrange, & Reinberg, 1996). Cet assemblage de protéines dans la région 5' en amont d'un gène donné définit la position du TSS. L'assemblage progressif du PIC implique de nombreuses phosphorylations et conduit au recrutement de l'ARN polymérase II au TSS (Thomas & Chiang, 2006).

Le complexe d'élongation se forme lorsque la transcription débute. Pour cela, l'ARN polymérase II doit d'abord être libérée du PIC *via* de multiples modifications post-traductionnelles au sein des répétitions d'heptapeptides qui constituent le domaine C terminal de sa grande sous-unité (CTD). L'initiation de la transcription requiert en particulier la phosphorylation des sérines 5 par CDK7, une sous unité de TFIIH. L'élongation effective nécessite ensuite la phosphorylation de la sérine 2 de son CTD par CDK9 (positive transcription elongation factor-b, P-TEFb), qui permet la transcription au delà des cinquante

premières bases, puis la déphosphorylation progressive de la sérine 5 par plusieurs complexes de phosphatases (Hsin & Manley, 2012).

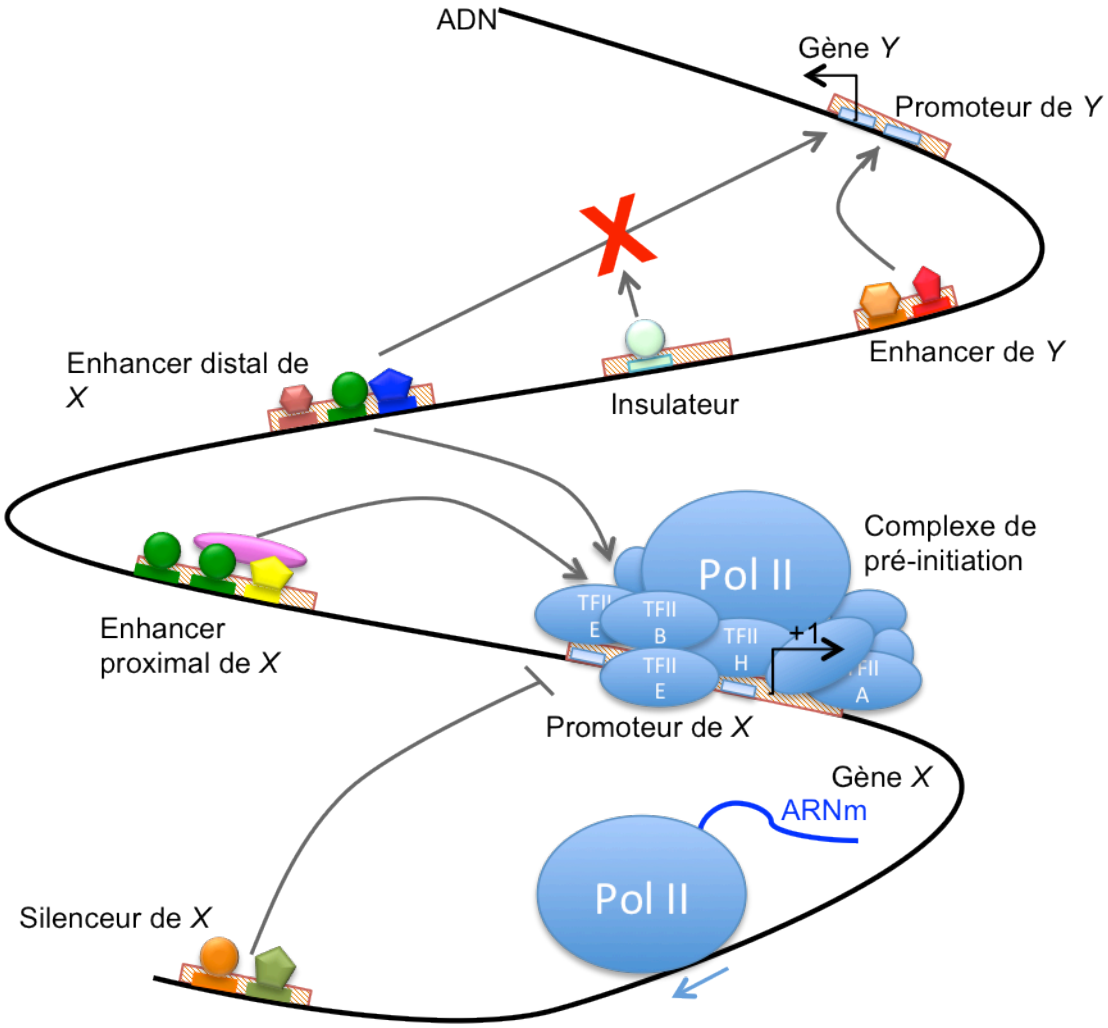


Figure I.2 : Les séquences *cis*-régulatrices en version « cartoon ». Toutes les zones hachurées correspondent à des régions *cis*-régulatrices qui contiennent différents TFBS (rectangles colorés). L'ARN polymérase II (Pol II) et des dizaines de facteurs de transcription dont les facteurs généraux (TFIIx) sont recrutés au niveau du promoteur pour former le complexe de pré-initiation. Les enhancers et silenceurs, lorsqu'ils sont fixés par des facteurs de transcription (formes géométriques colorées), activent et répriment l'initiation de la transcription respectivement. Le gène *Y* est régulé par l'enhancer *Y*, un insulateur empêchant l'interaction entre l'enhancer distal de *X* et le promoteur de *Y*. Il est important de noter que le contact entre enhancers/silenceurs et promoteurs est physique.

2. Les régions *cis*-régulatrices

De nombreux travaux montrent que ces étapes d'initiation et d'élongation de la transcription sont contrôlées par des séquences d'ADN agissant en *cis*, c'est à dire localisées sur la même molécule d'ADN que le gène dont elles contrôlent l'expression. La grande majorité de ces séquences *cis*-régulatrices se situent dans les régions non-codantes du génome, bien que certaines puissent se situer également dans des régions codantes (Barthel & Liu, 2008; Lang, Gombert, & Gould, 2005; Ritter, Dong, Guo, & Chuang, 2012; Tümpel, Cambronero, Sims, Krumlauf, & Wiedemann, 2008). Ces séquences agissent via le recrutement d'une classe spécifique de protéines, les facteurs de transcription qui se fixent sur des séquences spécifiques d'ADN et assurent une communication avec la machinerie de transcription (Jacob & Monod, 1961) (Figure I.2).

Il existe plusieurs types de régions *cis*-régulatrices que je passerai en revue dans les paragraphes suivants : les promoteurs, sièges de la machinerie basale de transcription, les enhancers et silencers, qui régulent la localisation et le niveau de transcription d'un gène, et les insulateurs qui sont des « éléments frontières ».

a) Les promoteurs et l'initiation de la transcription

Le promoteur d'un gène correspond au site où s'assemble le complexe de pré-initiation (PIC) et comprend donc le(s) site(s) d'initiation de la transcription (TSS). Ce type d'élément est généralement très court, de l'ordre de 100 paires de bases. Les premiers promoteurs ont été identifiés dans les années 70 chez les procaryotes (Dhar, Weissman, Zain, Pan, & Lewis, 1974; Pribnow, 1975; Schaller, Gray, & Herrmann, 1975), sur la base de la présence immédiatement en amont du TSS d'une séquence conservée riche en AT (Pribnow box) (pour revue, voir Rosenberg & Court, 1979). De même, la comparaison des séquences en amont du site d'initiation de la transcription de plusieurs gènes eucaryotes a révélé la présence d'un motif similaire à la boîte de Pribnow, la boîte TATA, localisée environ 30 paires de bases en amont du TSS (Gannon et al., 1979), conduisant à l'identification des premiers promoteurs eucaryotes à proximité des gènes d'histone H2A chez la drosophile et chez plusieurs espèces d'oursins, et de l'ovalbumine chez le poulet (M. L. Goldberg, PhD thesis, Stanford University, 1979, Gannon et al., 1979; Kedes, 1979). À la boite TATA, se sont rapidement ajoutées d'autres

signatures et/ou séquences canoniques qui chacune ne concerne qu'une minorité de promoteurs et qui peuvent être spécifiques des espèces considérées (Kadonaga, 2012; Ohler & Wassarman, 2010). Chez les vertébrés, par exemple, 70% des promoteurs annotés se situent au sein d'îlots CpG, des régions riches en dinucléotides CG globalement absentes des génomes d'invertébrés (Deaton & Bird, 2011). Même au sein d'un taxon, de nombreux types de promoteurs peuvent coexister. Ainsi, une analyse plus fine a révélé une grande diversité d'architectures et de logiques dans les promoteurs vertébrés en fonction des classes fonctionnelles des gènes régulés et du moment au cours du développement où ils sont exprimés (Carninci et al., 2006; Forrest et al., 2014; Haberle et al., 2014). Enfin, les promoteurs sont caractérisés par une structure chromatinienne particulière, notamment l'enrichissement de certaines modifications post-traductionnelles de l'histone H3, dont la triméthylation de la lysine 4 (H3K4me3) (Heintzman et al., 2007).

b) Les enhancers et les silencers activent et répriment l'activité des promoteurs

Les enhancers et les silencers contrôlent le profil spatiotemporel d'expression d'un gène en communiquant avec le promoteur du gène pour, respectivement, activer ou réprimer son activité, c'est à dire le recrutement du PIC, ou le début de l'élongation de la transcription. Ces séquences *cis*-régulatrices sont localisées à une distance variable en amont ou en aval du gène régulé, généralement dans une région intergénique ou intronique (E. M. Blackwood, 1998; Ong & Corces, 2011; Spitz & Furlong, 2012).

i) Les enhancers

Le premier enhancer a été découvert dans les séquences du virus SV40 en 1981 (Banerji, Rusconi, & Schaffner, 1981; Benoist & Chambon, 1981; Moreau et al., 1981). Cette classe d'éléments augmente (« enhance ») la transcription d'un (ou plusieurs) gène(s) en communiquant avec le promoteur. Les éléments identifiés font généralement quelques centaines de paires de bases de long et conservent leur activité hors de leur contexte génomique et chromatinien endogène, ce qui permet de tester facilement leur activité dans des constructions synthétiques où ils sont placés en amont d'un promoteur et d'un gène - dit rapporteur - dont l'activité peut facilement être mesurée. Néanmoins, ce type d'expérience teste uniquement la capacité d'une séquence à activer la construction et, dans certains cas, peut être

sujet à des artefacts (Marinić, Aktas, Ruf, & Spitz, 2013; Wai et al., 2001). Arnold et ses collaborateurs rapportent en particulier qu'un tiers des enhancers identifiés par une méthode à grande échelle basée sur le même principe, STARR-seq (Self-Transcribing Active Regulatory Region Sequencing), sont normalement activement réduits au silence dans les cellules où ils ont pourtant été identifiés (Arnold et al., 2013).

La définition initiale des enhancers postulait que leur activité était indépendante de leur orientation, distance ou position (5' ou 3') par rapport au promoteur qu'ils contrôlent. Cette première définition est sans doute trop stricte : il existe, par exemple, des enhancers qui ne fonctionnent que dans une seule direction (Hozumi et al., 2013), et même le niveau d'activité du premier enhancer découvert dépendait de sa distance au promoteur (Moreau et al., 1981). L'expression de la plupart des gènes exprimés de manière restreinte est contrôlée par des enhancers, mais il semble également que de nombreux gènes de ménage soient, au moins chez la drosophile, contrôlés par l'action d'enhancers (Arnold et al., 2013). On estime à plusieurs centaines de milliers le nombre d'enhancers dans un génome de mammifère, soit plus de dix fois le nombre de gènes codants (Bernstein et al., 2012; Shen et al., 2012; Zhu et al., 2013). Seule une petite minorité (~10000) est active dans chacun des types cellulaires ayant été étudiés. Je reviendrai plus loin sur cette surprenante abondance.

Les enhancers fonctionnent comme des plateformes de fixation et d'intégration de l'action des facteurs de transcription (TF), une classe spécifique de protéines qui reconnaissent et se fixent sur des séquences d'ADN plus ou moins spécifiques (E. M. Blackwood, 1998). Chaque enhancer contient plusieurs sites de fixation pour les facteurs de transcription (TFBS), et c'est la combinaison des activités régulatrices des facteurs ainsi recrutés, qui détermine l'activité de l'élément (Arnone & Davidson, 1997). Les facteurs de transcription peuvent agir soit comme activateurs de transcription, soit comme répresseurs, certains pouvant avoir les deux types d'activité en fonction du contexte cellulaire dans lequel ils agissent (Aza-Blanc, Ramírez-Weber, Laget, Schwartz, & Kornberg, 1997; Roose & Clevers, 1999). Si la fixation d'activateurs de la transcription est nécessaire à la fonction d'un enhancer, la définition de ses frontières spatio-temporelles d'activité peut faire intervenir des répresseurs agissant à courte distance. Ainsi l'enhancer "modèle" qui contrôle l'expression d'*evenskipped* dans sa deuxième bande d'expression au cours du développement précoce de la drosophile fixe à la fois les activateurs Hunchback et Bicoid et les répresseurs à court rayon d'action Giant et Krüppel, les seconds restreignant l'activité des premiers (Gray & Levine, 1996).

Une complication supplémentaire provient du fait que l'ADN des cellules eucaryotes n'est généralement pas "nu" : de nombreuses protéines de la chromatine le recouvrent, et l'organisent en structures appelées nucléosomes, dans lesquelles le brin d'ADN est enroulé autour d'un cœur de 8 histones 2 et histones 3 (Campos & Reinberg, 2009). Les nucléosomes et les facteurs de transcription sont donc en compétition pour la fixation sur l'ADN. Les modifications post-traductionnelles des histones ayant un effet important sur l'organisation de la chromatine, certaines modifications post-traductionnelles sur les histones de la chromatine sont enrichies autour des enhancers, en particulier la monométhylation de la lysine 4 de l'histone 3 (H3K4me1) et l'acétylation de la lysine 27 de cette même histone (H3K27ac) (Bulger & Groudine, 2011). Ces modifications sont associées à une structure de chromatine ouverte et accessible aux facteurs de transcription, ce qui a été confirmé par des expériences de sensibilité à la DNAse 1 (Shlyueva, Stampfel, & Stark, 2014).

De nombreux enhancers sont localisés à quelques kilobases de leur gène. Chez les vertébrés et la drosophile, au moins, certains enhancers, notamment ceux des facteurs de transcription et signaux extra-cellulaires contrôlant le développement embryonnaire, peuvent agir à grande distance de leur(s) promoteur(s) cible(s). Par exemple, chez la drosophile, l'enhancer responsable de l'expression du locus *cut* dans les mécanorécepteurs et les soies non innervées, est situé à 80kb en amont de son promoteur (Jack, Dorsett, Delotto, & Liu, 1991). Plus distant encore, l'enhancer responsable de l'expression de *Shh* dans les membres pendant le développement de la souris est situé à 1 Mb de son gène-cible (Lettice et al., 2003). Lorsque le couple enhancer-promoteur est distant, plusieurs gènes et donc plusieurs promoteurs peuvent être localisés entre ces deux éléments. Cet entrelacement entre parties codantes des gènes et séquences régulatrices agissant à grande distance met à mal la proposition initiale que le génome est constitué d'unités fonctionnelles juxtaposées, chacune correspondant à un gène. Elle pose d'autre part la question de la spécificité des interactions entre promoteurs et enhancers dont je parlerai par la suite.

Si d'autres modèles ont initialement été proposés (E M Blackwood & Kadonaga, 1998), le modèle du looping est actuellement le plus soutenu par l'évidence expérimentale pour expliquer l'action à distance des enhancers et les faibles contraintes sur la distance entre enhancers et promoteurs (Kleinjan & van Heyningen, 2005). Il a été construit en cartographiant les contacts entre différentes régions du génome, par des expériences dites de capture de conformation chromatinienne (CCC ou 3C, et ses adaptations à plus grande échelle (4C, HiC))

(van Steensel & Dekker, 2010). Ces expériences ont révélé que la structure tridimensionnelle de la chromatine joue un rôle fondamental dans le contrôle de l'expression génétique et suggèrent que les enhancers établissent des contacts spécifiques avec leur(s) promoteur(s) cible(s), par le biais de la formation de boucles d'ADN (Amano et al., 2009; Bulger & Groudine, 2011; Deng et al., 2012). L'établissement de boucles entre promoteurs et enhancers est requis et suffisant pour l'initiation de la transcription du gène, mais elle ne semble cependant pas suffire à l'étape d'élongation de la transcription (Deng et al., 2012). A noter que la présence de l'enhancer et de son, ou de ses, promoteur(s) cible(s) sur une même molécule d'ADN n'est pas strictement requise : dans de rares cas, un enhancer peut même être localisé sur un autre chromosome, agissant ainsi en tant que séquence *trans*-régulatrice (Gohl, Müller, Pirrotta, Affolter, & Schedl, 2008).

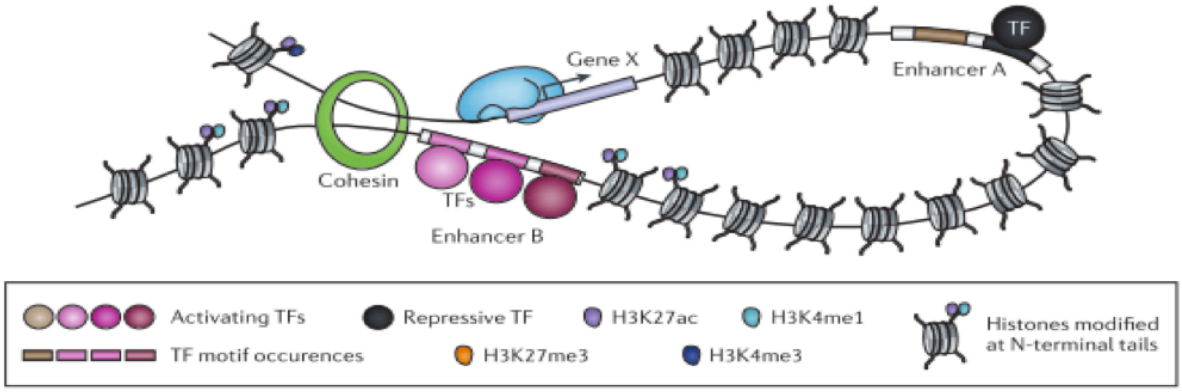


Figure I.3 : Interaction enhancer-promoteur : le modèle du looping. La transcription du gène X est activée par le contact entre son enhancer B et son promoteur, permis par la formation d'une boucle. La cohésine semble impliquée dans la formation de cette boucle. Le contexte chromatinien (marques d'histones associées aux enhancers actifs autour de l'enhancer B), la densité en nucléosomes et/ou l'implication d'autres protéines fixant l'ADN (le facteur de transcription répressif fixé sur l'enhancer A par exemple) permettent le contrôle spatiotemporel de l'établissement de ces boucles (d'après Shlyueva et al., 2014).

Les facteurs de transcription n'interagissent généralement pas eux-mêmes directement avec les éléments de la machinerie basale de transcription situés sur le promoteur. Ils recrutent des complexes co-activateurs, comme Mediator (Conaway & Conaway, 2011), ou corépresseurs, comme groucho/TLE, NurD (Cunliffe, 2008; Jennings & Ish-Horowicz, 2008) ou CtBP (Chinnadurai, 2007). La formation de boucles implique également un complexe protéique inattendu, la cohésine, initialement identifiée pour son rôle dans la cohésion des chromatides

sœurs et qui a la particularité de pouvoir enserrer plusieurs brins d'ADN (Kagey et al., 2010; Schmidt, Schwalie, et al., 2010). Certains de ces complexes, comme Mediator, agissent comme un pont entre les facteurs de transcription et le domaine C terminal de l'ARN-polII (Plaschka et al., 2015). D'autres modifient la chromatine en acétylant les histones, comme p300/CBP (Holmqvist & Mannervik, 2013) ou, au contraire, en les désacétylant, comme le complexe NurD (Cunliffe, 2008). L'ouverture de la chromatine par acétylation de la lysine 27 de l'histone H3 peut faciliter la fixation d'autres TFs (Francetic et al., 2012).

α) Enhanceosomes et tableaux d'affichage : deux modèles extrêmes de fonctionnement

Si l'activité des enhancers résulte de la fixation de facteurs de transcription, plusieurs modes de lecture de la présence de ces facteurs ont été décrits. Arnosti et Kulkarni (Arnosti & Kulkarni, 2005) ont ainsi décrit deux classes d'enhancers : l'enhanceosome et le « tableau d'affichage » (flexible billboard), dont l'organisation et la flexibilité véhiculent différents types d'information.

Dans les enhanceosomes, l'agencement précis des TFBS est primordial et la moindre variation dans l'ordre, l'orientation ou la distance entre sites, annihile activité de l'élément. La fonctionnalité de ces éléments repose sur la coordination et la coopérativité des facteurs fixés. Ce type d'enhancer a été décrit pour la première fois chez les mammifères en étudiant l'activation du gène de l'interféron-β en réponse à une infection virale (Thanos & Maniatis, 1995). Chacun des huit facteurs de transcription se fixant sur cet enhancer est indispensable à son activité : ils se fixent de manière coopérative pour former un complexe protéique qui agit comme une seule unité de régulation (Panne, 2008). Ainsi, les enhanceosomes « *traitent l'information des facteurs de transcription qui s'y fixent pour donner une réponse unitaire* » (Arnosti & Kulkarni, 2005), et répondent de manière hautement non linéaire à la présence de ces facteurs, à la manière d'un interrupteur (Papatsenko & Levine, 2007).

Des études récentes montrent qu'il existe une hiérarchie temporelle dans le recrutement des facteurs de transcription sur un enhanceosome (Chen et al., 2014), mais de manière inattendue, les différents TFs fixés sur l'enhanceosome de l'interféron-β établissent peu de contacts entre eux malgré leur proximité. Leur coopération implique donc probablement la déformation de

l'hélice d'ADN par les premiers facteurs qui faciliteraient la fixation des suivants et/ou le recrutement coopératif de coactivateurs comme p300/CBP (Panne, 2008).

Plusieurs enhancers développementaux semblent satisfaire à la définition d'un enhanceosome (Luster & Rizzino, 2003; Ruan et al., 2009). La présence d'éléments ultraconservés non codants chez les vertébrés, dont la séquence a été conservée à plus de 70% sur plusieurs centaines de millions d'années, pourrait aussi faire penser à un mode d'action de type enhanceosome (Harmston, Baresic, Lenhard, & B, 2013). Cette classe d'enhancer est néanmoins probablement très minoritaire (Spitz & Furlong, 2012).

Au contraire des enhanceosomes, l'organisation des sites de fixation des facteurs de transcription dans les enhancers de type « tableau d'affichage » est très flexible. Ainsi, Hare et ses collègues ont montré que les enhancers d'*even skipped* contiennent des blocs conservés de 20 à 30 paires de bases seulement entre différentes espèces de mouches. Ces blocs sont enrichis en TFBS homologues dont l'organisation n'est pas nécessairement conservée (Figure I.4 D) (Hare, Peterson, Iyer, Meier, & Eisen, 2008). L'existence d'éléments régulateurs ayant une architecture conservée n'est pas l'apanage des organismes multicellulaires : une étude chez la levure a montré que le changement d'orientation d'un TFBS dans un promoteur n'affecte son activité que dans 8% des cas (Sharon et al., 2012). La majorité des enhancers identifiés, et notamment les enhancers développementaux, semblent être architecturalement peu contraints (Borok, Tran, Ho, & Drewell, 2010).

Ces deux classes d'éléments correspondent probablement aux deux extrêmes d'un continuum de types d'enhancers. Dans de nombreux enhancers développementaux, les facteurs de transcription ne se lient pas tous de façon coopérative, mais l'organisation de leurs sites de fixation est en partie contrainte, par exemple pour faciliter leurs interactions (Spitz & Furlong, 2012). Ainsi, Makeev et ses collègues (Makeev, Lifanov, Nazina, & Papatsenko, 2003) ont par exemple montré que l'arrangement des sites de fixation de certains TFs dans les enhancers de drosophile était contraint, avec une périodicité de l'ordre d'un tour d'hélice qui placerait les facteurs en regard l'un de l'autre. Erceg et ses collègues ont eux montré que, dans des embryons de drosophile, l'organisation de deux types de TFBS au sein d'un enhancer synthétique est plus ou moins contrainte en fonction du tissu considéré. Ainsi, changer l'espacement entre des sites de fixation pour les facteurs Tin et pMad affecte plus l'activité de l'enhancer dans les cellules cardiaques que dans le mésoderme viscéral (Erceg et al., 2014).

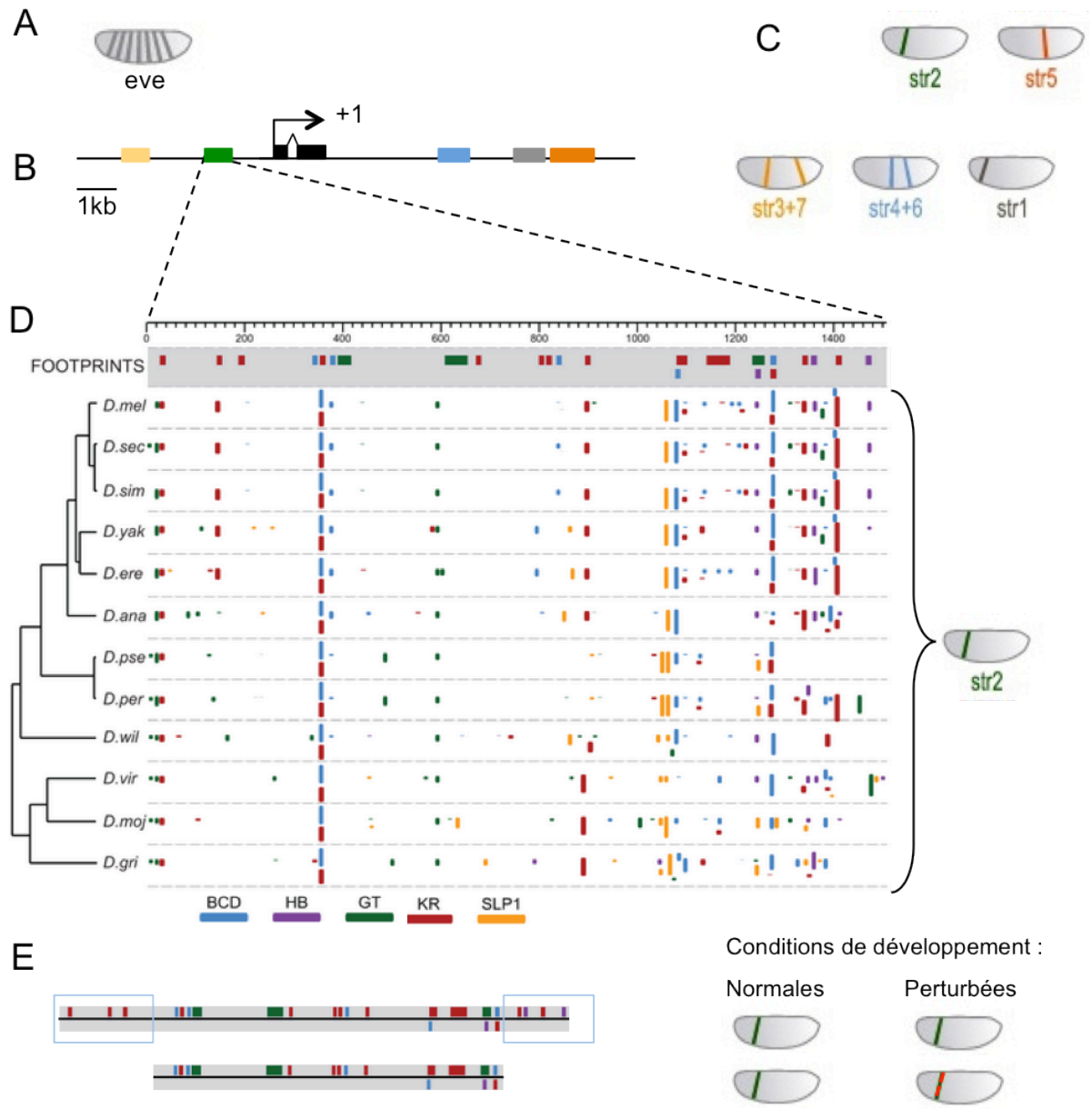


Figure I.4 : Les enhancers d'*even-skipped* (*eve*): modularité, conservation des motifs et redondance fonctionnelle des TFBS. A. Profil d'expression d'*eve* au stade blastoderme. **B.** Structure du locus d'*eve*. Chaque rectangle représente un enhancer d'*eve*. Leur couleur correspond au profil d'expression qu'ils génèrent. **C.** Le profil d'expression d'*eve* est la somme des activités de ses enhancers : chaque bande ou paire de bandes est générée par un enhancer différent (str1 à 7). **D.** Conservation des TFBS de l'enhancer str2 d'*eve* dont le profil d'activité est conservé chez les drosophiles. La séquence de cet enhancer n'est globalement pas conservée entre les différentes espèces de drosophiles, mais elles contiennent les mêmes TFBS, dont l'organisation est plus ou moins ré-arrangées en fonction de leur distance évolutive, soulignant la flexibilité de cet enhancer de type « tableau d'affichage ». Les rectangles représentent les sites de fixation pour cinq facteurs de transcription régulant cet enhancer. Leur couleur correspond à l'identité du facteur de transcription

reconnaissant ce motif (voir légende en bas), et leur hauteur représente leur affinité prédite. Le rectangle gris contient tous les sites de fixation effectivement fixés *in vitro*. E. En conditions de développement normales, l'enhancer str2 complet ou « minimal » (schématisés à gauche, même code couleur que D pour les TFBS) génère un profil d'expression correct pour la seconde bande (schématisé à droite). Cependant, la version tronquée ne parvient pas à établir correctement ce même profil. Les TFBS contenus dans les cadres bleus ne sont pas nécessaires à l'activité de l'enhancer mais participent à sa robustesse. (Adapté de Hare et al., 2008; Ludwig, Manu, Kittler, White, & Kreitman, 2011; Wilczyński & Furlong, 2010).

De nombreux enhancers pourraient donc avoir un niveau de contrainte architecturale intermédiaire entre les enhanceosomes et les « tableaux d'affichage », et ces contraintes pourraient dépendre du tissu considéré. Papatsenko et Lévine (Papatsenko & Levine, 2007) ont proposé que les enhancers les plus contraints répondent à des concentrations de facteurs plus faibles, pour lesquels la coopération entre facteurs est critique pour l'amplification du signal. Les enhancers les moins contraints répondraient eux à des facteurs suffisamment exprimés pour se fixer sans devoir coopérer. Notre compréhension des déterminants de ces contraintes reste actuellement très fragmentaire.

ß) Intégration de l'activité d'enhancers multiples : une diversité de paysages régulatoires

Les enhancers agissant à distance le font rarement de manière isolée. Ainsi, l'analyse de nombreuses lignées, dites "enhancer traps", qui insèrent des couples promoteur-gène rapporteur dans les génomes pour détecter la présence d'enhancers, a révélé que de nombreux enhancers avec des profils d'activité partiellement redondants sont localisés autour des gènes de développement (Kikuta et al., 2007). Ces gènes sont donc contrôlés par de véritables "paysages régulatoires" (regulatory landscapes) identifiés dans le contexte de l'embryon. L'intégration des différents éléments de ces paysages a été étudiée dans un petit nombre de cas.

Denis Duboule et ses collègues ont ainsi identifié un "archipel régulatoire" (regulatory archipelago) dans le désert génique localisé à proximité du complexe HoxD, du côté centromérique (Montavon et al., 2011; Spitz, Gonzalez, & Duboule, 2003). Cette région d'environ 600kb contrôle la transcription du complexe *HoxD* dans les doigts en développement. Elle contient un nombre important d'ilots de régulation dont les activités, partiellement

redondantes, sont coordonnées par l'établissement de nombreuses boucles entre ilots et avec le promoteur (Montavon et al., 2011). Une situation analogue a été décrite pour le contrôle de la transcription du gène FGF8 dans les membres (Marinić et al., 2013). Là encore de nombreux éléments sont présents, avec des activités partiellement redondantes mais aussi un entrelacement entre enhancers actifs dans différents tissus. Schwarzer et Spitz (Schwarzer & Spitz, 2014) proposent que le contrôle des contacts chromatiniens établis entre ces éléments permet d'intégrer leur activité, par exemple en multipliant le nombre d'interactions productives entre enhancers et promoteurs (Perry, Boettiger, & Levine, 2011), ou par les actions successives de ces éléments, les premiers ouvrant la chromatine pour préparer l'action des suivants (Andrey et al., 2013).

Une nouvelle classe d'éléments régulateurs, les super enhancers, pourraient constituer un cas extrême de "paysage régulatoires". Alors que la majorité des enhancers occupent un segment de chromosome de quelques centaines de bases à quelques kilobases, et que les archipels peuvent couvrir des distances de près d'un mégabase, plusieurs études ont identifié dans des cellules souches embryonnaires murines un petit nombre (~200) de régions génomiques de taille intermédiaire, allant jusqu'à 50kpb, contenant des regroupements de sites de fixations pour des facteurs de transcription et un enrichissement inhabituellement élevé pour le co-activateur de la transcription Mediator et pour les marques chromatiniennes associées aux enhancers actifs (H3K4me1, H3K27Ac) (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). Les auteurs ont appelé ces régions des « super-enhancers » et avancent que de telles séquences, dont l'activité serait spécifique d'un type cellulaire, joueraient un rôle clef le contrôle des réseaux géniques de maintien de l'identité cellulaire. Ces séquences sont généralement localisées à proximité de gènes de spécification, comme *Oct4* ou *Sox2* dans les cellules souches embryonnaires de souris ou d'oncogènes. Cependant, ces super-enhancers ne sont pas définis au niveau fonctionnel. Il serait notamment intéressant de savoir s'ils peuvent être découpés en sous-unités fonctionnelles comme les archipels et d'identifier les réseaux de contacts chromatiniens établis au sein de ces régions et avec leur promoteur cible. Les super-enhancers pourraient simplement correspondre à des archipels particulièrement compacts, pouvant agir comme une même unité de régulation, ce qui expliquerait l'importance de l'enrichissement en Mediator (Pott & Lieb, 2015). Si les raisons de cette super organisation des enhancers restent mystérieuses, elle pourrait faciliter, de manière pratique, l'identification d'enhancers impliqués dans le maintien de l'identité cellulaire.

iv) Les silencers

Les silencers ont initialement été identifiés chez la levure comme des éléments ayant des propriétés opposées à celles des enhancers, et pouvant ainsi réprimer l'expression d'un gène à distance et indépendamment de leur orientation (Brand, Breeden, Abraham, Sternglanz, & Nasmyth, 1985). Ils ont également été détectés dans les génomes de métazoaires où ils ont une action modulaire sur l'activité des enhancers et/ou promoteur du gène qu'ils contrôlent (Yuh & Davidson, 1996). Ils ont également une action de répression à longue distance (Gray & Levine, 1996). Pourtant, ces séquences ont globalement reçu moins d'attention et leur nombre au sein d'un génome animal reste inconnu, peut-être parce que leur identification est moins simple expérimentalement.

L'activité des silencers résulte de la fixation de répresseurs agissant à longue distance (Courey & Jia, 2001), via le recrutement de corépresseurs, dont Groucho/TLE ou Sir2. Ces protéines sont à l'origine de la formation de complexes protéiques incluant des modificateurs de la chromatine, en particulier des déacétylases. L'analyse de l'action du répresseur à grand rayon d'action, Hairy, sur le gène *Fuji-tarazu* (*Ftz*) chez la drosophile, indique que ce répresseur est capable de déacétyler la chromatine sur de longues distances (L. M. Li & Arnosti, 2011), peut-être par un mécanisme de "propagation" dirigé par Groucho (Martinez & Arnosti, 2008).

c) Vers une vision unifiée des enhancers et des promoteurs?

Les définitions initiales des promoteurs et des enhancers ont conduit à l'idée que ces éléments étaient de natures fondamentalement différentes. Leurs différences conceptuelles s'estompent actuellement (Andersson, Sandelin, & Danko, 2015).

Premièrement, au promoteur central (core promoter), sur lequel s'assemble la machinerie basale de transcription, s'est ajouté le promoteur proximal (proximal promoter) situé en amont du promoteur central, généralement à moins de 200pb du TSS et qui partage avec les enhancers la capacité à recruter des facteurs de transcription (Ohler & Wassarman, 2010). Ainsi promoteurs et enhancers semblent pouvoir fixer des protéines similaires. Le promoteur proximal pourrait jouer un rôle dans la spécificité de l'interaction entre promoteurs et enhancers (Gehrig et al., 2009; Zabidi et al., 2014) ou dans la capacité du promoteur à répondre à des

enhancers distants (Wen, Crawford, & Locker, 1993). Ainsi, d'un élément à l'activité générique essentiellement positionnelle, on est passé à un concept de promoteur pouvant jouer un rôle dans le contrôle de l'activité du gène et dans la spécificité d'action des séquences *cis*-régulatrices agissant à plus grande distance. Les promoteurs peuvent même jouer un rôle d'enhancer (Leung et al., 2015) et porter des marques d'histones spécifiques de ces éléments.

Inversement, certains enhancers sont capables d'agir comme promoteurs (Banerjee, Kim, & Kim, 2014). Il existe en particulier une classe d'enhancers transcrits, qui ont la capacité de recruter l'ARN polymérase II pour produire de courts ARNe (Lai & Shiekhattar, 2014) qui participent à la fonction de l'enhancer.

d) Limitation du champs de régulation des enhancers

Si certains enhancers peuvent agir à très grande distance, leur action est restreinte au contrôle de gènes ou ensembles de gènes spécifiques. Ils ne régulent pas nécessairement les gènes les plus proches, et peuvent, comme nous l'avons vu plus haut, être entrelacés avec des gènes dont ils ne contrôlent pas la transcription. Cette spécificité d'action est contrôlée au moins à deux niveaux, celui de la séquences nucléotidiques des promoteurs et enhancers, et celui du repliement tridimensionnel des chromosomes dans le noyau (van Arensbergen, van Steensel, & Bussemaker, 2014). Il est intéressant de noter, néanmoins, que la spécificité ne doit pas toujours être parfaitement contrôlée : ainsi les gènes *Luna Park* et *evx* sont sous le contrôle de l'archipel de régulation de l'expression dans les doigts du complexe *Hoxd* (Spitz et al., 2003) sans pour autant être impliqués dans ce processus. Il est probable que l'expression de ces gènes dans ce tissu est simplement neutre, et que la cellule ne subit donc que peu de pression pour empêcher cette expression.

i) Impact de la séquences des enhancers et promoteurs

Un exemple classique de l'interaction spécifique entre enhancer et promoteur, aussi appelée « compatibilité biochimique », est celui, chez la drosophile, de *gsb* et *gsbn (gooseberry et gooseberry neuro),* deux gènes voisins divergents et exprimés de façons différentes : leurs enhancers respectifs sont localisés dans la région de 10kb qui sépare les TSS des deux gènes et sont spécifiques de l'un ou l'autre gène (X Li & Noll, 1994). Chez les mammifères, au sein du

groupe de gènes ß-globine, la région de contrôle *cis*-régulatrice (Locus Control Region LCR) n'agit qu'avec un seul gène ß-globine à la fois en fonction du stade de développement de l'embryon (Tolhuis, Palstra, Splinter, Grosveld, & de Laat, 2002).

Une étude récente chez la drosophile a confirmé que la spécificité enhancer/promoteur est un phénomène général qui dépend de leur séquence (Zabidi et al., 2014). Cette étude suggère notamment que les promoteurs des gènes de ménage et des gènes impliqués dans les programmes développementaux ont des spécificités d'interaction avec des enhancers distinctes. Cette étude confirme des études antérieures, indiquant que certains enhancers interagissent préférentiellement avec des promoteurs centraux contenant certains motifs, comme les boites TATA ou les DPEs (Butler & Kadonaga, 2001; Ohtsuki, Levine, & Cai, 1998). Certaines séquences du promoteur proximal pourraient également stabiliser les interactions avec des enhancers spécifiques ou localisés à grande distance (Calhoun, Stathopoulos, & Levine, 2002).

ii) Rôle de l'architecture 3D des chromosomes

Nous avons vu plus haut que les enhancers peuvent interagir avec leurs promoteurs via la formation de boucles d'ADN. Le développement des technologies permettant de capturer la conformation 3D de la chromatine à grande échelle (5C, Hi-C) a permis de cartographier les interactions à longue distance entre éléments régulateurs, et gènes. Ce type d'approche a révélé l'existence de plusieurs niveaux hiérarchiques d'organisation de la chromatine (Phillips-Cremins, 2014) (Figure I.5).

A l'échelle des chromosomes, les premières expériences de Hi-C chez l'homme ont révélé l'existence de grands territoires chromosomiques de plusieurs dizaines de mégabases établissant des contacts de manière préférentielles et des contacts intra-chromosomiques entre petits chromosomes riches en gènes (Lieberman-Aiden et al., 2009). De plus, les régions de chromatine ouvertes ou fermées ont tendance à ségréger dans des compartiments distincts d'une taille médiane de 3Mb.

Avec l'augmentation de la résolution des techniques de caractérisation des contacts chromatiniens, des domaines plus petits (~1Mb), les domaines d'association topologiques (Topology associating Domains, ou TADs) ont été identifiés. Les TADs couvrent plus de 90%

du génome, et peuvent chevaucher deux compartiments. Ils sont caractérisés par un enrichissement en contacts intra-domaines et peu de contacts avec d'autres TADs (Dixon et al., 2012; Nora et al., 2012). Logiquement, les boucles enhancers-promoteurs sont fréquentes au sein des TADs (van Arensbergen et al., 2014).

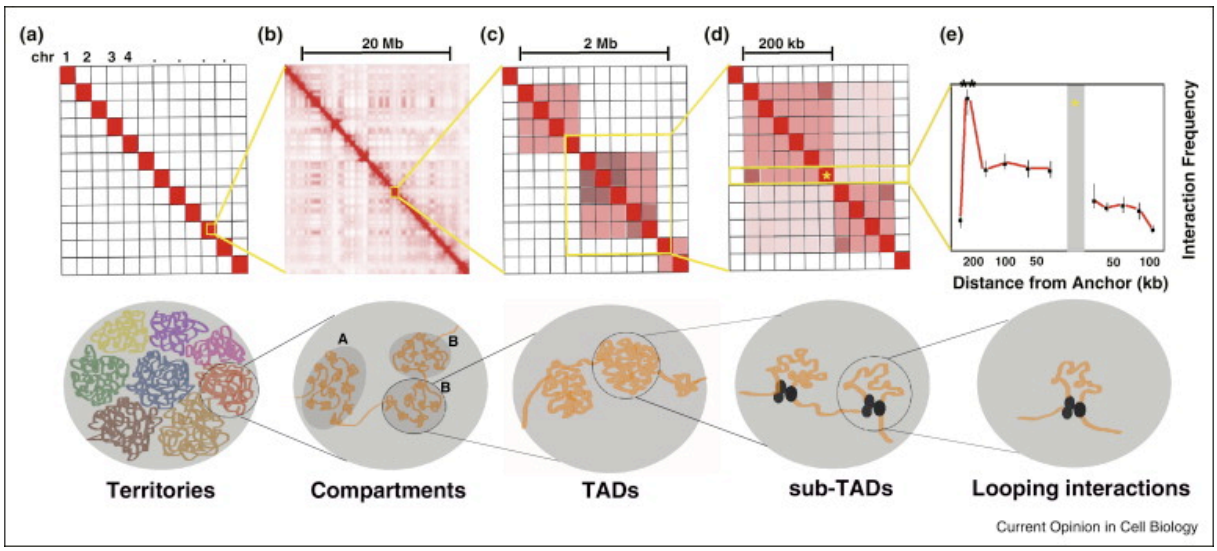


Figure I.5 : Les différents niveaux d'organisation de la chromatine. (a) Les territoires chromosomiques. **(b)** Les compartiments 'A' et 'B' au sein d'un territoire chromosomique. **(c)** Groupes de TADs au sein des compartiments A et B. **(d)** Sous-TADs au sein de TADs. **(e)** interactions intra-chromosomales pas looping au sein et entre sous-TADs et TADs. Les diagrammes (Heatmaps) illustrent les données d'interaction de la chromatine, l'intensité des signaux correspondant à l'intensité du rouge dans chaque case (Phillips-Cremins, 2014).

Cette organisation hiérarchique correspond à des fonctions distinctes.

Les compartiments de chromatine ouverts pourraient rapprocher des gènes co-exprimés dans ce qui a été baptisé des usines de transcription (transcription factories), localisées dans des régions précises du noyau, et marquées par les formes multiphosphorylées de l'ARN polymérase II (Gibcus & Dekker, 2013). Inversement, les compartiments réprimés sont associés à la lamine périnucléaire (Reddy, Zullo, Bertolino, & Singh, 2008). Cette localisation nucléaire apparait dynamique (Peric-Hupkes et al., 2010). Ainsi, au cours de la différenciation des cellules B, un domaine entier est réprimé et associé à la lamine dans les cellules pré-pro B, puis activé et relocalisé à l'intérieur du noyau sans les cellules pro B (Lin et al., 2012).

Les TADs semblent invariants entre types cellulaires. Ils pourraient jouer un rôle dans la délimitation des paysages régulatoires mentionnés plus haut et contraindraient ainsi les régions du génome que chaque enhancer ou promoteur a besoin de scanner pour trouver son ou ses

partenaires. De fait, les gènes au sein d'un TAD ont tendance à être plus co-exprimés qu'entre TADs (Nora et al., 2012). Certains clusters de gènes exprimés de manière concertée et dynamique, comme les gènes du complexe HoxD (Andrey et al., 2013; Noordermeer et al., 2014; Vieux-Rochas, Fabre, Leleu, Duboule, & Noordermeer, 2015), ou les gènes Six (Gómez-Marín et al., 2015), sont situés à la charnière entre deux TADs voisins. Dans le cas des gènes *HoxD*, la transition d'expression entre les gènes exprimés précocement, responsables des structures proximales du membre, et les gènes exprimés plus tardivement dans les doigts est assurée par la transition des marques activatrices (H3K27Ac) d'un TAD à l'autre. Au sein des TADs, on trouve à la fois des boucles d'ADN stables, ainsi que d'autres qui semblent être spécifiques d'un processus développemental ou de différentiation. Certaines des boucles permettant l'interaction entre promoteurs et enhancers existent avant l'activation des gènes concernés, facilitant peut-être l'activation rapide des gènes en réponse à un stimulus extracellulaire (DeMare et al., 2013).

iii) Les insulateurs, éléments “frontières”

Comment contrôler la position des frontières entre domaines de régulation, et donc l'interaction entre enhancers et promoteurs ? Plusieurs concepts, issus d'expériences assez anciennes, notamment celui d'insulateur, coexistent avec des concepts plus récents issus d’analyses de conformation 3D de la chromatine, comme celui de TAD. Les relations précises entre ces différents concepts restent à préciser. Les insulateurs ont d’abord été identifiés comme des éléments empêchant l’activation d’un promoteur par un enhancer lorsqu’il est situé entre les deux (Kellum & Schedl, 1991). Les insulateurs peuvent protéger un gène de l’effet de répression de la transcription produit par la chromatine condensée inactive avoisinante, mais leur mode d’action principal semble être l’établissement de contacts entre deux insulateurs, qui permettrait la formation de boucles, incompatibles avec les interactions promoteurs-enhancers non désirées. La fonction des insulateurs ferait donc également appel au concept de boucle d'ADN.

Une idée émergente est que des combinaisons distinctes de protéines architecturales agissent pour établir les frontières des TADs et des boucles intra TADs (Phillips-Cremins et al., 2013). Ces protéines incluent en particulier le facteur CTCF, la Cohésine, et le complexe co-activateur Mediator. Ainsi la combinaison de marques CTCF et Cohésine marquerait les frontières entre

TAD, alors qu'à l'échelle sub-TAD (~100kb), les boucles seraient stabilisées par les interactions Cohésine/Mediator. Certaines de ces interactions semblent fonctionnellement importantes. Ainsi, la frontière entre les deux TADs situés de part et d'autre du complexe des gènes *Six* est évolutivement conservée à l'échelle des deutérostomes, et inclue des sites tête-bêche pour le facteur CTCF. La délétion de la région frontière entre ces deux TADs affecte la spécificité des interactions entre enhancers et promoteurs (Gómez-Marín et al., 2015). Enfin, notons que CTCF, l'unique facteur d'isolation identifié chez les vertébrés est absent des génomes de nématodes, de levures et de plantes (Ong & Corces, 2011). Chez la drosophile quatre facteurs d'isolation supplémentaires ont été identifiés (Schoborg & Labrador, 2010).

Le modèle précédent a d'ailleurs été remis en question récemment chez la drosophile, la présence de sites pour des insulateurs comme CTCF ou Su(W) étant un moins bon indicateur de la frontière entre TADs que la présence de marques de chromatine active (Ulianov et al., 2015).

Globalement, la mise au point de méthodes à très haut débit, que ce soit pour identifier des enhancers, comme STARR-seq, des promoteurs, comme le ChIP-seq pour H3K4me3, ou pour caractériser la structure 3D de la chromatine a révolutionné notre compréhension des interactions entre promoteurs et enhancers. Les modèles proposés au cours des dernières années sont souvent séduisants, mais il ne faut pas oublier que nous n'avons encore que très peu de recul sur ces données et ne pas faire de généralisations trop ambitieuses. Il serait notamment intéressant de savoir quel est l'impact de la taille du génome sur les mécanismes mis en œuvre. Les leçons collectées en analysant les grands génomes de mammifères s'appliquent-elles également à des organismes ayant de plus petits génomes, comme la drosophile, le nématode, ou les ascidies ?

e) Les séquences *cis*-régulatrices et à la robustesse du programme développemental

Nous avons vu précédemment que l'expression de nombreux gènes régulateurs est contrôlée par des paysages régulatoires complexes. D'où vient cette complexité, quelle est son origine ?

L'analyse de l'activité de nombreux enhancers a révélé que lorsqu'un gène est exprimé de façon complexe dans plusieurs territoires ou à différents moments, l'activité de chacune de ses

séquences régulatrices est généralement restreinte à un petit nombre de territoires et limitée dans le temps (Arnone & Davidson, 1997). La logique d'action des séquences régulatrices est donc de nature modulaire. La région régulatrice du gène *even-skipped* (*eve*) de la drosophile constitue un exemple frappant de cette modularité. Le profil d'expression d'*eve* consiste en sept bandes perpendiculaires à l'axe antéro-postérieur de l'embryon au stade blastoderme (quelques heures après fécondation). La plupart des bandes sont générées par des modules *cis*-régulateurs indépendants répondants à des combinaisons de facteurs de transcription distincts (Figure I.4 A-C) (Fujioka, Emi-Sarker, Yusibova, Goto, & Jaynes, 1999; Sackerson, Fujioka, & Goto, 1999), ce qui a conduit à proposer que la nature faisait des bandes de manière "inélégante" par opposition à des mécanismes basés, par exemple, sur des modèles de réaction/diffusion, comme les modèles de Turing (Akam, 1989).

Cette modularité n'explique pas à elle seule la multiplicité des séquences *cis*-régulatrices contrôlant l'activité des gènes régulateurs du développement. L'analyse des séquences régulatrices au sein des paysages régulatoires a ainsi révélé la présence fréquente de plusieurs enhancers, d'activité similaire, et agissant de manière additive ou partiellement redondante (pour revue voir Barolo, 2012). Cette redondance fonctionnelle partielle des enhancers pourrait augmenter la robustesse du programme aux variations génétiques présentes dans une population naturelle, ou à des conditions environnementales extrêmes (pour revue, Lagha, Bothma, & Levine, 2012).

Ainsi, chez la drosophile, l'analyse des profils de fixation des facteurs de transcription développementaux par immunoprécipitation de la chromatine a révélé l'existence de régions assez éloignées du gène et fixant les même facteurs de transcription que les enhancers plus proches déjà identifiés. Ces régions ont effectivement des activités très similaires aux enhancers proximaux et ont été qualifiées de « shadow enhancers », car elles semblent vivre dans l'ombre de ces premiers enhancers. En conditions standard de développement, la délétion de l'un ou l'autre de ces enhancers "redondants" n'affecte pas le profil d'expression du gène cible. Par contre, en cas de perturbation environnementale ou génétique, la stabilité du profil d'expression requiert la présence simultanée des deux enhancers. Ainsi, chez la mouche, le locus *shavenbaby* est régulé par plusieurs enhancers redondants. Ce locus est important pour la spécification des trichomes sur la cuticule dorsale de l'embryon et de la larve. À température optimale pour le développement (25°C), la délétion individuelle de ces enhancers affecte peu la disposition des trichomes. Mais de nombreux trichomes sont perdus quand les embryons

portant ces délétions se développent à des températures extrêmes, élevées (32°C) ou basses (17°C), ou avec un niveau réduit de signalisation Wingless requise pour le développement des trichomes. Les embryons contenant tous les enhancers sont robustes à ces changements (Frankel et al., 2010). Des conclusions similaires ont été obtenues dans le cas des enhancers du gène *Snail*, chez la drosophile également (Perry, Boettiger, Bothma, & Levine, 2010).

La redondance informationnelle se retrouve aussi à l'échelle intra-enhancer. Les enhancers développementaux mesurent en général quelques centaines de paires de bases et contiennent souvent plusieurs sites de fixations sur l'ADN pour de multiples facteurs de transcription. S'il est possible de définir fonctionnellement un enhancer minimal (c'est à dire le plus petit fragment capable de récapituler qualitativement l'activité de l'enhancer dans des expériences de transgénèse), les régions avoisinantes contiennent souvent des copies additionnelles de ces TFBS, et la définition des extrémités d'un enhancer est actuellement assez arbitraire. Certains de ces TFBS sont fonctionnellement redondants et leur délétion ne semble pas affecter l'activité de l'enhancer dans des conditions optimales de développement. Encore une fois, l'exemple de l'enhancer de la deuxième bande d'*even skipped* est un cas d'école. L'enhancer complet mesure environ 800 paires de bases, dont plus de 300 peuvent être délétées sans conséquence majeure sur l'activité qualitative de l'enhancer, bien qu'elles contiennent plusieurs TFBS fonctionnels. *In vivo,* néanmoins, l'enhancer minimal ne permet un développement viable qu'en conditions optimales et les séquences avoisinantes sont nécessaires pour la robustesse de l'expression du gène (I.4 E) (Ludwig et al., 2011).

f) Importance des séquences régulatrices au cours de l'évolution

L'idée que le contrôle de l'expression des gènes puisse être soumis à la sélection naturelle est ancienne (pour revue, voir Wray, 2007). Déjà Jacob et Monod (Monod & Jacob, 1961) suggéraient que des mutations au sein des opérateurs pourraient jouer un rôle dans l'évolution. La découverte de King et Wilson montrant que les protéines humaines et de chimpanzé sont pratiquement identiques malgré la différence phénotypique entre ces deux espèces, a renforcé l'idée que la différence d'expression des gènes pouvait jouer un rôle évolutif majeur (King & Wilson, 1975).

L'importance relative pour l'évolution des espèces des mutations dans les séquences codantes et dans les séquences régulatrices non codantes fait l'objet d'un vif débat. Certains auteurs comme Carroll prônent un rôle majoritaire des séquences régulatrices dans l'évolution des morphologies embryonnaires (Carroll, 2008). D'autres insistent sur la contribution majoritaire des séquences codantes, que ce soit *via* des mutations changeant la fonction des protéines ou *via* d'autres mécanismes comme la duplication génique et la sous-fonctionnalisation des deux paralogues (Hoekstra & Coyne, 2007). Une tentative de synthèse entre ces deux visions suggère que l'évolution utilise les mutations codantes ou non codantes dans différentes proportions dans différents groupes, à différentes échelles de temps et à différents niveaux dans la hiérarchie de contrôle développemental (D. L. Stern & Orgogozo, 2008). Notons de plus que les séquences protéiques des facteurs de transcription peuvent eux aussi évoluer de manière adaptative, une autre manière de produire des changements dans les réseaux de *cis*-régulation (Lynch et al., 2008; Sayou et al., 2014). J'illustrerai plus loin la contribution des séquences régulatrices aux processus évolutifs, sans chercher à la comparer à celle des séquences codantes (mais voir par exemple Wittkopp, Haerum, & Clark, 2004).

Deux raisons sont souvent invoquées pour expliquer la sélection de mutations dans les séquences *cis*-régulatrices : la codominance de leurs effets et la modularité de leur action. La transcription des deux allèles d'un même gène est généralement indépendante et additive. Une mutation dans une région *cis*-régulatrice aura donc généralement un effet à l'état hétérozygote, alors que la majorité des mutations codantes sont elles récessives. Cette propriété rend les mutations *cis*-régulatrices plus visibles par le processus de sélection naturelle et facilite donc leur sélection. La modularité de l'action des séquences régulatrices permet, elle, de n'affecter la fonction d'un gène que dans une partie de son domaine d'expression et de limiter ainsi la pléiotropie de ce type de mutations. De fait, de nombreuses mutations identifiées grâce à des cribles pour des phénotypes développementaux ou des QTLs (Quantitative Trait Loci), identifiés en lien avec des traits évolutifs, affectent des séquences *cis*-régulatrices (Maurano et al., 2012). Ces mutations jouent un rôle dans la divergence entre espèces. Elles peuvent aussi jouer un rôle dans la convergence évolutive. La coévolution du mimétisme chez deux espèces de papillons distantes de plus de 65 millions d’années est ainsi due à un polymorphisme nucléotidique dans des régions homologues non-codantes situées près du locus de *WntA* (Gallant et al., 2014). Enfin, si la discussion ci-dessous se concentre sur l'évolution des séquences régulatrices, les facteurs de transcription peuvent eux aussi évoluer de manière adaptative (Lynch et al., 2008; Sayou et al., 2014).

Les mutations sélectionnées par l'évolution n'affectent parfois qu'un seul nucléotide. Ainsi, la mutation d'un site de fixation pour le facteur de transcription Dsx chez *Drosophila takahashi*, est suffisante pour rendre monomorphique un trait sexuellement dimorphique (Shirangi, Dufour, Williams, & Carroll, 2009). Certains types de sites de fixation pour des facteurs de transcription, pour les protéines Hox notamment, semblent d'ailleurs être soumis à des contraintes aussi fortes que les séquences codantes (Vernot et al., 2012). De même, un petit nombre de mutations *cis*-régulatrices dans le gène *ebony*, qui contrôle la pigmentation de l'abdomen, ont été impliquées dans l'adaptation de populations de *Drosophila melanogaster* à l'altitude (Rebeiz, Pool, Kassner, Aquadro, & Carroll, 2009).

Dans d'autres cas, les mutations sélectionnées consistent dans la délétion d'un enhancer complet. Ainsi, lors de l'adaptation de l'épinoche à la vie en lac, des macromutations récurrentes et indépendantes d'un enhancer du facteur de transcription *Pitx1* ont entraîné la réduction très nette des plaques osseuses sur les flancs et l'absence d'épines pelviennes (Chan et al., 2010). De même, la comparaison des génomes de l'homme et du chimpanzé indique l'absence chez l'homme de centaines de séquences fortement conservées entre les chimpanzés et les autres mammifères, suggérant la perte de nombreuses séquences régulatrices depuis la divergence avec la lignée conduisant au chimpanzé, dont certaines peuvent être reliées à l'acquisition de traits humains (McLean et al., 2011).

Enfin, des insertions de séquences peuvent également contribuer à l'évolution des profils d'expression et des phénotypes. Ainsi, l'invention du placenta chez les mammifères pourrait avoir résulté de nombreux événements d'insertion du transposon Mer20. Les copies insérées portent des signatures chromatiniennes d'enhancers actifs et pourraient avoir conduit à l'expression de plus de 1000 gènes dans cette nouvelle structure (Emera & Wagner, 2012; Lynch, Leclerc, May, & Wagner, 2011). Plus récemment dans l'histoire des mammifères, des insertions de transposons pourraient avoir contribué à la remarquable diversification de ce tissu entre groupes de mammifères et à l'évolution convergente de l'expression du gène de la prolactine dans l'endomètre (Emera et al., 2012).

Notons enfin qu'en plus de leur importance dans l'évolution de la lignée humaine (Siepel & Arbiza, 2014), les variations des séquences *cis*-régulatrices sont associées à de nombreuses maladies génétiques. 80% des variations observées dans les séquences non-codantes se situent

dans des séquences *cis*-régulatrices actives au cours du développement (Maurano et al., 2012; Vernot et al., 2012).

3. Les facteurs de transcription et leur spécificité de fixation sur l'ADN

Nous avons vu plus haut que les séquences *cis*-régulatrices fonctionnent par le truchement des protéines qui s'y fixent de façon spécifique comme les facteurs de transcription, ou aspécifique comme les histones. Dans cette nouvelle partie, je présenterai la structure et le mécanisme d'action des facteurs de transcription. J'introduirai également les méthodes utilisées *in vitro* et *in vivo* pour caractériser la spécificité de reconnaissance de l'ADN par ces protéines, et les différents modèles de représentation de cette spécificité.

a) Une grande variété de protéines fixant l'ADN

Les facteurs de transcription ont une structure modulaire qui comprend généralement un ou plusieurs domaines de fixation à l'ADN et un ou plusieurs domaines d'interaction avec d'autres protéines. Le premier permet à la protéine de se fixer sur ses TFBS (généralement de 5 à 12pb) avec une affinité dépendant de la séquence reconnue. Les seconds ont différents rôles, comme par exemple les interactions entre protéines, la localisation nucléaire ou la réception de messages hormonaux. Certains facteurs de transcription peuvent être bi-fonctionnels, combinant activité enzymatique et activité de facteur de transcription (Kozhevnikova et al., 2012). Le nombre de facteurs de transcription dans un génome varie d'un facteur 5 entre les bactéries (*E. coli*, 314 facteurs, (Pérez-Rueda & Collado-Vides, 2000)) et les mammifères (environ 1500 facteurs, (Vaquerizas, Kummerfeld, Teichmann, & Luscombe, 2009)). Cette diversification des facteurs de transcription contribue à la complexification du contrôle de l'expression génétique dans les animaux.

En dépit de leur grand nombre, les facteurs de transcription identifiés appartiennent à un nombre assez limité (~30) de familles protéiques définies par la présence de domaines structuraux de reconnaissance de l'ADN similaires (Wingender, Schoeps, & Dönitz, 2013). L'existence de plusieurs classes structurelles de facteurs de transcription illustre la diversité des solutions permettant l'interaction protéine-ADN. Dans la plupart des cas, néanmoins, c'est une

structure secondaire simple, généralement une hélice alpha, qui est responsable de l'interaction avec la double hélice d'ADN.

Les facteurs de transcription agissent via le recrutement de cofacteurs et peuvent agir comme activateurs ou répresseurs. Chacune de ces activités peut être associée à un motif peptidique. Ainsi, les répétitions de glutamines sont souvent trouvées dans des activateurs (Atanesyan, Günther, Dichtl, Georgiev, & Schaffner, 2012), alors que les motifs WRPW sont associés à la répression (A. L. Fisher, Ohsako, & Caudy, 1996). Une étude récente indique que l'activité de nombreux facteurs de transcription dépend du contexte *cis*-régulateur et cellulaire dans lequel ils agissent (Stampfel et al., 2015).

b) Détermination de la spécificité de fixation à l'ADN des facteurs de transcription, et de leurs sites de fixation dans les génomes.

La connaissance de l'affinité avec laquelle un facteur de transcription va lier telle ou telle séquence est cruciale pour la compréhension du fonctionnement d'un enhancer. Des méthodes *in vitro* et *in vivo* ont été développées pour identifier les séquences auxquelles se fixent les facteurs de transcription (Pour revue, voir Levo & Segal, 2014). Le principe général des méthodes *in vitro* est de révéler l'interaction en solution entre une protéine et une courte séquence d'ADN double brin. Les méthodes *in vivo* font, elles, appel à l'immunoprécipitation de fragments de chromatine, à partir d'un anticorps spécifique du facteur d'intérêt. Ces méthodes ont permis d'améliorer notre connaissance du répertoire des spécificités et des sites de fixation sur l'ADN des facteurs de transcription, de leur évolution.

i) Méthodes *in vitro*

Ces méthodes visent à caractériser les préférences en terme de séquence des facteurs de transcription. Ils consistent généralement à incuber en solution des protéines recombinantes avec de courts fragments d'ADN double brin, puis à visualiser l'interaction ou à sélectionner les fragments d'ADN sur la base de leur affinité.

Le retard sur gel (Electrophoretic Mobility Shift Assay ou EMSA)

Le retard sur gel est l'approche historique utilisée pour démontrer une interaction directe entre une protéine et un fragment d'ADN double brin (Fried & Crothers, 1981; Garner & Revzin, 1981; Hellman & Fried, 2007) . Cette méthode repose sur le fait que la mobilité d'un fragment d'ADN sur un gel de polyacrylamide est réduite s'il est lié par une ou plusieurs protéines. Techniquement, il s'agit de faire migrer, en conditions non-dénaturantes, des sondes (courts fragments d'ADN) marquées (par radioactivité ou fluorescence), seules et en présence d'une ou plusieurs protéines. Le résultat est d'abord visuel : l'ADN libre forme la bande qui migre le plus, et les complexes ADN-protéine(s) sont plus ou moins retardés en fonction de la taille et du nombre de protéines fixées (Figure I.6). L'analyse des images obtenues permet une quantification très précise. Le retard sur gel peut être réalisé sur des fragments d'ADN de différentes tailles, à partir de protéines ou domaines protéiques purifiés à partir d'un extrait protéique plus complexe tel qu'un extrait nucléaire.

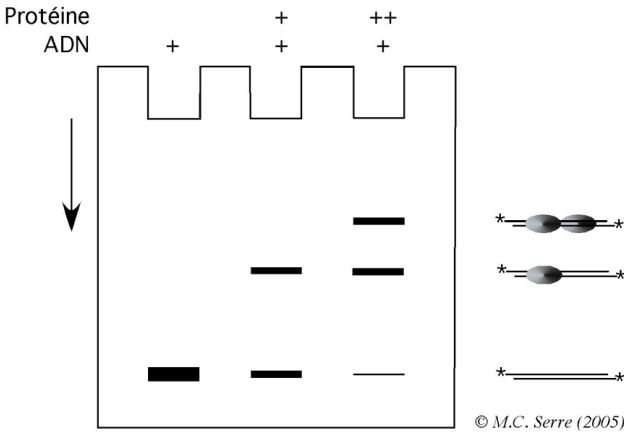


Figure I.6 : Principe du retard sur gel.
L'EMSA(Electrophoretic Mobility Shift Assay), est la technique la plus simple qui soit pour étudier une interaction entre un acide nucléique et des protéines. Elle repose sur le fait qu'un complexe ADN-protéine migrera moins vite dans un gel non-dénaturant qu'un ADN nu. Ce retard de migration permet de juger au premier coup d'œil si une séquence particulière d'ADN a été reconnue et liée par une protéine. L'illustration présente le cas de deux sites portés sur un même fragment. Si les deux sites fixent la même protéine, l'apparition du second complexe sera fonction de la concentration en protéine dans la réaction. (M.C. Serre, Université Paris-Sud).

Cette méthode permet de vérifier non seulement l'interaction d'une protéine avec une séquence d'ADN, mais aussi, par des expériences de mutations ponctuelles de la sonde, de déterminer l'importance des nucléotides dans cette séquence pour la fixation de la protéine. Elle permet également d'étudier la compétition ou la coopération entre protéines et de déterminer des constantes thermodynamiques de fixation (Man & Stormo, 2001). Plus récemment des méthodes plus quantitatives basées sur la résonnance plasmon de surface ont été développées (Stockley & Persson, 2009).

Le SELEX (Systematic Evolution of Ligands by EXponential enrichments : évolution systèmatique des ligands par enrichissement exponentiel)

Les expériences de retard sur gel nécessitent d'avoir une certaine connaissance *a priori* de la spécificité de fixation sur l'ADN de la protéine considérée. Le SELEX est une technique utilisée pour décrire l'affinité relative de liaison à l'ADN d'une protéine dont on ne connait rien de la spécificité de fixation sur l'ADN (Ellington & Szostak, 1990). Une protéine ou domaine protéique, recombinant(e) et fusionné(e) avec une étiquette reconnaissable, est incubé(e) avec un ensemble d'oligonucléotides double brin dégénérés. Les oligonucléotides fixés par la protéine sont sélectionnés par pull down avec l'étiquette de la protéine et séquencés. Afin de réduire le nombre de faux positifs, plusieurs cycles d'amplifications par PCR puis de sélection sont généralement nécessaires pour obtenir un enrichissement suffisant. Des protocoles de SELEX-seq automatisés pour l'identification à grande échelle des sites de fixation de facteurs de transcription ont été mis au point (Jolma et al., 2010). Le séquençage à haut débit permet de quantifier avec une très grande précision l'affinité relative de la protéine pour les différentes séquences liées. Couplées à des méthodes informatiques, elles permettent l'identification des spécificités de séquence d'un très grand nombre de facteurs de transcription. Ainsi, des données de SELEX-seq ont permis de déterminer les séquences consensus liées par plusieurs centaines de facteurs de transcription chez l'Homme (Jolma et al., 2013) ou chez la drosophile (Nitta et al., 2015). Une version séquentielle, permet même de caractériser la fixation coopérative de couples de facteurs de transcription (Jolma et al., 2015).

Autres méthodes

Il existe d'autres méthodes comme la technique du simple hybride (yeast one hybrid ou YIH) (J. J. Li & Herskowitz, 1993) ou les puces à ADN pour les protéines qui lient l'ADN (Protein-Binding DNA microarray ou PBM) (Berger & Bulyk, 2006). La première technique permet de ne pas avoir à produire de protéine recombinante *in vitro*, la seconde est une alternative, moins précise, au SELEX-seq. La combinaison des deux méthodes a notamment permis de réaliser un atlas de spécificité des facteurs de transcription du nématode *Caenorhabditis elegans* (Narasimhan et al., 2015). La combinaison de ces méthodes a aussi permis de montrer la conservation évolutive générale des spécificités de fixation sur l'ADN de plus de 1000 protéines issues d'espèces distantes (Weirauch et al., 2014).

Les techniques utilisées pour identifier les spécificités de liaison à l'ADN des facteurs de transcription génèrent un très grand nombre de séquences qui doivent être analysées. Il existe plusieurs méthodes plus ou moins informatives de représenter les TFBS en modélisant de façon quantitative l'affinité des TFs pour leurs sites de fixation: on peut utiliser une séquence consensus, une matrice, un logo de séquence, des k-mers, etc... (Figure I.7). L'enjeu est de prendre en compte le maximum d'information contenu dans ces résultats, comme par exemple l'interdépendance de différentes positions, sans pour autant augmenter la complexité des programmes de prédiction (pour revue Mathelier & Wasserman, 2013).

Résultats obtenus : vers un atlas compréhensif des spécificités de fixation sur l'ADN des facteurs de transcription

Combien de motifs sont reconnus par des facteurs de transcription au sein d'un génome ? Les expériences *in vitro* de caractérisation de la spécificité de reconnaissance des facteurs de transcription de toutes les familles majeures, et dans plusieurs systèmes modèles, ont permis de dresser un véritable lexique de fixation potentielle de facteurs de transcription, qui couvre probablement de l'ordre de 50% de l'ensemble des facteurs. De ce travail méthodique, consultable par exemple *via* la base de données Jaspar (Mathelier et al., 2015), émergent certaines règles.

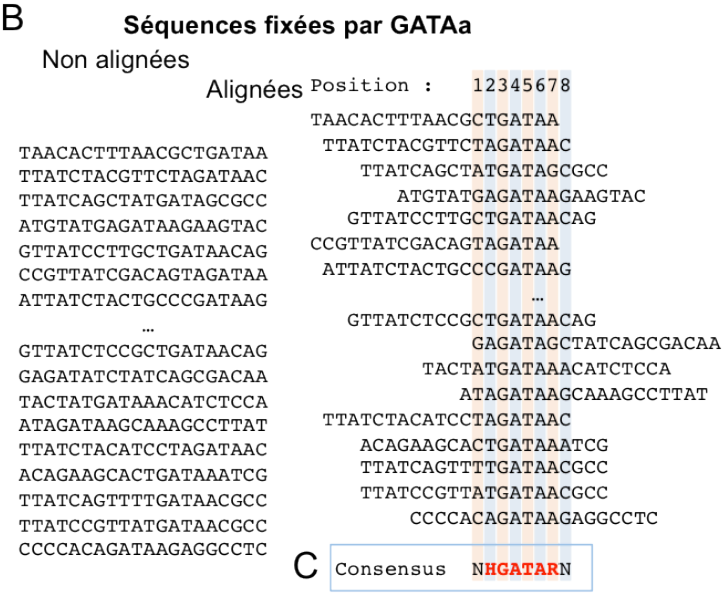
Premièrement, la diversité des spécificités de reconnaissance de l'ADN au sein de chaque famille structurelle est variable. Par exemple, les facteurs de la famille ETS reconnaissent des séquences très similaires (Wei et al., 2010), alors que les facteurs ayant des motifs de type doigts de Zinc C2H2, ont une grande diversité de reconnaissance (Garton et al., 2015; Najafabadi et al., 2015; Persikov et al., 2015). Ainsi, le nombre de motifs reconnus par les facteurs de transcription est plus petit que le nombre de facteurs de transcription, même si certains facteurs semblent avoir plusieurs modes de fixation sur l'ADN. Deuxièmement, l'analyse à grande échelle de la spécificité de fixation sur l'ADN d'un grand nombre de facteurs de transcription, indique que les spécificités de fixation sur l'ADN de différentes familles sont distinctes, même si un petit nombre de protéines appartenant à des familles distinctes peuvent avoir des spécificités de fixation identiques (Jolma et al., 2013). Enfin, la spécificité de fixation sur l'ADN d'un facteur de transcription est généralement conservée au cours de l'évolution et donc partagée avec tous ses paralogues et orthologues. Les données obtenues chez une espèce métazoaire peuvent donc généralement être utilisées dans toutes les autres (Nitta et al., 2015;

Weirauch et al., 2014). Il y a néanmoins des exceptions intéressantes de facteurs de transcription dont la spécificité a changé de manière adaptative, et l'importance de l'évolution de la spécificité de fixation sur l'ADN des facteurs de transcription ne doit pas être négligée (Lynch & Wagner, 2008; Lynch et al., 2008; Sayou et al., 2014).

A

Code	Représente
A	Adénine
G	Guanine
C	Cytosine
T	Thymine
Y	Pyrimidine (C ou T)
R	Purine (A ou G)
W	Weak (A or G)
S	Strong (G or C)
K	Keto (T or G)
M	Amino (C or A)
D	A, G, T
V	A, C, G
H	A, C, T
B	C, G, T
N (ou X)	Any base
-	Délétion

B



C

D

Position	1	2	3	4	5	6	7	8	9	10	11
A	1423	2782	708	0	4000	27	3887	3550	799	1432	1487
C	560	1633	31	0	0	29	0	4	681	897	829
G	1242	1235	10	4000	0	109	6	383	2296	1360	1099
T	775	424	1177	0	0	3835	107	63	224	311	585

Nombre d'apparition de chaque nucléotide à chaque position

E

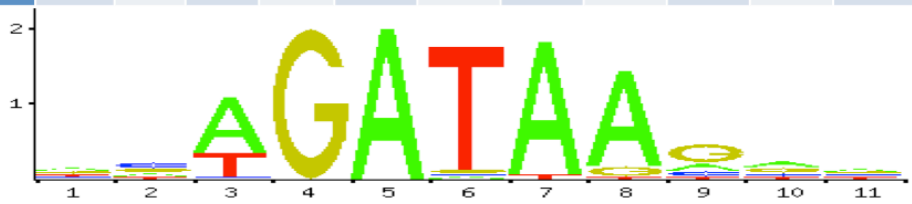


Figure I.7 : Décrire un site de fixation de facteur de transcription. Le code IUPAC de nomenclature des nucléotides **(A)** permet de résumer l'information contenue dans l'alignement de séquences fixées par un facteur de transcription **(B)** en une séquence consensus **(C)**. Cette information peut apparaître dans une matrice de fréquence à chaque position **(D)**, ou de façon beaucoup plus visuelle dans un logo de séquence **(E)**. L'exemple concerne la séquence reconnue par les facteurs de transcription GATA. Les séquences de l'alignement ont été choisies arbitrairement et proviennent de données de SELEX pour Gata4/5/6 chez *Ciona intestinalis* (Nitta et al, en préparation). La matrice et le logo correspondent à des données de ChiP-Seq pour GATA 1 chez la souris provenant de la base de donnée JASPAR.

Cet atlas de spécificité de fixation sur l'ADN, couplé au développement de logiciel de modélisation des sites et de recherche à l'échelle du génome (Slattery et al., 2014), constitue un outil indispensable pour la recherche de sites de fixation dans des enhancers identifiés, permettant de catégoriser leur activité. Cet atlas permet notamment de relier le "lexique régulatoire" - l'ensemble des motifs extraits à l'échelle génomique des sites protégés de l'action de la DNAse1 identifiés par DNase-seq (Neph et al., 2012; Pique-Regi et al., 2011) - à l'ensemble des facteurs de transcription.

Enfin, la connaissance de la spécificité de fixation sur l'ADN d'un jeu de facteurs et de leur concentration au sein du noyau peut être utilisée pour construire des modèles thermodynamiques de réseaux de régulation, comme cela a été montré avec succès par Segal et collègues, dans le cas du réseau de segmentation chez la drosophile (Segal, Raveh-Sadka, Schroeder, Unnerstall, & Gaul, 2008). La recherche de regroupements de sites prédits de fixation pour un set précis de facteurs de transcription peut aussi permettre l'identification d'enhancers ayant une spécificité d'action précise (Khoueiry et al., 2010; Markstein, Markstein, Markstein, & Levine, 2002).

ii) Approches *in vivo* : ChIP-on-chip et ChiP-seq

Des techniques permettant d'identifier les séquences génomiques sur lesquelles se fixe un facteur de transcription *in vivo* ont donc été développées, basées sur l'immunoprécipitation de fragments d'ADN fixés *in vivo* par des facteurs de transcription. Ces fragments contenant les sites de fixation à l'ADN sont ensuite identifiés à grande échelle, initialement par hybridation sur puce (ChIP-on-chip), et depuis quelques années, séquencés à haut débit (ChIP-seq). Ces expériences sont réalisées sur des extraits cellulaires provenant d'un seul ou de plusieurs tissus, voire d'organismes entiers. Ils peuvent donc être spécifiques du type cellulaire et de la période de développement étudiés. Ces techniques ont notamment été utilisées à grande échelle lors des projets ENCODE et ModENCODE visant à décoder la syntaxe *cis*-régulatrice dans les organismes modèles majeurs (J. B. Brown & Celniker, 2015; Sloan et al., 2015).

Résultats obtenus : émergence d'une syntaxe cis-régulatrice

Ces expériences haut-débit *in vivo* ont permis de confirmer et de quantifier la prévalence de mécanismes initialement identifiés par des expériences fonctionnelles à petite échelle de dissection de séquences régulatrices.

Ces expériences ont révélé que les spécificités de fixation sur l'ADN *in vitro* et *in vivo* des facteurs de transcription sont généralement similaires en terme de séquences (Orenstein & Shamir, 2014) (nous verrons plus bas que ce constat doit être mitigé lorsque l'on prend en compte la structure tridimensionnelle de l'hélice d'ADN). Deux différences existent néanmoins entre les deux sets de données. D'une part, la fixation collaborative de plusieurs facteurs *in vivo* peut conduire à l'identification de sites correspondants à la fixation d'hétérodimères. D'autre part, on observe le recrutement *in vivo* de certains facteurs à des sites génomiques dont la séquence ne correspond pas au site reconnu par ce facteur *in vitro*. Ce phénomène peut parfois constituer jusqu'à 45% des signaux d'une expérience de ChIP-seq. Il a été étudié en détail par Wassermann et ses collègues : ces sites de fixation inattendus sont généralement localisés dans des régions de chromatine ouverte, fortement enrichies en sites, dits "zinger", ressemblant à ceux des protéines CTCF, JUN, ETS et THAP11 (Worsley Hunt & Wasserman, 2014). La raison de cet enrichissement, qui pourrait refléter un artefact de la technique (notamment du à l'étape de cross-linking) ou le recrutement indirect préférentiel des facteurs étudiés, est pour le moment assez mal comprise. Il est intéressant de noter néanmoins que des régions dites HOT (pour Highly Occupied Target), qui fixent un grand nombre de facteurs de transcription sans pour autant contenir leur sites de fixation, ont dans leur grande majorité une activité enhancer endogène spécifique d'un tissu chez la drosophile (Kvon, Stampfel, Yáñez-Cuna, Dickson, & Stark, 2012). Ces régions disparaissent lorsque l'étape de crosslinking est omise, suggérant que l'association d'un grand nombre de facteurs à des séquences régulatrices pour des gènes fortement exprimés pourrait être artefactuelle (Kasinathan, Orsi, Zentner, Ahmad, & Henikoff, 2014). Malgré ces limitations, l'analyse et l'intégration informatique des données issues d'expériences d'immunoprécipitation de chromatine ont aussi eu un fort impact sur notre compréhension de l'architecture *cis*-régulatrice au cours du développement.

Si le recrutement de facteurs de transcription à des *loci* qui ne contiennent pas de sites canoniques pour ces facteurs est plus fréquent que prévu, le recrutement de facteurs à des sites contenant des sites de fixation mais ne semblant pas avoir d'activité enhancer a aussi été une

surprise de ces approches (W. W. Fisher et al., 2012; Xiao-yong Li et al., 2008). Alors que l'on s'attendait à ce qu'un facteur de transcription ne contrôle l'expression que de quelques centaines de gène cibles, le nombre de sites fixés est généralement de l'ordre de plusieurs milliers, voire de la dizaine de milliers. Un des challenges du domaine est donc à la fois de comprendre les raisons de la forte densité de fixation et d'identifier la minorité de sites fonctionnels.

Les expériences de cartographie de sites *in vivo* ont confirmé que la plupart des sites de fixation de facteurs de transcription sont localisés dans des régions de chromatine ouverte identifiées par des expériences de protection à la DNAse, FAIRE ou ATAC-seq (Thurman et al., 2012). L'analyse de la dynamique d'occupation dans le temps a permis de distinguer plusieurs modes d'interactions entre facteurs de transcription et chromatine. Un petit nombre de facteurs, dits pionniers, sont capables de se fixer sur des sites normalement occupés par des nucléosomes, en déplaçant ceux-ci (Magnani, Eeckhoute, & Lupien, 2011; Zaret & Carroll, 2011). Un nombre plus important de facteurs sont dits sédentaires, car ils se fixent de manière stable sur la chromatine ouverte. Enfin, une troisième catégorie, majoritaire, les facteurs migrants, se fixent de manière plus sporadique sur leurs sites de fixation lorsque la chromatine est ouverte, peut-être parce qu'ils ont besoin pour se fixer de coopérer/interagir avec d'autres facteurs dont la disponibilité est limitative (Sherwood et al., 2014).

La cartographie *in vivo* a également confirmé que la plupart des enhancers sont constitués de regroupements (clusters) de sites de fixation sur l'ADN pour plusieurs facteurs de transcription (He et al., 2011; Wunderlich & Leonid, 2009). Dans certains cas, il a été possible d'identifier des facteurs "partenaires" qui ont tendance à se fixer sur des sites voisins à de nombreux *loci*, ce qui a permis de mieux comprendre la logique combinatoire donnant leur spécificité aux éléments. La co-localisation de sites occupés par le facteur de transcription MAD, relai de la signalisation Bmp, et de facteurs spécifiques de certains lignages a ainsi permis de déchiffrer la logique de la réponse spécifique tissulaire à un signal inducteur générique (Mullen et al., 2011; Trompouki et al., 2011). En combinant une approche d'immunoprécipitation de facteurs agissant dans le même réseau génétique avec des approches informatiques d'apprentissage, Zinzen et ses collègues (Zinzen, Girardot, Gagneur, Braun, & Furlong, 2009) ont pu prédire l'activité spatio-temporelle d'enhancers mésodermiques chez la drosophile sur la base de la combinatoire d'occupation par 5 facteurs de transcription (Twist, MEF2, Tinman, Bagpipe et Biniou). La présence d'un regroupement de sites de fixation à un locus donné permet généralement d'augmenter la confiance dans la fonctionnalité des interactions identifiées (He et

al., 2011), tout comme la nature temporellement dynamique du profil de fixation (Wilczyński & Furlong, 2010).

Les approches d'immunoprécipitation de chromatine ont aussi permis de mieux décrire l'évolution des séquences régulatrices. Ces approches ont permis d'identifier les enhancers "fantômes" et leur importance pour la robustesse des programmes développementaux (voir plus haut). Chez la drosophile, le niveau d'affinité des sites de fixation des facteurs Hox est également conservé et impliqué dans la robustesse des profils d'expression des gènes développementaux (Crocker et al., 2015).

Au cours de l'évolution, la position des sites de fixation de facteurs de transcription sur l'ADN peut changer, phénomène très fréquent appelé "turn-over" (ou rotation). Des modes d'évolution contrastés de ce phénomène ont été décrits chez les vertébrés et la drosophile. Chez les vertébrés, la comparaison des profils d'occupation de 5 facteurs de transcription (dont CTCF) dans le foie montre une grande divergence, malgré la conservation de la spécificité de fixation *in vitro* sur l'ADN de ces facteurs : moins de 10% des sites de fixation du facteur CEBPA sont conservés entre l'oppossum et l'homme (Schmidt, Wilson, et al., 2010). Cette divergence extensive des sites occupés par des facteurs de transcription a été confirmée par des expériences indépendantes de cartographie de la chromatine ouverte par DNase-seq (Stergachis et al., 2014). Par contraste, la position des sites *in vivo* de fixation de facteurs de transcription est beaucoup mieux conservée chez les drosophiles que chez les vertébrés, même au sein de groupes contenant des espèces ayant une distance évolutive similaire (He et al., 2011). Les raisons de cette différence entre espèces restent mystérieuses, la non-fonctionnalité de la plupart des sites de fixation détectés chez l'homme serait une explication possible. On peut également noter que les expériences chez les vertébrés ont été faites à partir de tissu adulte, alors que les expériences chez la drosophile sont faites avec des embryons. Mais il est aussi possible que la divergence des résultats provienne du simple choix des facteurs analysés.

4. Structure de la chromatine, structure locale de l'ADN et transcription

L'ADN est une molécule qui a une structure physique tridimensionnelle, ce que l'on a tendance à oublier quand, par exemple, on analyse des résultats de séquençage. Cependant, la structure locale de la double hélice et l'organisation de la chromatine sont intrinsèquement

liées à la régulation de la transcription car elles peuvent notamment jouer sur l'affinité d'un facteur de transcription pour son site de fixation et sur l'accessibilité de l'ADN aux facteurs de transcription.

a) Chromatine et transcription

i) Les nucléosomes

Dans le noyau des cellules eucaryotes, l'ADN est compacté sous forme de chromatine. Le premier niveau d'organisation est l'enroulement de l'ADN autour d'octamères d'histones (composés de deux hétérodimères histone2A-histones2B et de deux autres hétérodimères histone3-histone4 (Richmond & Davey, 2003)). 147 paires de bases sont enroulées autour de chacune de ces structures globulaires et le produit de l'association entre un fragment d'ADN et un oligomère d'histones est un nucléosome, d'un diamètre d'environ 10nm. La chromatine est ainsi constituée de la succession régulière de nucléosomes. Deux nucléosomes adjacents sont reliés par un ADN de liaison (DNA linker) sur lequel peut se fixer l'histone de liaison H1 ou son variant H5 qui facilite la condensation de la chromatine (Hergeth, Schneider, Hergeth, & Schneider, 2015). La fréquence d'apparition des nucléosomes sur l'ADN permet de distinguer l'euchromatine de l'hétérochromatine et donne une information de permissivité pour la transcription des gènes qui leur sont associés. L'euchromatine a une conformation lâche due à des nucléosomes plus dispersés et est associée principalement à des gènes actifs, tandis que l'hétérochromatine est compactée par un niveau supérieur d'organisation de ses nucléosomes dans lequel sont impliquées d'autres protéines et est associée principalement à des gènes réprimés (Bassett, Cooper, Wu, & Travers, 2009). Comme nous l'avons vu plus haut, ces deux conformations correspondent à des compartiments distincts dans le noyau (Lieberman-Aiden et al., 2009).

On peut distinguer l'hétérochromatine constitutive de l'hétérochromatine facultative dont la dynamique est associée à celle des gènes régulés qui y sont localisés. Différents complexes de protéines sont impliqués dans leur formation, par exemple HP1 pour la l'hétérochromatine constitutive et Polycomb pour l'hétérochromatine facultative pendant la différenciation et le développement (Saksouk, Simboeck, & Déjardin, 2015). L'ouverture de la chromatine peut être initiée par la fixation de facteurs de transcription pionniers, tels que GATA1 et FoxA,

capables de se fixer et d'ouvrir localement la chromatine condensée afin de faciliter la fixation d'autres protéines (Zaret & Carroll, 2011).

A l'échelle du génome, le positionnement des nucléosomes est déterminé de façon dynamique par les enzymes de remodelage de la chromatine, les protéines fixant l'ADN (facteurs de transcription, PIC, ARNpolII) (Struhl & Segal, 2013) et la séquence nucléotidique, qui influe indirectement sur la formation de nucléosomes, même si les octamères d'histones n'ont pas une spécificité de contact avec l'ADN (Sathyapriya, Vijayabaskar, & Vishveshwara, 2008). Nous avons déjà abordé cette problématique plus haut lors de la discussion sur les facteurs pionniers, sédentaires ou migrants, par exemple.

En plus des techniques d'analyse de la chromatine *in vivo* associées au séquençage haut débit comme la DNase-seq (Kharchenko et al., 2011; Thurman et al., 2012), le FAIRE-Seq (Giresi, Kim, McDaniell, Iyer, & Lieb, 2007) ou, plus récemment l'ATAC-seq (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013), il existe également des méthodes *in silico* pour prédire le positionnement et l'occupation des nucléosomes sur un fragment d'ADN à partir de sa séquence. Ces approches, basées sur l'analyse des propriétés physiques locales de l'hélice, ne peuvent pas tenir compte de l'état dynamique de la cellule et de la régulation de la transcription. Elles ne fournissent donc qu'une information générique sur les régions les plus probablement occupées par des nucléosomes, mais permettent d'avoir une vision sur les séquences régulatrices complémentaires de celles transmises par les expériences *in vivo*.

Par exemple, Segal et ses collègues ont développé un modèle probabiliste basé sur les séquences de sets de 199 mononucléosomes chez la levure et de 177 nucléosomes naturels chez le poulet (Segal et al., 2006). Chez la levure et le poulet, les nucléosomes sont préférentiellement répartis dans des régions contenant des dinucléotides "AA"/"TT"/"AT" avec une périodicité de 10pb environ. Ce motif reflèterait une signature pour la présence de nucléosomes à la fois *in vivo* et *in vitro*. Ce modèle estime d'abord la distribution des dinucléotides à chaque position pour des centaines de séquences nucléosomales obtenues expérimentalement. Ces probabilités sont converties en énergie libre apparente et un modèle thermodynamique prenant en compte la compétition thermodynamique entre les nucléosomes, l'encombrement stérique et la concentration en nucléosomes a été développé. Ce modèle "glisse" le long de la séquence d'ADN et estimerait la probabilité que chaque paire de bases soit occupée par un nucléosome(Segal et al., 2006). D'autres méthodes semblent plus

performantes pour prédire le positionnement des nucléosomes (Field et al., 2008; Kaplan et al., 2009) mais leur efficacité varie en fonction de la région du génome considérée : ces trois méthodes sont par exemple comparables lorsqu'il s'agit de prédire la position des nucléosomes sur des promoteurs murins (Liu et al., 2013). Toutefois, si elle ne permet pas forcément de prédire l'emplacement des nucléosomes, la signature dinucléotidique utilisée par Segal et ses collègues en 2006 corrèle avec l'activité d'enhancers chez la drosophile et l'ascidie (Khoueiry et al., 2010).

ii) Les modifications post-traductionnelles des histones et la compaction de la chromatine

Les histones portent des marques épigénétiques sous forme de modifications covalentes de leur queue N-terminale (acétylation, phosphorylation, ubiquitination, méthylation, ADP ribosylation, glycolysation, sumoylation) associées à différentes fonctions *cis*-régulatrices (Zentner & Henikoff, 2013). L'analyse du rôle fonctionnel des ces modifications a conduit à proposer l'existence d'un code d'histone qui déterminerait la structure des nucléosomes et ainsi l'état de compaction chromatinien des séquences considérées (Jenuwein & Allis, 2001). Les promoteurs sont enrichis en modifications post-traductionnelles de l'histone H3, notamment la triméthylation de la lysine 4 (H3K4me3) (Heintzman et al., 2007). Les enhancers sont également enrichis en histone H3 monométhylée sur la lysine 4 (H3K4me1) et en histone H3 acétylée sur la lysine 27 (H3K27ac) (Bulger & Groudine, 2011). Ces marques sont dynamiques, apportées et retirées par des enzymes spécifiques qui agissent de concert avec les facteurs de transcription et leurs protéines associées. Les histones acétylées sont reconnues par des protéines qui contiennent un bromodomaine, telles que des lysine-acéthyl-transférases et des re-modeleurs de la chromatine ; elles peuvent donc être impliquées dans la régulation de la conformation de la chromatine. La méthylation des lysines est reconnue par des protéines contenant des domaines protéiques Tudor ou Chromo par exemple. En particulier, la méthylation de la lysine 27 de l'histone 3, associée à la répression de l'expression des gènes, est reconnue par le complexe Polycomb (Zentner & Henikoff, 2013).

Il est donc essentiel de prendre en compte dans l'analyse des enhancers ce niveau supplémentaire de régulation de la transcription que sont le positionnement des nucléosomes et leurs modifications épigénétiques. Pourtant, notons que nous sommes parfois confrontés au problème de la poule et de l'œuf. Nous ignorons souvent l'enchainement précis des évènements

qui conduisent à un enhancer actif, dont la chromatine est ouverte et sur lequel les facteurs de transcription sont fixés. Dans certains cas, c'est probablement la fixation de facteurs pionniers qui conduit à l'ouverture de la chromatine. Dans d'autres cas, l'ouverture de la chromatine *via* l'action d'enzymes de modification des histones est peut-être l'élément initial. Dans d'autres cas, enfin, il est probable que ce soit la compétition thermodynamique entre facteurs de transcription et nucléosomes, éventuellement biaisée par les propriétés physiques locales de la double hélice qui soit déclenchante (Barozzi et al., 2014).

b) Rôle de la structure locale de l'hélice d'ADN dans son interaction avec les facteurs de transcription

Nous avons jusqu'à maintenant principalement focalisé notre attention sur la séquence primaire de la molécule d'ADN. Des données récentes suggèrent que les propriétés physiques de l'hélice d'ADN doivent également être prises en compte lors de l'étude des éléments *cis*-régulateurs.

Des approches basées sur la conservation de la structure 3D de la molécule d'ADN ont montré que certaines de ses propriétés sont sous pression de sélection (pour revue Parker & Tullius, 2011). L'identification des régions, dont la structure est contrainte évolutivement, a permis de détecter deux fois plus de régions conservées que les approches traditionnelles de conservation de séquence nucléotidique ; la même forme de l'hélice pouvant être générée par différentes séquences d'ADN. Ainsi 12% des bases du génome humain sont évolutivement contraintes du point de vue de la structure de l'ADN et ces régions sont enrichies en éléments fonctionnels, dont des enhancers (Stephen C. J. Parker, Hansen, Abaan, Tullius, & Margulies, 2009). Cet enrichissement y est plus important que dans les séquences conservées au niveau de la séquence uniquement.

Ainsi, la forme locale de l'ADN est importante pour au moins certains aspects des éléments fonctionnels du génome : dans les séquences codantes, la conservation des triplets de nucléotides est généralement couplée à la conservation de la forme de l'hélice. En effet, dans la double hélice d'ADN, les atomes du squelette situés sur les brins complémentaires les plus proches de part et d'autre du petit sillon, et donc impliqués dans la structure de l'hélice, sont séparés par trois nucléotides (Rohs 2009). Mais la plus grande partie des bases sous contraintes évolutives sont localisées dans des régions non-codantes (ENCODE). Dans le génome humain,

des SNPs associés à différents phénotypes entrainent des changements importants de la forme locale de l'ADN (Stephen C. J. Parker et al., 2009). Ces changements de structure, affectant des fonctions biologiques dans des régions non-codantes évolutivement conservées, suggèrent que cet effet est local et que la topographie de l'ADN à un locus donne une information de permissivité pour une fonction régulatrice qui peut être soumise à des pressions de sélection.

Nous savons depuis longtemps que les facteurs de transcription reconnaissent des séquences nucléotidiques spécifiques, plus ou moins dégénérées (voir plus haut). Cette reconnaissance se fait principalement par interaction avec le grand sillon de l'hélice (Seeman, Rosenberg, & Rich, 1976). La cristallisation de complexes protéine-ADN a révélé qu'en plus de ce mode de reconnaissance, certains facteurs de transcription reconnaissent également la forme locale de la double hélice *via* une interaction avec son petit sillon (Joshi et al., 2007). Ce second niveau de reconnaissance pourrait conférer aux facteurs de transcription une plus grande spécificité de reconnaissance qu'initialement anticipé (Zhou et al., 2015).

White et ses collègues ont étudié chez la souris des séquences contenant des TFBS potentiels pour le facteur de transcription à homéodomaine *Crx,* fixés et non fixés par cette protéine à l'échelle du génome (White, Myers, Corbo, & Cohen, 2013). Ils ont montré que le potentiel *cis*-régulateur des séquences fixées par le facteur de transcription réside dans des caractéristiques de séquences très locales et non dans leur contexte génomique : un taux de GC élevé corrèle avec la présence de sites de fixation occupés. Un fort taux de GC peut être associé avec plusieurs déterminants de l'activité d'un enhancer comme la présence de nucléosomes (Kaplan et al., 2009) ou la profondeur du petit sillon (Bishop et al., 2011). Le calcul de la taille du petit sillon a permis de montrer que les sites effectivement liés par *Crx* ont un sillon plus profond que les sites potentiels non fixés, à affinité théorique de séquence égale. Ainsi, la taille du petit sillon pourrait être une signature très localisée qui permettrait de distinguer les TFBS fonctionnels et non fonctionnels (White et al., 2013).

La forme et les propriétés physiques de l'hélice d'ADN dépendent elles-mêmes de sa séquence (Rohs et al., 2010), mais une étude récente combinant biologie structurale et analyse de spécificité de fixation sur l'ADN a permis d'identifier les résidus de la protéine Hox src impliqués dans la reconnaissance du petit sillon par l'hétérodimère exd-src. La mutation de ces résidus affecte la spécificité globale de reconnaissance des sites sans altérer la reconnaissance de la séquence du site, indiquant que la séquence et la forme sont lues de manière indépendante

(Abe et al., 2015). Cette étude révèle par ailleurs que l'inclusion de paramètres de forme améliore les performances des modèles informatiques de reconnaissance de sites cibles par exd-src. Enfin, la même équipe a montré que dans de nombreuses familles de facteurs de transcription, l'environnement immédiat de la séquence reconnue, qui impacte la structure de l'hélice mais ne contient qu'une faible information de séquence, joue un rôle important dans la fixation du facteur (Dror, Golan, Levy, Rohs, & Mandel-Gutfreund, 2015).

Dans l'ensemble, ces données suggèrent que la reconnaissance de leurs sites de fixation par des facteurs de transcription doit intégrer à la fois la séquence du site et la structure 3D locale de l'hélice d'ADN. Des modèles thermodynamiques basés sur des données de cristallographie de courtes séquences d'ADN ont été développés pour prédire la forme et les propriétés physiques d'une l'hélice d'ADN en fonction de sa séquence (Bishop et al., 2011; Broos et al., 2013; Durán et al., 2013).

B. Les signatures des enhancers

1. Identification des enhancers

Il existe de nombreuses méthodes pour prédire des séquences *cis*-régulatrices. Ces dernières années, grâce au coût décroissant du séquençage haut débit notamment, elles ont été utilisées à l'échelle de génomes entiers. Mais la seule séquence des régions ainsi identifiées ne permet pas de prédire leur activité – lorsqu'elles en ont. C'est pourquoi, chacune de ces prédictions doit être testée (Hardison & Taylor, 2012).

a) Méthodes de prédictions des séquences *cis*-régulatrices fonctionnelles

Les propriétés des enhancers et de la chromatine associée listées plus haut sont utilisées pour tenter de prédire tous les enhancers actifs d'un génome, dans un tissu et/ou à un stade de développement spécifique (Figure I.8) (pour revue, Shlyueva et al., 2014).

L'empreinte phylogénétique est une méthode qui repose sur l'hypothèse que les régions non codantes fonctionnelles sont évolutivement conservées car elles sont importantes pour l'organisme et, notamment, impliquées dans la régulation de l'expression des gènes : de telles séquences pourraient donc être des enhancers (Tagle et al., 1988). Toutefois, cette méthode ne

permet pas de distinguer la nature des éléments non codants conservés et potentiellement fonctionnels et, l'activité d'enhancers peut être conservée indépendamment de leur séquence (Weirauch & Hughes, 2010).

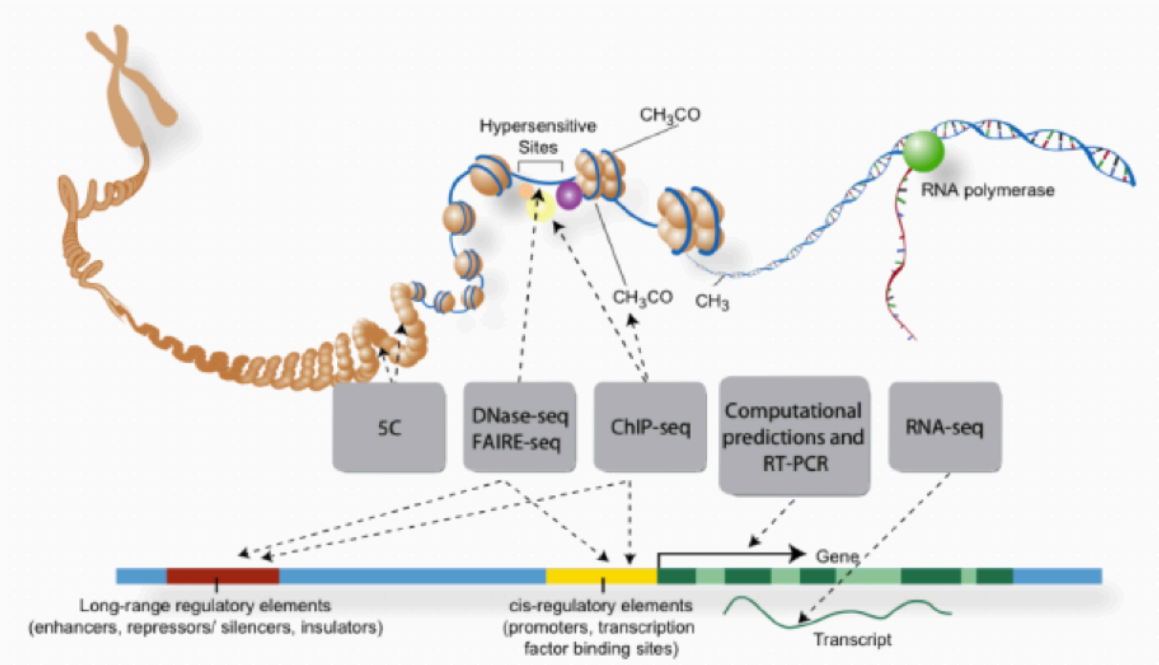


Figure 1.8 : **Méthodes utilisées pour prédire les séquences régulatrices à l'échelle d'un génome.** Le séquençage d'ARNm provenant de différentes sources permet de savoir à quel moment du développement et/ou dans quel tissu un gène est actif. Plusieurs stratégies *in vivo* sont utilisées pour prédire les éléments *cis*-régulateurs. Les régions fixées par des facteurs de transcription et du co-activateur p300 sont mises en évidence par ChIP-séq. Des expériences d'hypersensibilité à la DNAse I et de FAIRE-séq révèlent l'accessibilité de l'ADN. L'étude de la méthylation de l'ADN ou des données de ChIP-séq de marques d'histones spécifiques permettent de définir différents états chromatiniens. Enfin, les données de 5C identifient les interactions entre différentes parties du génome. A ces techniques *in vivo* s'ajoutent les méthodes de prédiction bioinformatiques. (Crédits : site du projet ENCODE http://genome.ucsc.edu/ENCODE)

Différentes approches listées plus haut permettent de prédire *in silico* les sites de fixation potentiels des facteurs de transcription sur la base de la séquence génomique, ou de localiser sur le génome la fixation *in vivo* de facteurs de transcription donnés. Les regroupements contenant des combinaisons spécifiques de sites de fixations de facteurs de transcription sont de bons candidats.

Enfin, des méthodes *in vivo* d'analyse de l'ADN génomique permettent de localiser les nucléosomes, les différentes modifications d'histones et les autres protéines fixées à l'ADN associées à l'activation de la transcription (Hardison & Taylor, 2012).

Ces approches sont généralement combinées et permettent d'annoter les génomes et d'y prédire des régions fonctionnelles comme le font les consortiums des projets ENCODE (Encyclopedia of DNA Elements) chez l'homme, modENCODE chez la drosophile et le ver et mouseENCODE chez la souris. Combiner plusieurs signatures augmente l'efficacité des prédictions des enhancers dont l'activité doit toutefois être validée expérimentalement. (Dunham et al., 2012; Shen et al., 2012; Zhu et al., 2013)

b) Validation expérimentale de l'activité d'un enhancer

Traditionnellement, l'activité d'un enhancer est mise en évidence grâce à une construction rapporteur. Le fragment d'ADN testé est cloné dans un plasmide en amont d'un promoteur et d'un gène rapporteur (*LacZ* encodant la ß-galactosidase par exemple). Ce plasmide est ensuite introduit *in vivo* dans un organisme et permet de visualiser l'activité spatiotemporelle de la région testée. Il peut également être transfecté dans des lignées cellulaires, permettant un suivi plus quantitatif de son activité. Ces expériences permettent de déterminer si la région testée est suffisante pour générer un profil d'expression donné. Le nombre d'enhancers putatifs validés par cette méthode est limité par l'étape de clonage et n'excède pas quelques douzaines par projet.

Les approches de type ChIP-seq ou DNase-seq ont permis d'identifier un grand nombre d'enhancers putatifs (10 000 à 150 000 selon le type cellulaire), mais la majorité de ces enhancers n'a pas encore pu être testée fonctionnellement. Ainsi, on ne sait généralement pas combien d'enhancers sont effectivement actifs dans tel ou tel tissu ou à un stade du développement donné, ni quel(s) gène(s) est (sont) régulé(s) par ces enhancers potentiels.

Ainsi, des méthodes haut débit ont été développées pour permettre de tester le nombre croissant de séquences *cis*-régulatrices prédites ou synthétiques. Elles permettent de tester simultanément l'activité de milliers de séquences *cis*-régulatrices produites par la synthèse massive d'oligonucléotides. Les deux principaux types d'approches reposent sur un système de

code barre transcrit et sur l'intensité de la fluorescence associée à l'expression du gène rapporteur respectivement (Figure I.9). Un « code-barres » transcrit unique associé à chaque région testée permet d'extraire leurs niveaux d'activités relatifs par séquençage de l'ARNm (Farley et al., 2015; Kheradpour et al., 2013; Nam & Davidson, 2012). L'enhancer-FACS-seq (eFS, Fluorescent Activated Cell Sorter followed by sequencing), repose sur le tri des cellules transfectées en fonction du niveau de fluorescence produit par le gène rapporteur. Cependant, aucune de ces méthodes ne permet de connaître le profil d'activité spatial des séquences testées (pour revue, Weingarten-Gabbay & Segal, 2014).

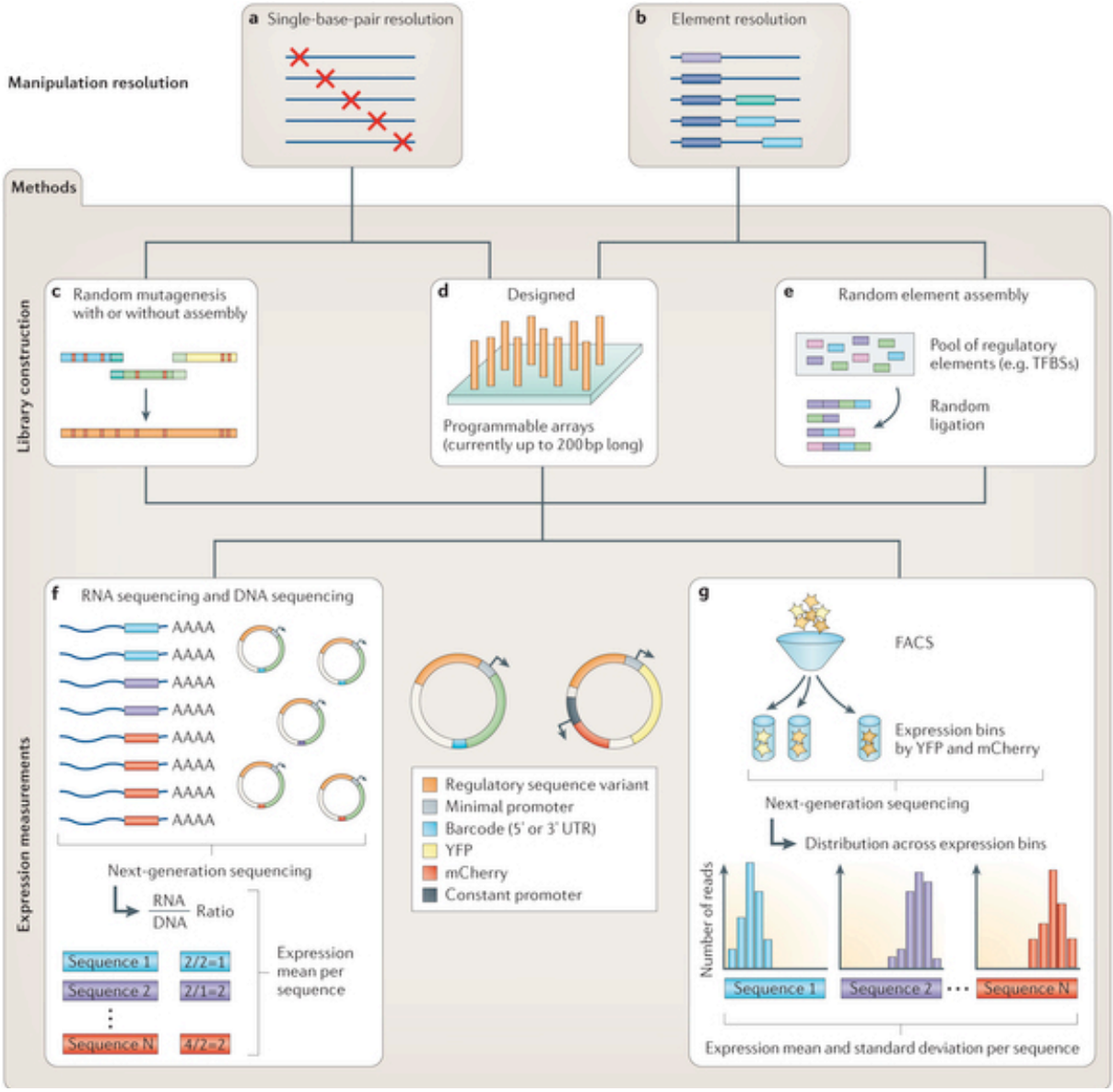


Figure I.9 : Stratégies des cribles à haut débit de l'activité enhancer. La première ligne décrit le type d'information pouvant être obtenu en fonction du type de constructions testées. La boîte « méthodes » présente les principales stratégies de construction de banques d'enhancers synthétique et de mesure de leur activité (Levo & Segal, 2014).

Le récent développement du STARR-Seq (Self-Transcribing Active Regulatory Region Sequencing) a permis de tester sans *a priori* le potentiel *cis*-régulateur de fragments d'ADN couvrant tout le génome de la drosophile (Arnold et al., 2013). L'ADN génomique est fragmenté, chaque fragment est cloné en aval d'un promoteur dans une construction rapporteur qui est ensuite transfectée dans différentes lignées cellulaires. La mesure du niveau relatif d'activité des séquences testées se fait par simple séquençage de l'ARNm, chaque enhancer activant sa propre transcription. Cette méthode permet de s'affranchir du travail de prédiction en amont et son débit est plus élevé que toutes les méthodes précédemment utilisées.

Ainsi, des milliers d'enhancers ont été identifiés chez différentes espèces-modèles telles que la drosophile (Arnold et al., 2013; Gisselbrecht et al., 2013), la souris (Patwardhan et al., 2012), l'Homme (Melnikov et al., 2012) ou l'oursin (Nam & Davidson, 2012). La validation des différents jeux de données permet d'optimiser les méthodes de prédiction. Toutefois, les enhancers endogènes identifiés, même spécifiques d'un seul tissu ou d'un seul stade développemental, sont très différents : les déterminants connus de l'activité d'un enhancer sont nombreux, pas toujours compris, et leurs apports respectifs généralement mal compris.

2. Comprendre la logique intrinsèque des enhancers

La section précédente a présenté un panorama global de nos connaissances sur les séquences régulatrices, en particulier les enhancers. Elle a illustré l'évolution progressive des approches utilisées, depuis les expériences initiales à petite échelle jusqu'aux approches à grande échelle de la dernière décennie, qu'elles soient descriptives, comme la cartographie de la chromatine et de son organisation tridimensionnelle et de la fixation sur l'ADN de facteurs de transcription individuels, ou fonctionnelles, comme l'identification à grande échelle de séquences régulatrices actives, via le STARR-seq par exemple.

Les dix dernières années ont indéniablement conduit à un raffinement quantitatif de notre compréhension du mode d'action des facteurs de transcription et des enhancers et à une meilleure estimation de l'importance de ces derniers au cours de l'évolution animale et dans les pathologies humaines. Elles ont enfin permis d'identifier un petit nombre de concepts nouveaux comme celui de paysage régulatoire, de domaine topologique, d'enhancer "de l'ombre", ou de reconnaissance de la forme de l'hélice d'ADN par les facteurs de transcription.

Malgré ces avancées, notre compréhension des enhancers demeure très parcellaire. La grande majorité des "règles" issues de ces études semble, comme c'est souvent le cas en biologie, ne concerner qu'une fraction des éléments caractérisés. La figure I.10 illustre ainsi que, parmi toutes les signatures de l'identité enhancer ou de l'activité d'un enhancer chez les métazoaires, les seules qui semblent vraiment nécessaires sont :

1) la présence de regroupement de sites de fixation pour certains facteurs de transcription (Identité);

2) leur occupation par des facteurs (activité), le recrutement de cofacteurs (activité) et l'établissement d'une communication avec un promoteur (activité).

Les autres signatures ne sont, elles, présentes que dans des sous-populations, partiellement chevauchantes, des enhancers.

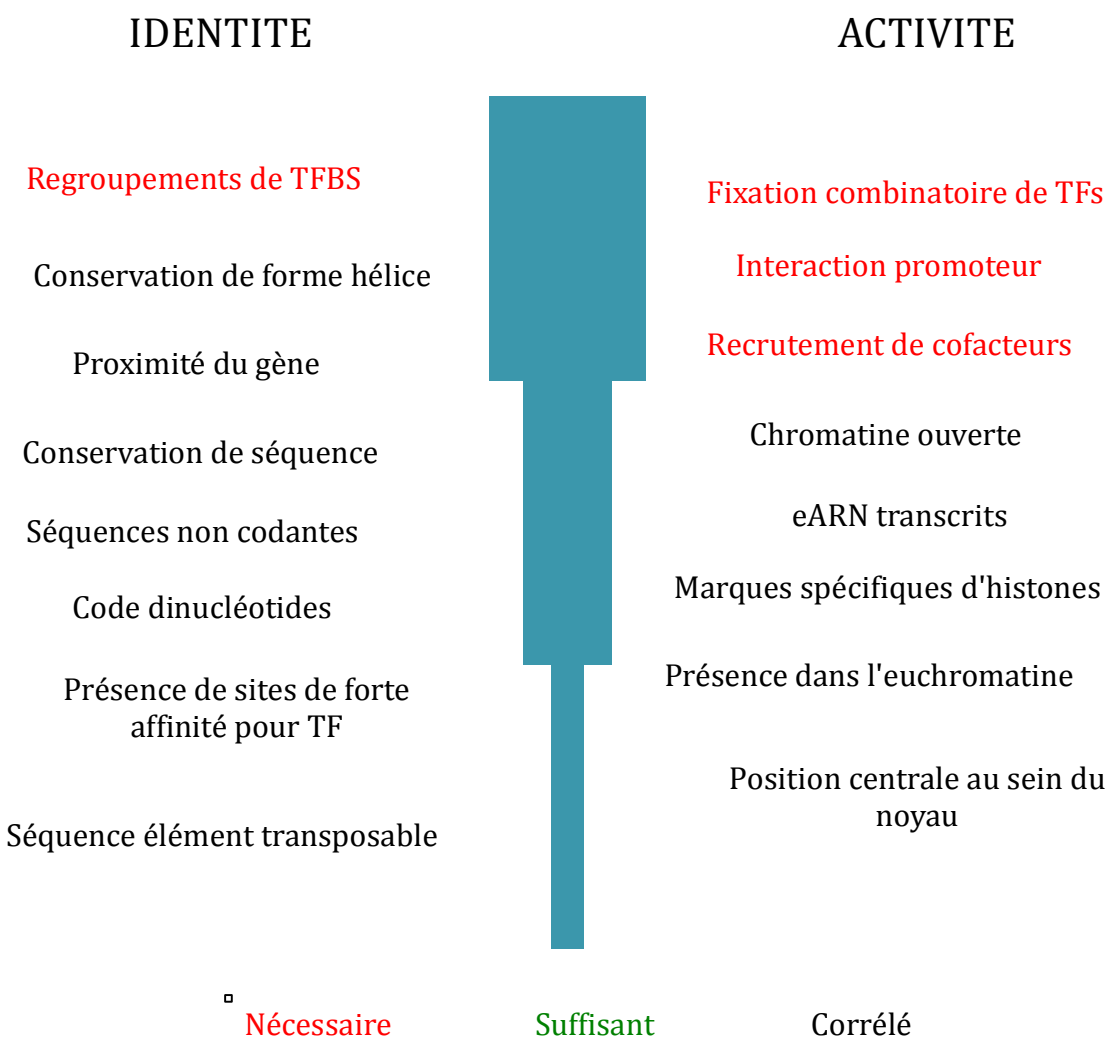


Figure I.10 : Règles générales permettant de détecter l'activité d'un enhancer (expérimental) ou d'inférer son identité (computationnel)

Ainsi à partir d'un enhancer naturel dont les sites de fixation sur l'ADN des facteurs de transcription sont parfaitement cartographiés, nous peinons toujours, malgré les progrès et les milliards d'euros investis, à construire un enhancer synthétique ayant un profil d'activité similaire à celui de l'élément naturel.

Plusieurs hypothèses pourraient expliquer ces difficultés. Il est premièrement possible que ce que nous appelons enhancer, au sens d'une classe homogène d'éléments, n'existe pas : il y aurait au sein de chaque espèce des classes multiples d'enhancers, chacune avec une logique propre, peut-être enrichie dans les éléments régulateurs de gènes impliqués dans une fonction cellulaire précise. Des éléments commencent à étayer cette hypothèse. Ainsi les éléments régulateurs agissant à grande distance semblent agir préférentiellement sur des gènes de contrôle du développement (Kikuta et al., 2007). Aussi, les enhancers développementaux et de gènes de ménage interagissent avec des promoteurs de différentes classes (Zabidi et al., 2014). Les enhancers les plus architecturalement contraints pourraient agir comme interrupteurs, alors que les moins contraints pourraient agir comme des potentiomètres précis permettant de répondre par exemple à des gradients de signalisation (Papatsenko & Levine, 2007). Des différences entre espèces apparaissent également, que ce soit au niveau des types de promoteurs que l'on y trouve (par exemple les CpG islands chez les vertébrés (Deaton & Bird, 2011) des signatures dinucléotidiques (Khoueiry et al., 2010), d'éléments transposables recrutés comme séquences *cis*-régulatrices (Teng, Firpi, & Tan, 2011), voire même de structure tridimensionnelle de la chromatine (Ea et al., 2015).

Le chantier reste donc immense. Il impliquera sans doute la multiplication et le raffinement des analyses descriptives et fonctionnelles à grande échelle, à l'échelle de la cellule unique (D. Wang & Bodovitz, 2010),analysées avec des méthodes statistiques et/ou d'apprentissage de plus en plus raffinées. Ces analyses à haut débit, néanmoins, établiront surtout des corrélations. Pour obtenir des relations de causalité, ces analyses devront être complétées par des analyses fonctionnelles détaillées sur un petit nombre d'éléments bien choisis. Mon travail de thèse a consisté à conduire une telle analyse sur l'un des enhancers les plus simples et les mieux compris chez les métazoaires, l'élément a chez l'ascidie *Ciona intestinalis*.

II. *Ciona intestinalis*, organisme modèle pour l'étude des séquences *cis*-régulatrices

Les ascidies ont émergé en tant qu'organismes modèles pour l'embryologie expérimentale grâce aux travaux de Laurent Chabry (1887) et Edwin Conklin (1905) (Chabry, 1887; Conklin, 1905), puis à la fin du XXème siècle, grâce aux brillantes applications des techniques de biologie moléculaire sur ces animaux par l'équipe de Nori Satoh, dont les trente années de recherche sur les ascidies ont été récompensées par la médaille Kovalevski, un prix international de la Société des Naturalistes de St Petersbourg (Kuratani, Wada, Kusakabe, & Agata, 2006).

Dans cette partie, je présenterai les différentes caractéristiques qui font de *Ciona intestinalis* un organisme modèle pour l'étude de la biologie du développement en général et pour la compréhension de la logique des enhancers pendant le développement en particulier.

A. Ciona intestinalis, une ascidie-modèle

D'abord décrites par Aristote sous le nom de "théthyons", les ascidies doivent leur nom actuel à Job Baster, médecin et naturaliste hollandais du XVIIIème siècle : il est dérivé du mot grec « ασκός » (*askós*, outre) car les pêcheurs méditerranéens les appelaient « outres de mer » (Cuvier, 1817). Et pour cause, les ascidies sont des invertébrés marins sessiles qui se nourrissent en filtrant l'eau de mer, l'aspirant par un siphon et la rejetant par un autre.
Les ascidies sont hermaphrodites et le corps des individus adultes est couvert d'une tunique de cellulose plus ou moins épaisse dont dérive le nom « tunicier » de leur sous-embranchement. Les tuniciers sont les seuls animaux capables de produire de la cellulose, un polysaccharide que l'on retrouve dans les matrices extracellulaires des plantes, des algues et des bactéries, et qui a sans doute été acquis par les tuniciers par un transfert horizontal de gène (Nakashima, Yamada, Satou, Azuma, & Satoh, 2004). Au sein des tuniciers, on distingue trois classes : les ascidies (solitaires ou coloniales), les larvacés et les thaliacés, qui se distinguent par d'importantes disparités de morphologie et de mode de vie, les deux dernières ayant un mode de vie pélagique. Les ascidies constituent le groupe le plus grand et le plus diversifié des tuniciers : près de 3000 espèces d'ascidies sont documentées parmi lesquelles plusieurs organismes modèles désormais bien établis comme *Ciona intestinalis*, *Halocynthia roretzi* et *Botryllus schlosseri* et d'autre encore émergents comme *Phallusia mammillata* ou *Molgula occulta* et *oculata* (Shenkar & Swalla, 2011) (figure II.1).

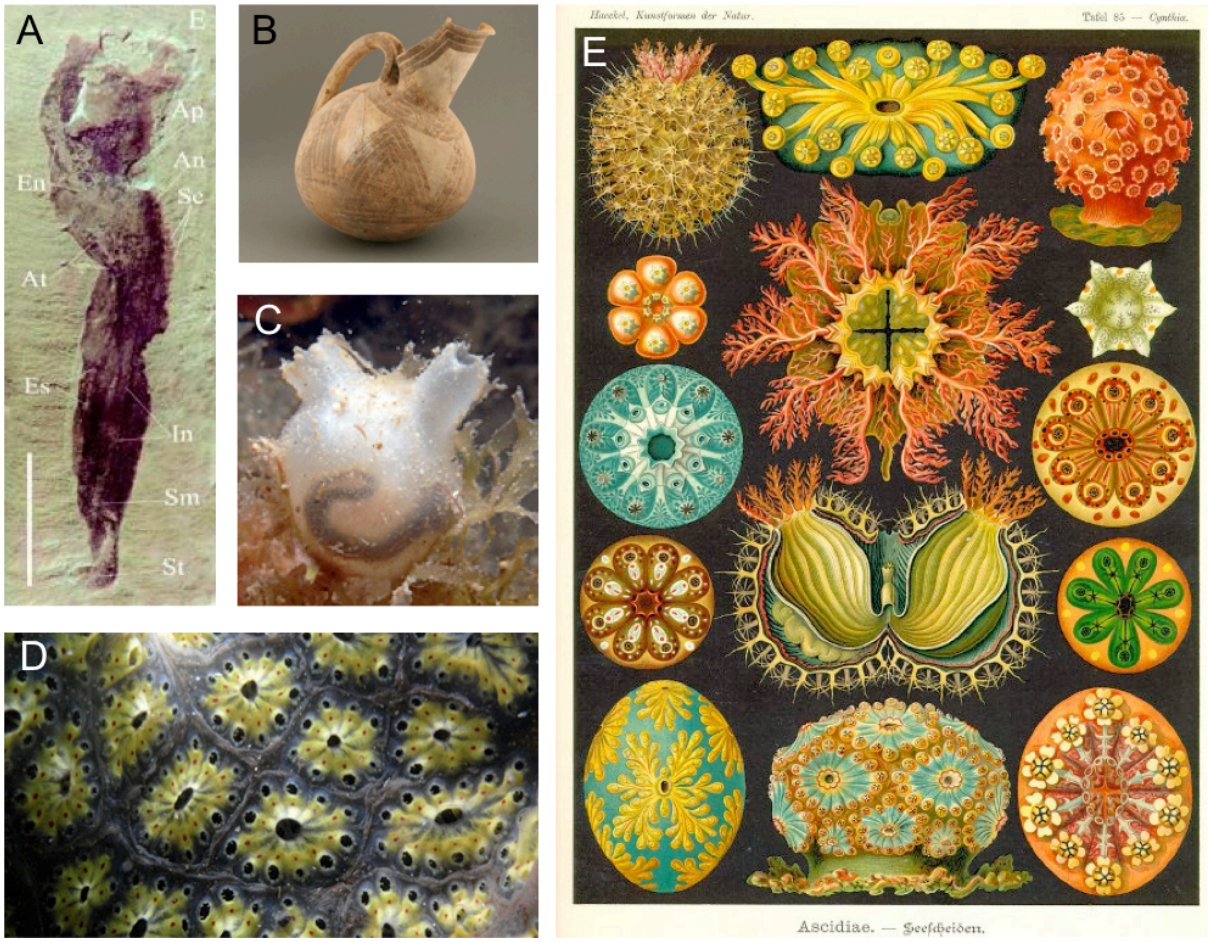


Figure II.1 Histoire et diversité des ascidies. A. *Shankouclava shankouense,* hypothétique fossile de tunicier du cambrien inférieur, trouvé en Chine et daté d'environ 543 millions d'années (Chen et al.,2003) **B**. ασκός (askos), outre grecque (2000-1850 B.C) **C**. *Molgula manhattensis,* une ascidie solitaire. **D.** *Botryllus schlosseri*, une ascidie coloniale **E.** Gravures de différentes espèces d'ascidies par le biologiste allemand Ernst Haeckel dans son ouvrage *Kunstformen der Natur* (L'art des formes de la nature), 1904. Crédits : *Molgula, botryllus* : Wilfried Bay-Nouailhat, http://www.mer-littoral.org ; Askos : musée archéologique de Kalamata, Grèce

Au dix-huitième siècle, Linné identifie et nomme l'espèce *Ciona intestinalis*, dont voici la description sommaire : « laevis alba membranacea » (formé d'une membrane lisse et blanche) (Linné & Salvius, 1758). C'est aujourd'hui l'espèce d'ascidie la plus étudiée en biologie du développement et celle qui a servi à toutes mes expériences pendant ma thèse. Cette espèce a un cycle de vie de deux à trois mois et se trouve dans toutes les mers tempérées du globe (Shenkar & Swalla, 2011). *Ciona intestinalis* est une espèce saisonnière dont la période de reproduction dépend vraisemblablement de la température de la mer : celles sur lesquelles j'ai travaillé, provenant de Roscoff, en Bretagne, sont généralement fertiles d'avril à novembre.

1. Phylogénie

Animalia, Eumetazoa, Bilateria, Deuterostomia, Chordata, Olfactores, Tunicata, Ascidiacea, Enterogona, Phlebobranchia, Cionidae

En 1789, Linné classe les ascidies au sein des mollusques (VERMES, MOLLUSCA) : comme Job Baster et Peter Simon Pallas, zoologiste et botaniste allemand, il voit une analogie entre l'huitre et les ascidies. A partir de la dissection d'ascidies conservées dans de « l'esprit de vin», Georges Cuvier confirma qu'il s'agissait de mollusques, s'apparentant plus particulièrement aux bivalves car dépourvues « d'organes de locomotion », « renfermées dans un sac à deux tuyaux » et ayant « une bouche dans le fond du sac à l'opposite du tuyau par lequel l'eau de mer pénètre, et de manière à ce que cette eau ne puisse y arriver qu'après y avoir arrosé les sacs branchiaux » (Cuvier, 1817).
Mais les différences de structure au niveau de ces « branchies » les apparentent plus étroitement aux Salpes qui ont le même type d'organisation interne. Ainsi, Lamarck définit le groupe des Tuniciers qui comprend les « tuniciers réunis ou botryllaires » (botrylles et pyrosomes) et les « tuniciers libres ou ascidiens » (salpes et ascidies), mais doute qu'il soit apparenté aux mollusques bivalves, car leurs organisations respectives ont très peu de similitudes (Lamarck, 1816).

C'est Alexandre Kowalevsky qui remarquera en 1866 des similarités entre les larves d'ascidies et les embryons de vertébrés (Kowalevsky, 1866). L'observation d'un tube neural dorsal et d'une notochorde est la preuve que les ascidies, comme les céphalocordés et les vertébrés, sont des chordés. De par leurs similarités morphologiques avec les vertébrés et leur plus grande complexité apparente, les céphalochordés sont considérés comme plus proches des vertébrés. Pendant près de 150 ans, les tuniciers vont conserver cette position dans la classification, à la base des chordés.

L'avènement de la génomique permet les études de phylogénie moléculaire et de montrer, en 2006, que les tuniciers sont le groupe-frère des vertébrés, et que tous deux forment le clade des Olfactores (Delsuc, Brinkmann, Chourrout, & Philippe, 2006) (figure II.2). Cette classification moléculaire est renforcée par des observations anatomiques : des cellules comparables aux cellules de la crête neurale des vertébrés ont été observées dans les larves de tuniciers mais pas dans celles des céphalocordés (Abitua, Wagner, Navarrete, & Levine, 2012). De par cette

position stratégique de groupe-frère des vertébrés, les ascidies seraient de bons modèles pour étudier des processus de développement fondamentaux et l'origine des vertébrés. Cependant, les tuniciers ne sont pas des "vertébrés simplifiés" car ils sont extrêmement dérivés, ce dont je discuterai dans les paragraphes sur leur génome et leur développement.

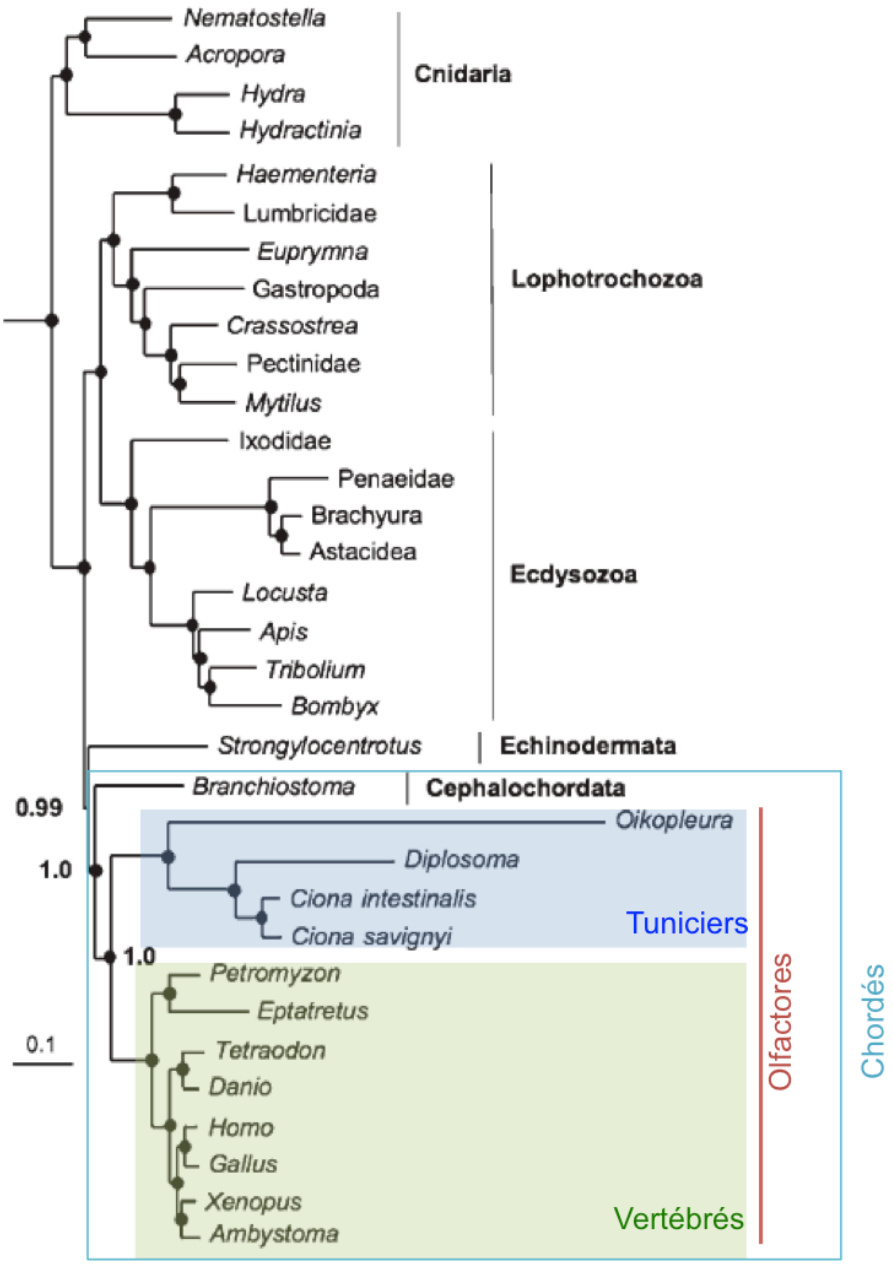


Figure II.2. Arbre phylogénétique montrant que les tuniciers sont le groupe-frère des vertébrés (Adapté de Delsuc, Tsagkogeorga, Lartillot, & Philippe, 2008))

Les relations phylogénétiques au sein des tuniciers ne sont pas encore complétement élucidées et il semblerait que les ascidies ne soient pas un groupe monophylétique (Tsagkogeorga et al., 2009). En effet, si les larvacés sont généralement considérés comme les tuniciers les plus basaux, les thaliacés semblent s'insérer au sein des ascidies (Lemaire, 2011)(Figure II.3).

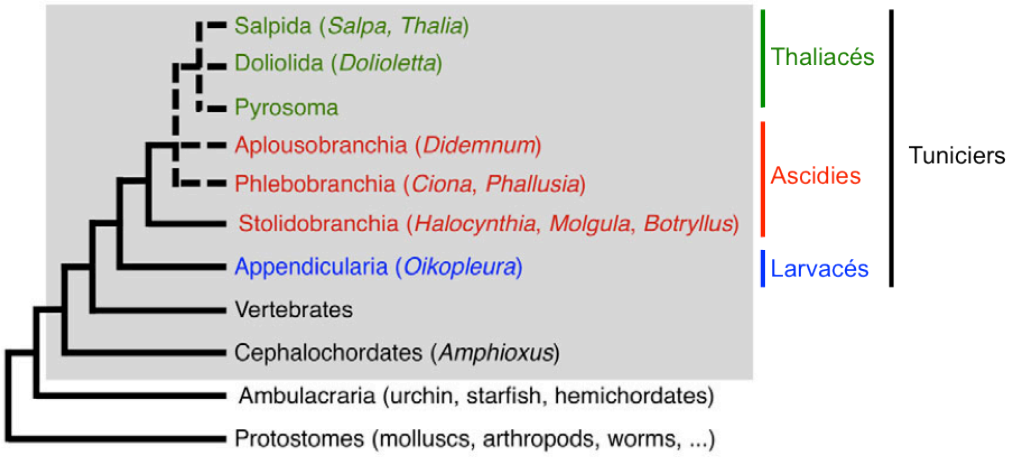


Figure II.3 Cladogramme illustrant les relations phylogénétiques au sein des tuniciers. Les rectangle grisé correspond à l'embranchement des chordés. Le sous-embranchement des tuniciers comprend les thaliacés (vert), les ascidies (rouge) et les larvacés (bleu). Le nom des principaux genres représentant chaque classe est entre parenthèses. Les relations phylogénétiques entre thaliacés et ascidies n'étant pas complétement résolues, elles sont représentées en pointillés. (Adapté de Lemaire, 2011)

Récemment, la subdivision de *Ciona intestinalis* en deux espèces a été proposée (Pennati et al., 2015) : *Ciona intestinalis «* anciennement de type B », sur laquelle j'ai travaillé demeure *Ciona intestinalis* (Linné & Salvius, 1758). *Ciona intestinalis «* anciennement de type A » devient *Ciona robusta*, décrite par Hoshinoa et Tokioka en 1967 (Hoshino & Tokioka, 1967). Par simplicité, et notamment pour pouvoir référer aux articles sur les génomes, je conserverai l'appellation « type A / type B » dans le reste de ma thèse.

2. Génomique

Le premier génome de tunicier, celui de *Ciona intestinalis* a été séquencé en 2002. C'est un génome de petite taille (160Mb environ), riche en AT (65%) et très compact qui comporte quatorze chromosomes. Sa densité en gènes (1 gène/7,5kb) pourrait être due à la faible présence d'éléments transposables, avec pour conséquence des séquences intergéniques et introniques compactes (Denoeud et al., 2010). Les ascidies n'ayant pas subi les deux phases de

duplication génomique typiques des vertébrés, elles possèdent donc moins de paralogues. (Dehal et al., 2002).

Comparés à l'hypothétique génome ancestral des chordés, les génomes des tuniciers sont petits et simplifiés par la perte de gènes et d'introns : ainsi plus de 200 gènes présents chez l'amphioxus et les vertébrés sont absents du génome de *Ciona intestinalis* (Putnam et al., 2008).

	Drosophila melanogaster	*Ciona intestinalis*	*Homo sapiens*
Taille du génome	120Mb	160Mb	3Gb
Nombre de paires de chromosomes	4	14	23
Duplication totale de génome	0	0	2
Nombre de gènes codants	14000	16000	30000
Nombre de facteurs de transcription	500	500	1300
Densité en gènes #	1 gène / 9kb	1 gène / 7,5kb	1 gène / 100kb
Nombre d'exons par gènes #	5	6,8	8,8
Diversité nucléotidique autosomale (% par site) *	1	1 à 4	0,1
GC%	55%	35% °	42%

Tableau II.1 Comparaison des caractéristiques principales du génome de *Ciona intestinalis* avec celles des génomes de drosophile et d'Homme (°Dehal 2002, #Irvine 2013, * Leffler 2012)

Le génome de *Ciona intestinalis* est très dynamique et l'ordre de ses gènes a connu de nombreux réarrangements, non seulement par rapport aux vertébrés, mais aussi par rapport à d'autres espèces du genre *Ciona* : la synténie entre *Ciona intestinalis* et *Ciona savignyi* est limitée à des blocs de moins d'un méga-base (Hill et al., 2008). Une forte diversité génétique se retrouve également au sein de l'espèce où l'on observe 1,2% et 4,6% de polymorphisme entre les deux allèles de l'individu séquencé chez *Ciona intestinalis* et *Ciona savignyi* respectivement (Dehal et al., 2002; Vinson et al., 2005). La diversité génétique entre individus est très importante : 1 à 2% et 3 à 5 % de polymorphisme chez *Ciona intestinalis* et *Ciona*

savignyi respectivement (Abdul-Wajid, Veeman, Chiba, Turner, & Smith, 2014). Ainsi, les ciones, et plus particulièrement *Ciona savigny* et *Ciona intestinalis* sont parmi les eucaryotes ayant le plus haut niveau de diversité nucléotidique autosomale, ce qui ne peut s'expliquer par la seule taille de leurs populations (Leffler et al., 2012).

Aujourd'hui, plusieurs consortiums que notre équipe coordonne ont permis d'obtenir les génomes de plusieurs espèces d'ascidies solitaires (*Ciona intestinalis* types A et B, *Halocynthia aurantium, Halocynthia roretzi, Phallusia mammillata et Phallusia fumigata*), et des études de génétique des populations voient le jour grâce au séquençage des génomes de plusieurs individus d'une même espèce, fournissant un atlas précieux du polymorphisme de *Ciona intestinalis* (Abdul-Wajid et al., 2014).

3. Embryologie

Les ascidies ont une reproduction panmictique synchronisée entre les individus par le lever du jour : elles rejettent quasi-simultanément leurs gamètes dans la mer où la fécondation et le développement ont lieu. Chez la cione, l'autofécondation est rendue impossible par histoincompatibilité (Morgan, 1944). Son développement embryonnaire dure 18h à 18°C (Hotta et al., 2007), puis la larve éclot pour nager librement avant de se fixer sur un corps marin et de subir une métamorphose drastique au cours de laquelle l'essentiel de ses tissus larvaires disparaissent (Karaiskou, Swalla, Sasakura, & Chambon, 2015).

a) Grandes étapes du développement embryonnaire

Avant même l'entrée du spermatozoïde dans l'œuf, ce dernier a déjà une polarité animale/végétative établie. Après la fécondation, qui a le plus souvent lieu dans l'hémisphère animal, le cytoplasme est réorganisé pendant les deux phases de ségrégation ooplasmique qui assurent la localisation des déterminants maternels au cours du premier cycle mitotique (Sardet, Paix, Prodon, Dru, & Chenevert, 2007). L'œuf subit ensuite une série de divisions cellulaires stéréotypées très conservées entre différentes espèces d'ascidies, même distantes phylogénétiquement. Par exemple, *Ciona intestinalis* et *Halocynthia roretzi*, distantes d'environ 400 millions d'années ont des plans de clivage quasiment identiques. Le plan sagittal, qui est le seul plan de symétrie de l'embryon est déterminé dès la première division

cellulaire. Le second clivage détermine l'axe antéro-postérieur. Enfin, la troisième division sépare les pôles animal et végétatif.

Ainsi, dès le stage 8-cellules, quatre grands lignages, a, b, A et B sont distingués, correspondant respectivement aux lignages antérieur-animal, postérieur-animal, antérieur-végétatif et postérieur-végétatif.

L'embryon subit ensuite une succession de divisions cellulaires et forme une blastula de 32 à 76 cellules. La gastrulation commence au stade 112 cellules chez *Ciona* (110 chez *Hallocynthia*) et se fait par invagination au pôle végétatif des cellules mésendodermiques qui produiront les organes internes. Puis la plaque neurale dorsale se referme sur elle-même au cours de la neurulation pour former le tube neural. Enfin, la queue s'allonge progressivement du stade bourgeon caudal jusqu'à l'éclosion de la larve qui ne comprend que 2600 cellules environ (Figure II.4).

b) Le développement stéréotypé de *Ciona intestinalis*

Les embryons d'ascidies se développent selon des plans de clivages invariants, au moins jusqu'à la gastrulation. Le faible nombre et les divisions stéréotypées des cellules des embryons d'ascidies ont permis de reconstruire le lignage cellulaire au cours des premiers stades développementaux (Chabry, 1887; Conklin, 1905) puis jusqu'au stade de la jeune gastrula où la grande majorité des cellules voit leur potentiel de différentiation restreint à un seul tissu (Nishida, 1987). Une fois les trois grands axes de l'embryons établis, au stade 8-cellules, chaque cellule est identifiée par une formule unique du type Xm.n(*), X étant la lettre correspondant au lignage dont elle est issue (A, a, B ou b) (Figure II.5). Le premier nombre, « m », désigne le rang du cycle cellulaire après la fécondation. Le second nombre, « n », identifie chaque cellule par rapport à sa position à l'issue de la division cellulaire de sa mère. La cellule Xm.n produira ainsi en se divisant les cellules X(m+1).(2n-1) et X(m+1).(2n), la surface externe de X(m+1).(2n-1) étant plus proche en coordonnées polaires du pôle végétatif. Enfin, l'astérisque permet de différencier les cellules symétriques de l'embryon.

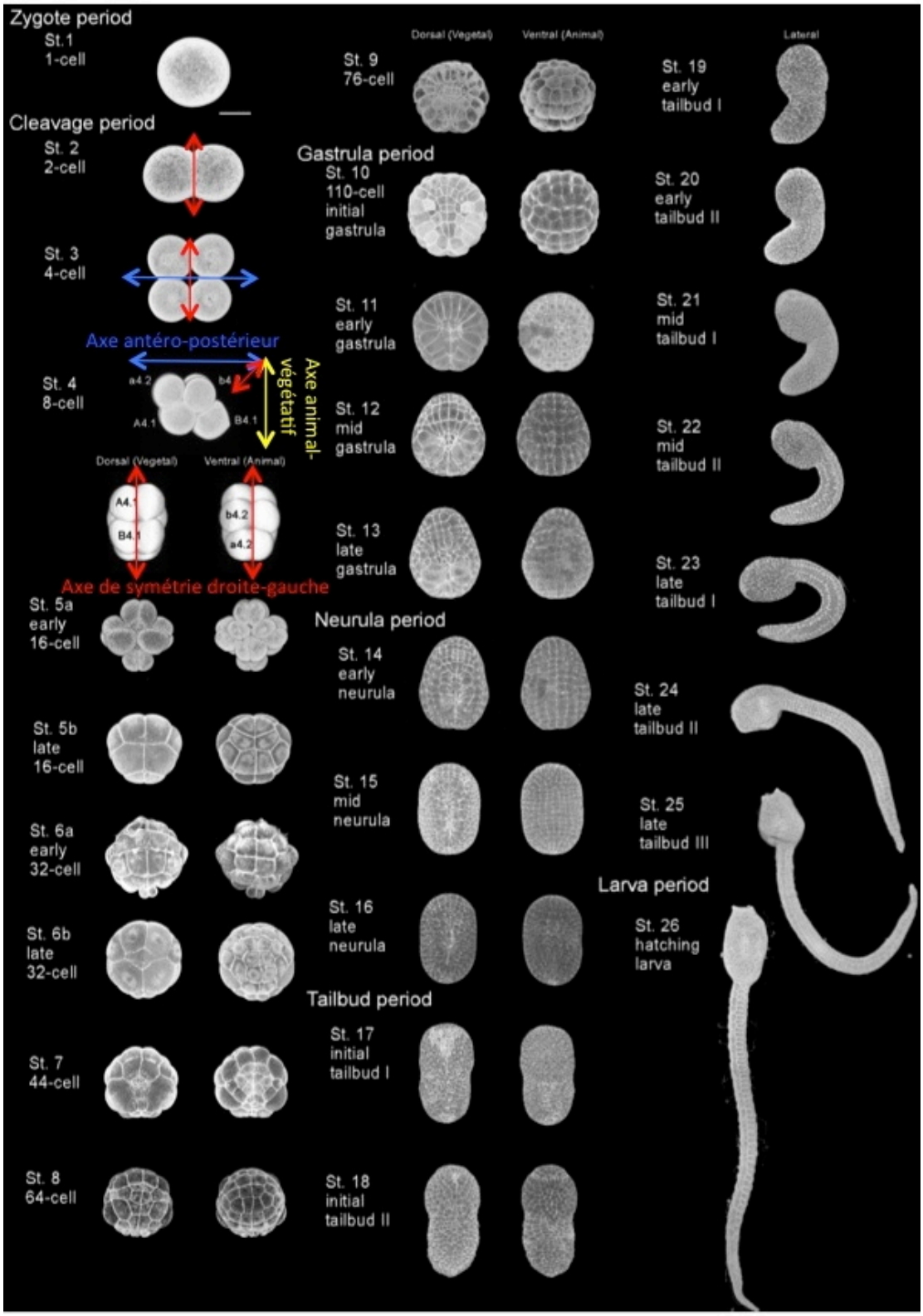


Figure II.4. Le développement des embryons de *Ciona intestinalis*. Images reconstituées en 3D des embryons de la cellule-œuf à la larve. Les trois axes de symétrie sont représentés en couleur jusqu'au stade huit cellules : axe sagittal en rouge, axe antério-postérieur en bleu, axe animal-végétatif en jaune. (Adapté de Hotta et al., 2007)

c) Spécification des destins cellulaires

Les résultats de Chabry et Conklin ont d'abord laissé penser que le développement des ascidies est mosaïque, contrôlé localement par l'héritage au cours des divisions successives de déterminants maternels localisés. C'est en effet le cas, par exemple, de la majorité des cellules musculaires de la queue pour lesquelles un réseau de gènes autonome à la cellule (cell-autonomous) a été reconstruit (Imai, Levine, Satoh, & Satou, 2006). Mais, à partir du stade 16 cellules la majorité des cellules n'ont pas un programme de division autonome et répondent à des signaux à courte distance secrétés par leurs voisines (Lemaire, 2009).
Certaines cellules sont restreintes dès le stade 32 cellules à ne former qu'un seul type cellulaire, c'est le cas des blastomères A6.1 et a6.6 par exemple, qui ne formeront respectivement que de l'endoderme et de l'épiderme. Au stade de la jeune gastrula, la plupart des cellules (102 sur 112) sont restreintes à ne former qu'un seul type de tissu, la larve comprenant moins de vingt tissus majeurs. Cette analyse a permis d'établir une carte des destins cellulaires présomptifs (Nishida, 1987) (Figure II.5).

Chez la cione, la larve est également très simple : elle comprend environ 2600 cellules. Ainsi, K. Hotta a pu reconstruire manuellement un embryon au stade bourgeon caudal (Hotta et al., 2007).

Plusieurs processus développementaux ont été étudiés au niveau transcriptionnel, comme la formation de la notochorde (Kawai, Takahashi, Nishida, & Yokosawa, 2005), de l'endoderme (Satou, Imai, & Satoh, 2001), du cœur (Christiaen, Stolfi, Davidson, & Levine, 2009; Christiaen et al., 2008) ou de l'induction neurale (Bertrand, Hudson, Caillol, Popovici, & Lemaire, 2003; Hudson & Lemaire, 2001), que je décrirai en détail dans la seconde partie.

4. Intérêts pratiques du modèle :

La collecte des gamètes de *Ciona intestinalis* est rapide et simple, et la fécondation *in vitro* permet d'obtenir plusieurs milliers d'embryons se développant de manière synchrone et stéréotypée. Le développement rapide des embryons de *Ciona intestinalis* (à 18°C, la gastrulation commence moins de 5h après la fécondation) et le petit nombre de cellules qui

constituent l'embryon (environ 2600 pour le têtard) permettent d'obtenir des résultats rapides, précis et un niveau de résolution cellulaire.

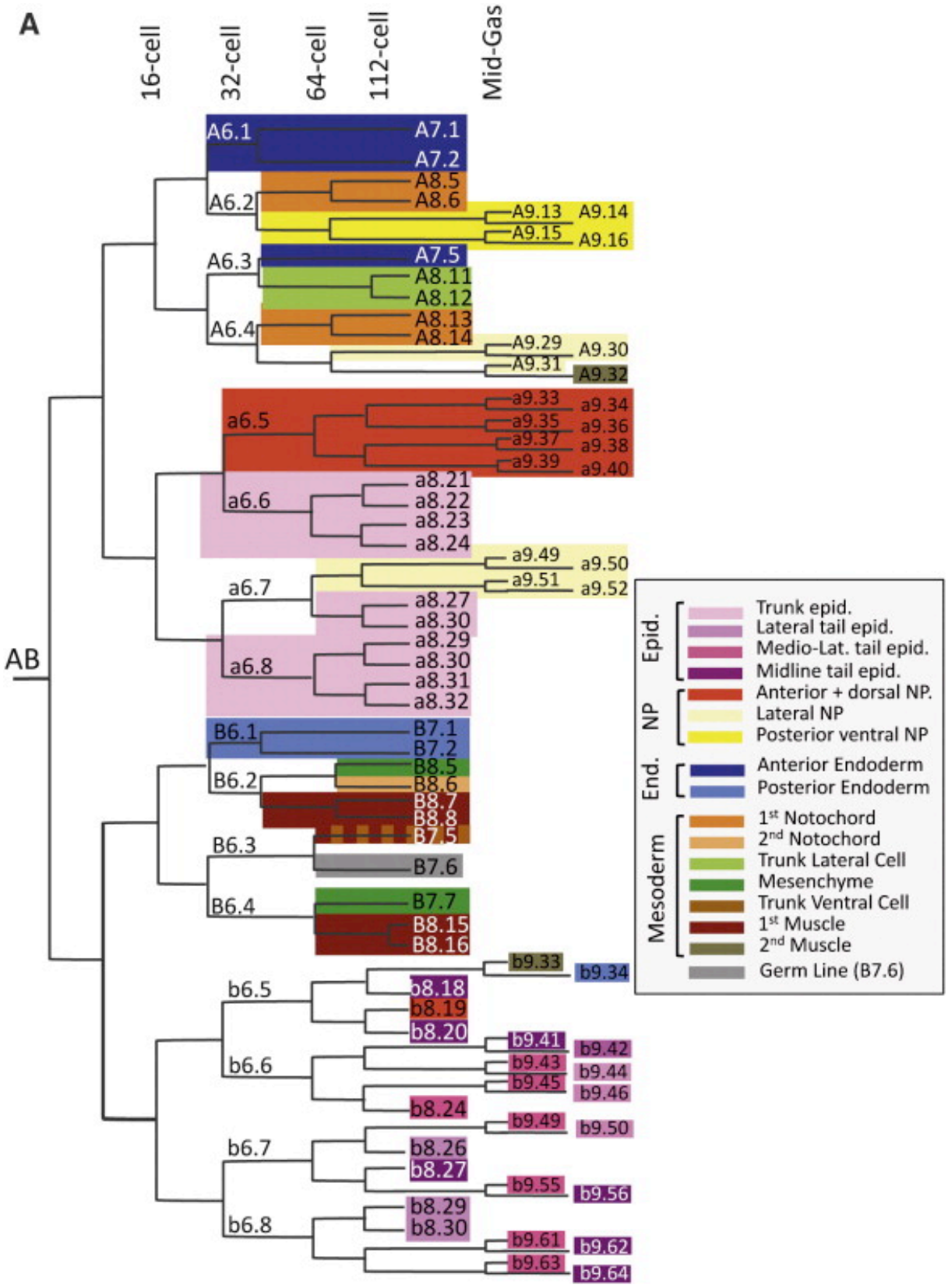


Figure II.5 : Arbre représentant le lignage cellulaire et les territoires embryonnaires présomptifs jusqu'au stade gastrula chez *Ciona intestinalis* (Lemaire, 2009).

Ces embryons peuvent être microinjectés avec des morpholinos, des ARNs messagers, mais aussi des TALENS ou des constructions CRISPR-Cas9, ou soumis à des inhibiteurs pharmacologiques de voies de transductions de signaux (Christiaen, Wagner, Shi, & Levine, 2009b; Stolfi, Gandhi, Salek, & Christiaen, 2014; Treen et al., 2014).

Mais c'est grâce à l'électroporation des zygotes que les ciones sont devenues des organismes modèles (Christiaen, Wagner, Shi, & Levine, 2009a; Corbo, Levine, & Zeller, 1997). Cette technique permet de générer des centaines d'embryons transgéniques en routine, et est particulièrement appropriée à l'étude des séquences *cis*-régulatrices : l'électroporation d'un gène rapporteur couplé à celles-ci permet en effet de visualiser l'activité d'un enhancer, et ce avec une résolution cellulaire grâce au lignage invariant.

Les embryons d'ascidies sont également adaptés pour des études moléculaires à moyenne échelle, grâce aux technique d'hybridation *in situ* (Imai, Hino, Yagi, Satoh, & Satou, 2004; Miwata et al., 2006; Satou, Takatori, et al., 2001) qui permettent de visualiser la localisation des ARNs, ou de microinjection de morpholinos antisens qui permettent d'étudier le rôle d'un gène par perte de fonction (Yamada et al., 2003). La combinaison de ces deux techniques a permis de déterminer les relations épistatiques entre plus d'une cinquantaine de gènes régulateurs précoces et la reconstruction d'un premier brouillon du réseau de régulation des gènes exprimés au cours des premiers stades embryonnaires de *Ciona intestinalis* (Imai et al., 2006).

Ainsi *Ciona intestinalis* est un modèle efficace pour étudier non seulement les liens entre les réseaux de gènes et la morphogénèse, mais aussi la logique *cis*-régulatrice elle-même.

B. Les régions *cis*-régulatrices chez les ascidies

Au milieu des années 90 parut la première étude d'un élément régulateur chez l'ascidie : l'équipe de Nori Satoh utilisa la technique de microinjection pour étudier la régulation du gène *HrMA4* exprimé spécifiquement dans les cellules du muscle dès le stade gastrula chez *Halocynthia roretzi* (Hikosaka Akira, Takehiro Kusakabe, 1994). Puis, en 1997, J. Corbo identifia un enhancer minimal dirigeant l'expression de *Brachyury* dans la notochorde en utilisant l'électroporation de constructions rapporteurs (Corbo et al., 1997). Grâce à cette technique, plusieurs centaines de régions *cis*-régulatrices ont été identifiées et caractérisées chez les ascidies, en majorité chez *Ciona intestinalis* (Irvine, 2013; Wang & Christiaen, 2012), ce qui a permis de dégager certaines propriétés de la *cis*-régulation chez les ascidies.

1. Méthodes d'identification des enhancers chez les ascidies

a) 3kb en amont du site d'initiation de la transcription

La majorité des enhancers identifiés chez la cione sont situés à moins de 1,5kb en amont du TSS (Satoh, Satou, Davidson, & Levine, 2003). Ainsi la méthode adoptée par Corbo en 1997 a constitué une référence pour l'identification des enhancers (Corbo et al., 1997). Il s'agit de cloner en amont d'un gène rapporteur les 3kb situés en amont du TSS, puis de raffiner la compréhension de la logique régulatrice sous-jacente par des délétions successives (Figure II.6). Cette méthode a certes permis d'identifier un grand nombre d'enhancers très proches du TSS, mais a peut-être introduit un biais dans l'identification de régions régulatrices plus proches du TSS chez les ascidies que chez les autres métazoaires. Des expériences de type ChIP-seq, ATAC-seq, ou 3C et dérivés permettraient d'identifier d'éventuels enhancers plus distaux.

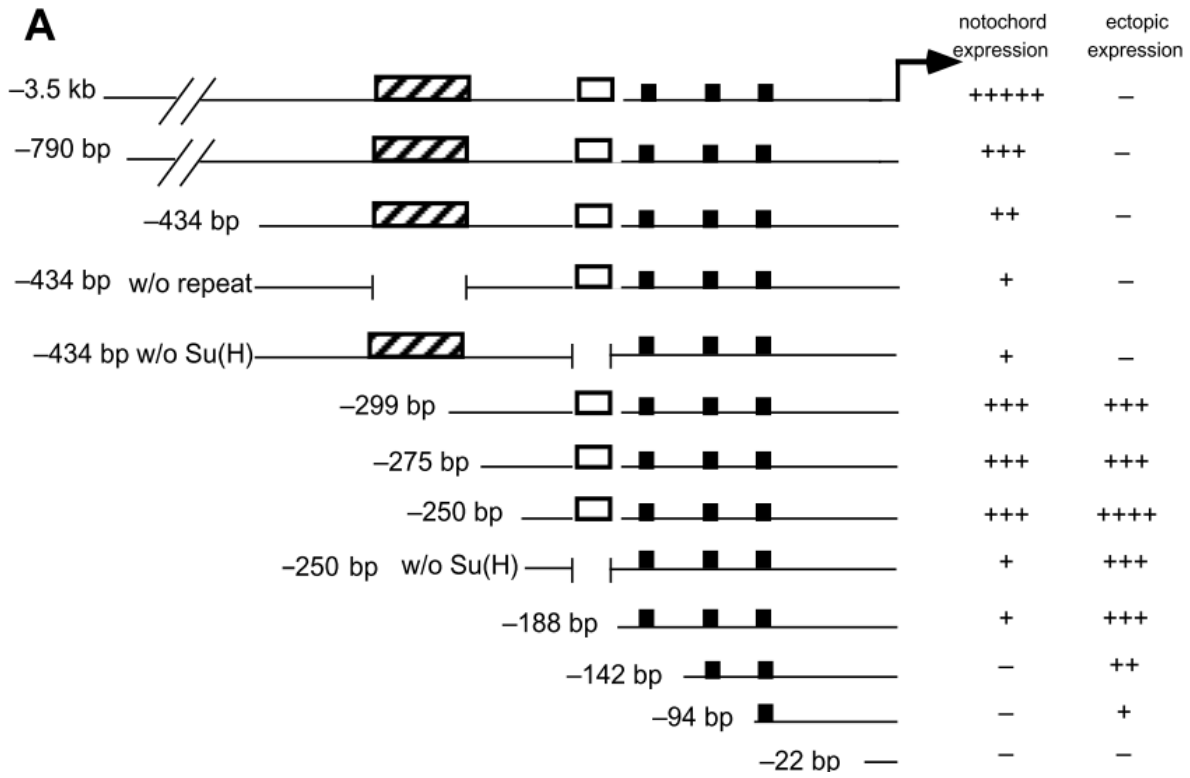


Figure II.6 : Les 3,5kb en amont du site d'initiation de la transcription de *Brachyury* contiennent plusieurs éléments *cis*-régulateurs. La dissection de cette région en testant l'activité des régions schématisées ci-dessus par électroporation a permis d'identifier une région importante pour réprimer l'expression de *Brachiury* hors de la notochorde (région autour du rectangle hachuré), une région importante pour son expression dans la notochorde (rectangle blanc) et trois TFBS qui semblent importants pour son expression dans le muscle et le mésenchyme (carrés noirs). Cette approche expérimentale a permis d'identifier et de caractériser un très grand nombre d'enhancers. (Corbo et al., 1997).

b) Clonage aléatoire

Avant 2002, en l'absence de génome annoté, une des seules possibilités pour tester l'activité *cis*-régulatrice de fragments génomiques était le clonage aléatoire. Cette technique "à l'aveugle" consiste à cloner en amont d'un promoteur basal et d'un gène rapporteur un fragment d'ADN génomique de 1 à 3kb choisi au hasard. De cette façon, sur 138 fragments de 1,7kb en moyenne, 11 enhancers actifs au cours du développement embryonnaire ont pu être identifiés par Harafuji *et al*., suggérant une fréquence moyenne d'un enhancer actif pendant cette période tous les 20-40kb chez *Ciona intestinalis* (Harafuji, Keys, & Levine, 2002). Dans une étude similaire, centrée sur les gènes *Hox*, Keys et ses collègues identifièrent 22 enhancers sur les 222 fragments testés (Keys et al., 2005).

Ces approches non biaisées ont un faible taux de réussite : 10% ou moins des séquences testées sont des enhancers.

c) Empreinte phylogénétique

De façon fortuite, le génome de la seconde espèce d'ascidie séquencée, *Ciona savignyi* (Small, Brudno, Hill, & Sidow, 2007) est à une distance évolutive de celui de *Ciona intestinalis* qui permet d'utiliser la conservation de séquence pour identifier des enhancers. En effet, le signal de conservation dans le non-codant est existant et non-saturé, ce qui permet de distinguer des régions non-codantes faiblement conservées (et probablement non fonctionnelles) et des régions non-codantes fortement conservées (et probablement fonctionnelles). Plusieurs enhancers ont pu être identifiés de cette manière dans plusieurs laboratoires, par exemple ceux des gènes *Pitx* (Figure II.7) (Christiaen, Bourrat, & Joly, 2005), *sFRP1/5* (Lamy, Rothbächer, Caillol, & Lemaire, 2006), *FoxF* (Beh, Shi, Levine, Davidson, & Christiaen, 2007), *ROR* (Auger et al., 2009) ou encore *Lef/Tcf* (Squarzoni, Parveen, Zanetti, Ristoratore, & Spagnuolo, 2011) (Pour review, Wang & Christiaen, 2012). Les enhancers identifiés ainsi ont généralement une taille de quelques centaines de paires de bases et il existe une relation quantitative significative entre la fonctionnalité et le degré de conservation des éléments non-codants (Johnson, Davidson, Brown, Smith, & Sidow, 2004). Il existe toutefois des cas particuliers qui échappent à cette règle.

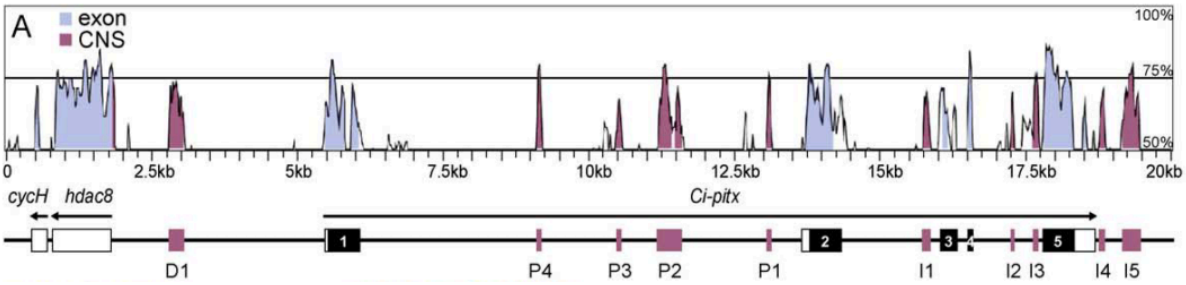


Figure II.7 : Les séquences des éléments *cis*-régulateurs de *Pitx* sont conservées entre *Ciona intestinalis* et *Ciona savignyi.* En haut apparait le taux de conservation de séquences entre les deux espèces. En bas est schématisée la région autour de *Pitx* : les rectangles noirs et blancs représentent les exons (dont la conservation apparaît en bleu), les rectangles mauves représentent les éléments conservés non codants (CNS, conservation en mauve). D1 et P2 sont des enhancers qui activent l'expression de *Pitx* dans le stomodéum, P3 est un module co-activateur, et les régions I1-5 pourraient être responsables de l'expression asymétrique de *Pitx.* (Christiaen et al., 2005).

En effet, l'identification d'enhancers par empreinte phylogénétique entre les deux ciones n'est pas possible pour tous les gènes. En effet, pour certains gènes, le niveau de conservation des régions non codantes avoisinantes n'est pas suffisant pour identifier clairement des séquences ayant potentiellement une activité enhancer. C'est le cas de certains enhancers de *FoxA-a* (Irvine, 2013). Le récent séquençage du génome de *Ciona intestinalis* type B (résultats non publiés du laboratoire), plus proche du type A que celui de *Ciona savignyi* pourrait permettre d'identifier des enhancers dans de telles régions. De plus, toutes les régions non-codantes conservées ne sont pas des enhancers embryonnaires. Des régions conservées n'ayant pas d'activité enhancer à un certain stade peuvent être actives plus tard pendant le développement ou même pendant ou après la métamorphose. Ces régions pourraient également être des éléments *cis*-régulateurs fonctionnels tels que des silencers ou des insulateurs dont la validation expérimentale est moins évidente. Enfin, certaines de ces régions pourraient constituer d'autres types d'éléments, comme des gènes non-codants.

d) Clusters de sites de fixation pour des facteurs de transcription

Une autre méthode utilisée chez les ascidies pour identifier des enhancers est la recherche de clusters de sites de fixation pour des facteurs de transcription dans les régions non-codantes de gènes co-exprimés. La logique sous-jacente est que des gènes exprimés dans les mêmes cellules ou tissus seraient régulés par les mêmes facteurs de transcription, et leurs enhancers contiendraient donc les mêmes sites de fixation pour ces facteurs de transcription (Haeussler, Jaszczyszyn, Christiaen, & Joly, 2010).

Kusakabe et ses collègues ont recherché les motifs sur-représentés dans les régions non-codantes d'une cinquantaine de gènes exprimés dans le muscle des larves de *Ciona intestinalis*. Ils ont ainsi identifié dans ces régions un enrichissement en sites de fixation pour les facteurs de transcription CREB et MRF (Kusakabe, Yoshida, Ikeda, & Tsuda, 2004).

Il est intéressant de noter que la majorité des clusters de sites de fixation potentiels de facteurs de transcription ne confère pas d'activité enhancer à la région du génome où ils se trouvent. S'ils sont nécessaires, leur seule présence n'est pas suffisante pour l'activité d'un enhancer, ce que je discuterai plus tard.

e) Analyse de la structure chromatinienne

De part le faible nombre de cellules présentes dans les embryons d'ascidies et l'absence de culture cellulaire pour les invertébrés marins, les approches analysant la structure chromatinienne n'ont pas été autant exploitées que chez d'autres organismes modèles.
Une approche de ChIP-seq comparative chez *Ciona intestinalis* et *Phallusia mammillata* au stade de la jeune gastrula a été menée au sein de l'équipe. Les combinaisons de modifications d'histones ont permis une annotation fonctionnelle des éléments régulateurs à l'échelle du génome chez la jeune gastrula (Gineste en préparation). Toutefois, l'obtention de données de ChIP-Seq requiert pour l'instant une très grande quantité d'embryons et cette méthode n'a pas encore été utilisée de façon systématique chez les ascidies.
Le récent développement de techniques haut débit de caractérisation de la structure chromatinienne nécessitant moins de matériel biologique, telles que le ChIP-seq sur une cellule (Rotem et al., 2015) ou l'ATAC-seq, devrait permettre de caractériser de nouveaux enhancers potentiels.

f) Combinaison de ces méthodes

Plusieurs études ont combiné certaines des méthodes listées ci-dessus.
Par exemple, Johnson et ses collègues ont cherché dans le génome de *Ciona intestinalis* des regroupements de sites de fixation potentiels de facteurs de fixation pour MyoD, CRE et Tbx6 caractéristiques des enhancers actifs dans les précurseurs musculaires précédemment identifiés. Ils ont sélectionné 269 enhancers candidats contenant au moins un segment de séquence conservé avec *Ciona savignyi* (plus de 75% d'identité sur plus de 20pb). Environ 30% de ces enhancers sont actifs dans des précurseurs des cellules musculaires (Johnson et al., 2005).
Khoueiry et ses collègues ont associé la conservation de séquence à la présence d'une combinaison de TFBS. Sur 20 séquences testées, 4 nouveaux enhancers neuraux ont pu être caractérisés (Khoueiry et al., 2010).
Haeussler et ses collègues ont testé des séquences enrichies en motifs dupliqués, conservées entre *Ciona savignyi* et *Ciona intestinalis* et situées aux alentours de gènes exprimés dans le système nerveux central antérieur. Sur les 23 séquences testées, 10 sont actives dans le système nerveux central antérieur et leurs différences de profils d'expression s'expliquent par la présence de TFBS additionnels, les seuls pentamères GATTA n'étant pas suffisants pour activer la transcription dans le neuroectoderme antérieur (Haeussler et al., 2010).

Ces approches combinées ont permis d'améliorer l'efficacité de la recherche d'enhancers chez l'ascidie, mais leur "rendement" n'a pour l'instant jamais dépassé les 45% (Tableau II.2). Le nombre croissant d'enhancers validés expérimentalement a permis de dégager certaines caractéristiques des enhancers d'ascidies, qui peuvent évoluer avec l'utilisation de nouvelles méthodes pour les identifier.

Méthodes et filtres	**Fragments testés**	**Taille (kpb)**	**Positifs**	**Enhancers identifiés**	**Références**
Clonage aléatoire	138	1,7	11 (8,0%)	5 (3,6%)	(Harafuji et al., 2002)
Clonage aléatoire + gènes *Hox*	222	~3	29 (13,1%)	21 (9,5%)	(Keys et al., 2005)
Coexpression de gènes + Clusters de TFBS + conservation	23	<2	7 (30,4%)	6 (26,1%)	(Johnson et al., 2005)
Gènes co-exprimés + Clusters de TFBS + conservation	23	<1	10 (43,5%)	10 (43,5%)	(Haeussler et al., 2010)
Clusters de TFBS + conservation	20	<0,2	7 (35%)	4 (25%)	(Khoueiry et al., 2010)

Tableau II.2 : Projets d'identification d'enhancers à l'échelle du génome. Combiner différentes méthodes pour identifier des enhancers augmente l'efficacité des prédictions (d'après Wang & Christiaen, 2012).

2. Caractéristiques des enhancers chez les ascidies

Les ascidies ont de petits enhancers (au moins pour ce qui est des enhancers minimaux) de l'ordre de quelques centaines de paires de bases, ce qui est généralement observé chez d'autres métazoaires (Levine, 2010). La dissection plus précise de certains enhancers par délétions successives a permis d'identifier des enhancers dits "minimaux" de 50 à 200pb contenant généralement des sites de fixation pour deux à quatre facteurs de transcription différents, actifs dans un seul ou quelques lignages cellulaires. Ces enhancers sont donc petits et relativement simples, comparés à la complexité combinatoire de la région *cis*-régulatrice d'*Endo16* chez

l'oursin qui implique au moins 30 sites pour 13 facteurs de transcription ayant des rôles activateurs ou répresseurs (Yuh & Davidson, 1996).

Les séquences *cis*-régulatrices des ascidies semblent plus proches des gènes qu'elles régulent que chez d'autres espèces : les enhancers sont généralement présents dans les régions inter-géniques adjacentes et/ou dans le premier intron lorsque ce dernier est de grande taille. Ainsi, la majorité des enhancers connus sont localisés à moins de 3kb en amont du site d'initiation de la transcription (TSS), (Tableau II.3), mais seul un petit nombre d'enhancers a été pris en compte dans cette étude. Actuellement, on ne sait pas si des enhancers agissant à longue distance existent dans les génomes d'ascidies, les types d'expériences ayant permis de les identifier chez d'autres espèces n'ayant pas encore été réalisés chez les ascidies. La compacité du génome n'explique pas cette proximité gène-enhancer, car chez la drosophile qui a un génome de taille similaire, des enhancers ont été trouvés à 40kb du TSS (Kvon et al., 2014). Ainsi, les contraintes expliquant cette proximité seraient spécifiques aux ascidies : leurs génomes ayant été beaucoup réarrangés, il y a sans doute un avantage sélectif à avoir un enhancer localisé près du TSS (Irvine, 2013). Mais on peut également imaginer que c'est l'inverse : le génome a pu être extensivement réarrangé en l'absence d'éléments *cis*-régulateurs à distance, la probabilité de séparer un gène de ses régions régulatrices étant limitée.

	***Drosophila melanogaster* ***	***Ciona intestinalis* °**	***Mus musculus*°**
Enhancers les plus distants du TSS	28% > 20kb, Jusqu'à plus de 100kb	12kb en amont 7,5kb en aval	123kb en amont 93kb en aval
Distance médianes des enhancers au TSS	10kb	630pb en amont 2,4kb en aval	>4,1kb en amont >6,3kb en aval

Tableau II.3 Les enhancers des ascidies sont plus proches des gènes qu'ils régulent que ceux des vertébrés et des drosophiles. Comparaison entre la *Drosophila melanogaster, Ciona intestinalis et Mus musculus.* (***** Kvon 2014, ° Irvine 2013)

Les enhancers d'ascidies semblent avoir une architecture flexible et peu contrainte au sein et entre espèces, rendant difficile l'extraction des règles de la logique *cis*-régulatrice sous-jacente. Ainsi, au sein d'une espèce, des enhancers ayant le même profil d'activité, sont très variables en termes de composition, organisation et contribution à l'activité de l'enhancer de leurs TFBS

: la présence de sites de fixations pour au moins un des facteurs de transcription CRE, MyoD ou Tbx6 étant leur seul point commun (Figure II.8) (Brown, Johnson, & Sidow, 2007). Dans ce cas, l'organisation des enhancers orthologues est très similaire entre *Ciona intestinalis* et *Ciona savignyi*.

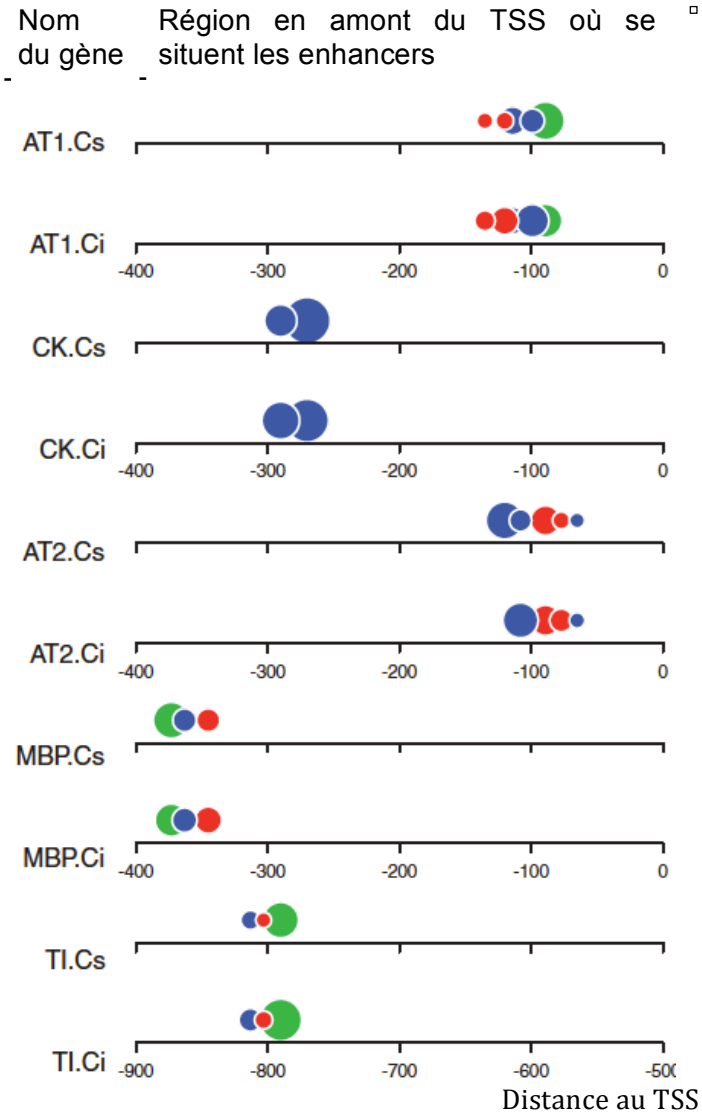


Figure II.8 : Les enhancers contrôlant l'activité de gènes dans le muscle chez *Ciona intestinalis* (Ci) et *Ciona savignyi* (Cs) présentent une grande variabilité structurale, mais la structure d'enhancers homologues est conservée entre les deux espèces. Les cercles représentent les sites de fixation pour CRE (rouge), MyoD (vert) et Tbx6 (bleu). La taille des cercles correspond à la contribution relative des TFBS à l'activité de l'enhancer. (Adapté de Brown 2007)

S'ils ne présentent aucune conservation de séquence, les enhancer responsables de l'expression de *Msxb* dans le lignage de b6.5 chez *Ciona intestinalis* et *Phallusia mammillata*, conservent leur activité lorsqu'ils sont testés chez l'autre espèce (Roure, Lemaire, & Darras, 2014).

Les mêmes observations ont été faites chez *Ciona intestinalis* et *Halocynthia roretzi* pour les enhancers responsables de l'expression conservée d'*Otx* dans les lignages neuraux a et b (Oda-Ishii, Bertrand, Matsuo, Lemaire, & Saiga, 2005). Tous deux possèdent des sites de fixation pour Ets1/2 et GATA4/5/6, nécessaires pour l'induction d'*Otx* par FGF (Bertrand et al., 2003), mais le nombre, la distance et l'orientation relative de ces TFBS diffèrent entre les deux enhancers orthologues (Figure II.9) (Oda-Ishii et al., 2005).

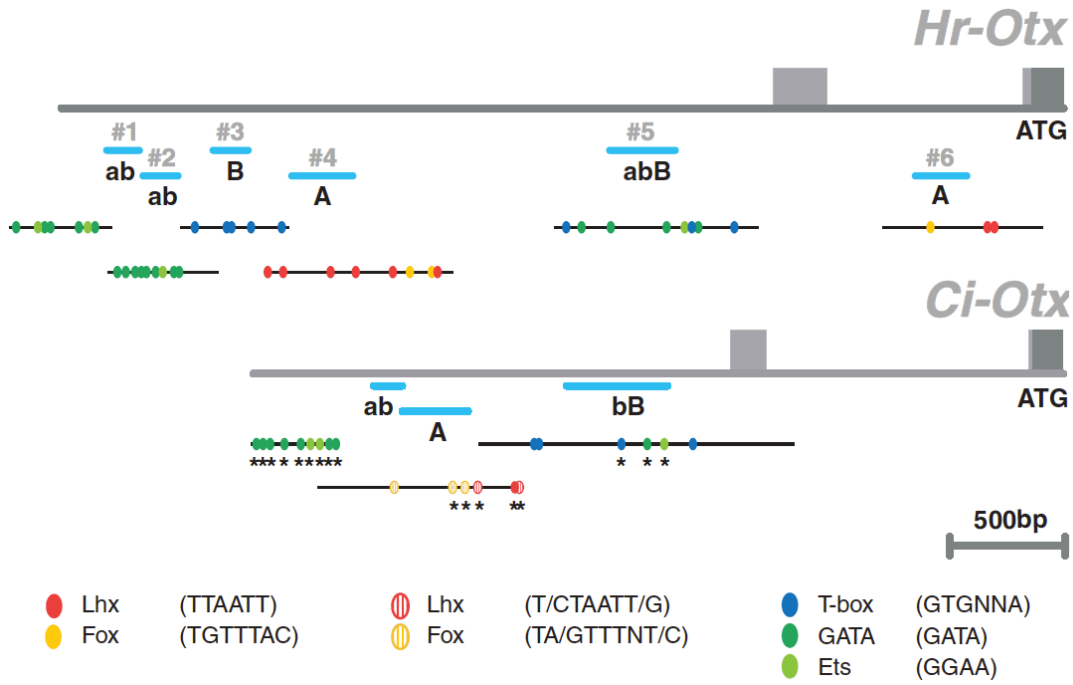


Figure II.9 : Les enhancers d'*Otx* de *Ciona intestinalis* (*Ci*) et *Halocynthia roretzi* (*Hr*) ont une activité conservée en dépit de leur grande diversité structurale. Les lignes grises représentent les régions 5' d'*Otx*. Les rectangles représentent les exons traduits (gris foncé) et non-traduits (gris clair). Les lignes bleues représentent les enhancers identifiées dans chaque espèce. Les lettres (A, a, B, b) représentent les lignages cellulaires où ces enhancers sont actifs. Les différents TFBS sont représentés par des ovales dont l'identité est définie par le code couleur en bas de la figure. Les astérisques indiquent les TFBS conservés entre *Ci* et *Cs* (Oda-Ishii et al., 2005).

De plus, dans le cas de l'enhancer neural d'*Otx* chez *Ciona intestinalis*, l'ordre, l'orientation et la distance entre les TFBS ne semblent pas contraints tant qu'au moins deux sites ETS et GATA sont présents et que deux TFBS sont séparés par au moins 5pb pour éviter l'encombrement stérique (Figure II.10) (Khoueiry et al., 2010).

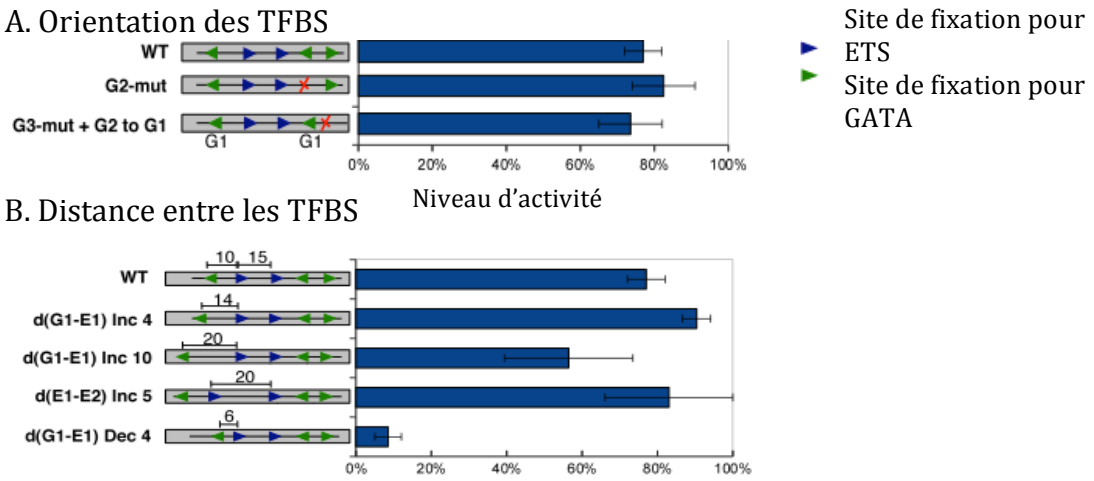


Figure II.10 : L'enhancer neural d'*Otx* semble avoir une architecture flexible : l'orientation d'un site de fixation pour un facteur de transcription (TFBS) **(A)** ou la distance entre deux TFBS **(B)** ne semble pas avoir d'incidence majeure sur l'activité de l'enhancer tant que deux TFBS sont séparés par au moins 5pb. À gauche sont schématisées les versions de l'enhancer testées dans une construction rapporteur. À droite, les histogrammes représentent le pourcentage d'embryons où l'enhancer est actif dans les précurseurs neuraux (Khoueiry et al., 2010).

Ces résultats suggèrent une flexibilité structurale de type "panneau d'affichage" (billboard) (Arnosti & Kulkarni, 2005; Spitz & Furlong, 2012). De tels enhancers serait donc plus difficiles à prédire sur la base de règles syntaxiques précises.

Mais ce n'est pas le cas pour tous les enhancers de *Ciona intestinalis* : une approche similaire a été menée par Passamaneck et ses collègues pour comprendre l'organisation des sites de fixation de *Brachyury* et *FoxA-a* dans l'enhancer activant l'expression de *Tune* dans la notochorde. La très grande majorité des modifications de l'organisation intrinsèque des TFBS diminuent ou annihilent l'activité de l'enhancer (Passamaneck et al., 2009) (Figure II.11).

Si plusieurs signatures de l'activité enhancer ont été identifiées, la compréhension de leur logique n'est pas suffisante pour prédire la présence d'enhancers à partir de la seule séquence de l'ADN génomique.

Dans la prochaine partie, je présenterai l'**élément a**, que j'ai utilisé pendant ma thèse comme enhancer-modèle pour mieux comprendre la logique de l'architecture d'un enhancer.

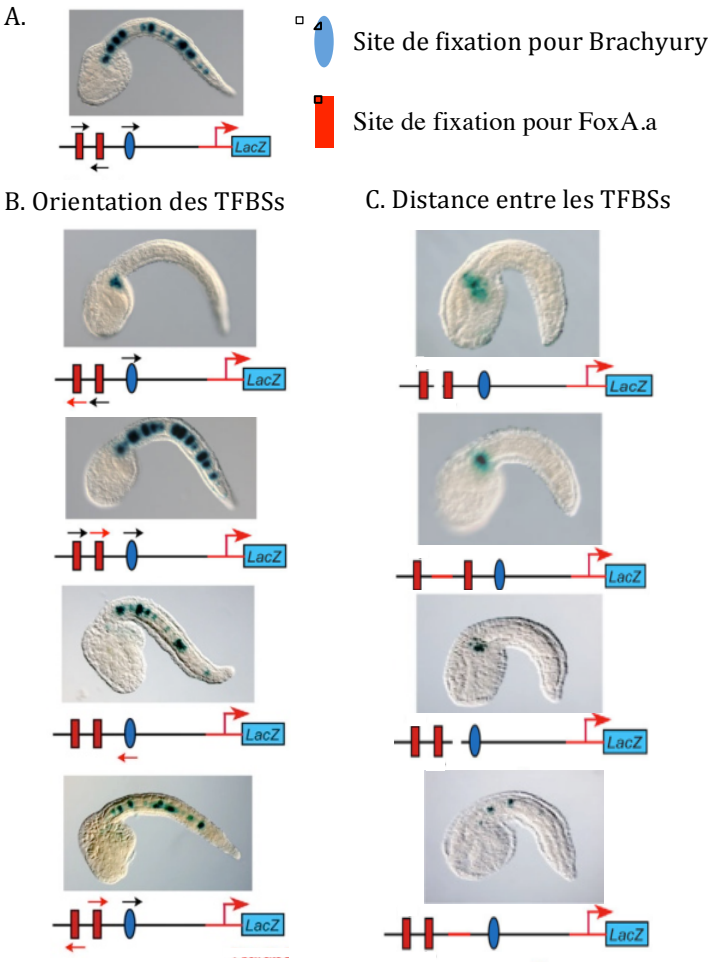


Figure II.11 : L'orientation (B) et la distance (C) entre les TFBS jouent un rôle dans l'activité de l'enhancer de Tune dans la notochorde (A).
Tous les embryons sont au stade du bourgeon caudal tardif et colorés par la ß-galactosidase. Sous les photographies sont schématisées les constructions rapporteurs qui ont été testées. Les sites de fixation de Brachyury et FoxA-a apparaissent respectivement en rouge et en bleu. Les flèches noires indiquent l'orientation naturelle des TFBS et les flèches rouges un changement de cette orientation. Enfin, les « coupures » correspondent à des délétions de 5pb et les traits rouges entre deux TFBS à des insertions de 10pb. Chaque embryon est représentatif de la majorité des embryons obtenus en électroporant chaque construction : dans la majorité des cas, l'activité dans la notochorde est altérée ou perdue. (D'après Passamaneck et al.,2009).

3. L'élément a, l'enhancer neural d'*Otx*

Chez l'ascidie, au stade du bourgeon caudal, le système nerveux central (CNS) contient la vésicule sensorielle, le cou, le ganglion viscéral et le tube nerveux caudal. Ces trois tissus seraient respectivement équivalents aux pro- et mésencéphales, au rhombencéphale et à la moelle épinière chez les vertébrés (Ikuta & Saiga, 2007). Chez *Ciona intestinalis,* le CNS, est formé par des cellules issues de trois paires de blastomères au stade 8 cellules (a4.2, b4.2 et A4.1) (Figure II.12) Les cellules végétatives issues du lignage A forment la partie postérieure de la vésicule sensorielle et les parties ventrales et latérales du ganglion viscéral et du tube nerveux caudal. Les cellules ectodermiques issues du lignages a forment la partie antérieure de la vésicule sensorielle et celles issues du lignage b forment la partie dorsale du ganglion viscéral et du tube nerveux caudal (Figure II.12).

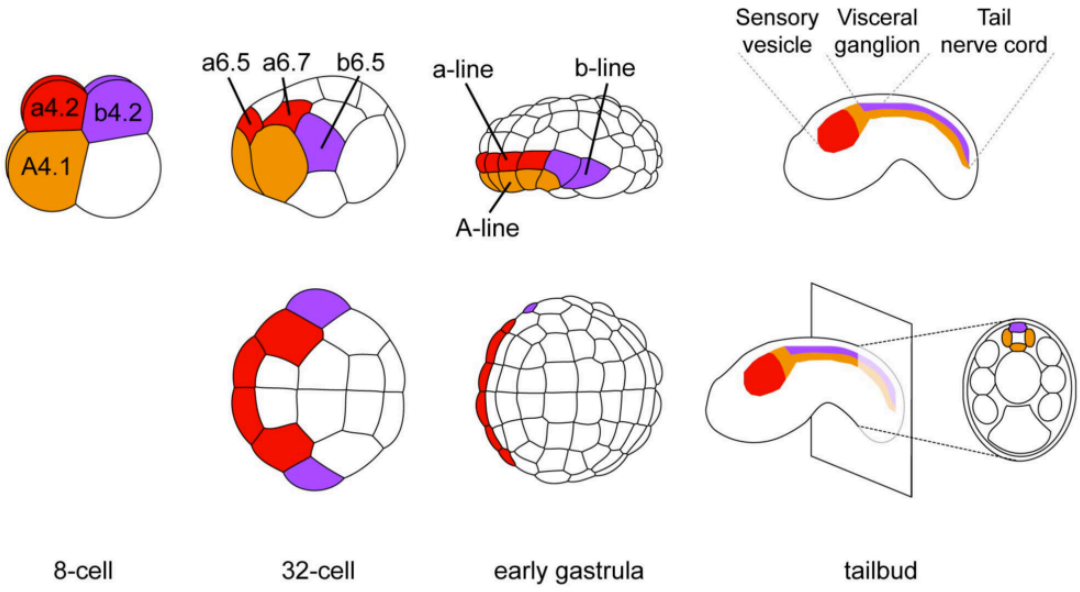


Figure II.12 Lignages cellulaires du système nerveux central (CNS) chez les ascidies. A chaque stade développemental, les cellules contribuant au CNS sont colorées suivant leur origine au stade 8 cellules. La coupe transversale montre l'organisation de la moelle épinière formée de cellules issues de lignages distincts : A et b.(Roure et al., 2014)

Si le système nerveux central issu des blastomères végétatifs est spécifié de façon autonome (Minokawa, Yagi, Makabe, & Nishida, 2001; Yasuo & Hudson, 2007), les précurseurs neuraux issus de cellules animales sont induits avant la gastrulation (Hudson & Lemaire, 2001; Reverberi & Minganti, 1946; Rose, 1939).

Dans cette partie, je m'intéresserai à l'induction neurale des blastomères du pôle animal et à la logique *cis*-régulatrice de l'enhancer neural précoce du gène *Otx*, le premier marqueur neural exprimé dans ces lignages.

a) Induction neurale

L'induction neurale est le phénomène de spécification de l'ectoderme en neuroectoderme pendant le développement embryonnaire animal (Wilson & Edlund, 2001). Chez les vertébrés, plusieurs voies de signalisation sont impliquées dans l'induction neurale, dont les voies FGF et BMP (Stern, 2005).

Chez les ascidies, l'induction neurale a lieu du stade 32-cellules au début de la gastrulation (Hudson & Lemaire, 2001). Des expériences menées sur *Halocynthya roretzi* ont montré que la voie BMP et ses antagonistes ne semblent pas impliqués dans l'induction neurale chez les ascidies (Darras et Nishida 2001). Par contre, chez *Ciona* (Bertrand et al., 2003), et peut-être *Halocynthia* (Tokuoka, Kumano, & Nishida, 2007), *FGF9/16/20* induit un destin neural dans certaines cellules ectodermiques.

Dès le stade 32 cellules, le destin de la paire de blastomères a6.5 est restreint à la plaque neurale antérieure, alors que la paire b6.5 contribue à la fois des destins neuraux et épidermiques dorsaux (Figure II.12). Pour comprendre les mécanismes moléculaires définissant un destin neural, C. Hudson et P. Lemaire montrèrent, par une série d'ablations et de recombinaisons de blastomères, que lorsque les lignages a et b sont explantés au stade 8-cellules, ils ne forment que de l'épiderme (Hudson & Lemaire, 2001). Le contact cellulaire entre les blastomères végétatifs des lignages A et B et les précurseurs neuraux animaux est requis à partir du stade 32 cellules pour l'activation des marqueurs neuraux animaux dans les blastomères a6.5 et b6.5, dont le plus précoce d'entre eux, *Otx* (Figure II.13 A). Les blastomères végétatifs A4.1 et B4.1 n'ont pas la même capacité à induire un destin neural : l'induction provient principalement de la descendance de la paire de blastomères végétatifs antérieurs A4.1 (Figure II.13 B) (Hudson & Lemaire, 2001). La réponse des cellules issues des lignages a et b à cette induction est aussi différente : les premières formeront la vésicule sensorielle antérieure et les secondes la partie dorsale du tube neural (Hudson & Lemaire, 2001; Rothbächer, Bertrand, Lamy, & Lemaire, 2007).

A la suite de ce travail, Bertrand et ses collègues montrèrent que FGF9/16/20, exprimé dans les cellules végétatives inductrices dès le stade 16 cellules, est l'inducteur neural endogène chez *Ciona intestinalis* (Bertrand et al., 2003). Lorsque FGF9/16/20 est inhibé par injection d'un morpholino antisens dirigé contre FGF9/16/20, les progéniteurs neuraux adoptent un destin épidermique. Par des expériences d'inhibition de la synthèse protéique, ces auteurs montrèrent également qu'*Otx*, dont l'action de spécification du tissu neural antérieur est conservée chez les bilatériens (Hirth, 2003), est une cible directe de la voie de signalisation FGF9/16/20. *Otx* est régulé directement par FGF9/16/20 via les facteurs de transcription maternels Ets1/2 et Gata4/5/6 (homologue aux GATA 4, 5 et 6 chez les vertébrés) (Bertrand et al., 2003). Gata4/5/6 est le facteur maternel déterminant l'identité animale dans l'embryon. Issu d'un ARN maternel ubiquitaire, son l'activité est restreinte à l'hémisphère animal par l'action antagoniste du

complexe β -catenin-TCF dans l'hémisphère végétatif (Rothbächer, Bertrand, Lamy, & Lemaire, 2007). Ets1/2 est un facteur de transcription maternel ubiquitaire qui médie l'action de FGF à la fois dans les cellules animales et végétatives (Bertrand et al., 2003; Miya & Nishida, 2003). *Otx* est donc induit à l'intersection des cellules où Gata4/5/6 est actif, et où Ets1/2 est activé en réponse à la voie de signalisation FGF (Rothbächer, Bertrand, Lamy, & Lemaire, 2007).

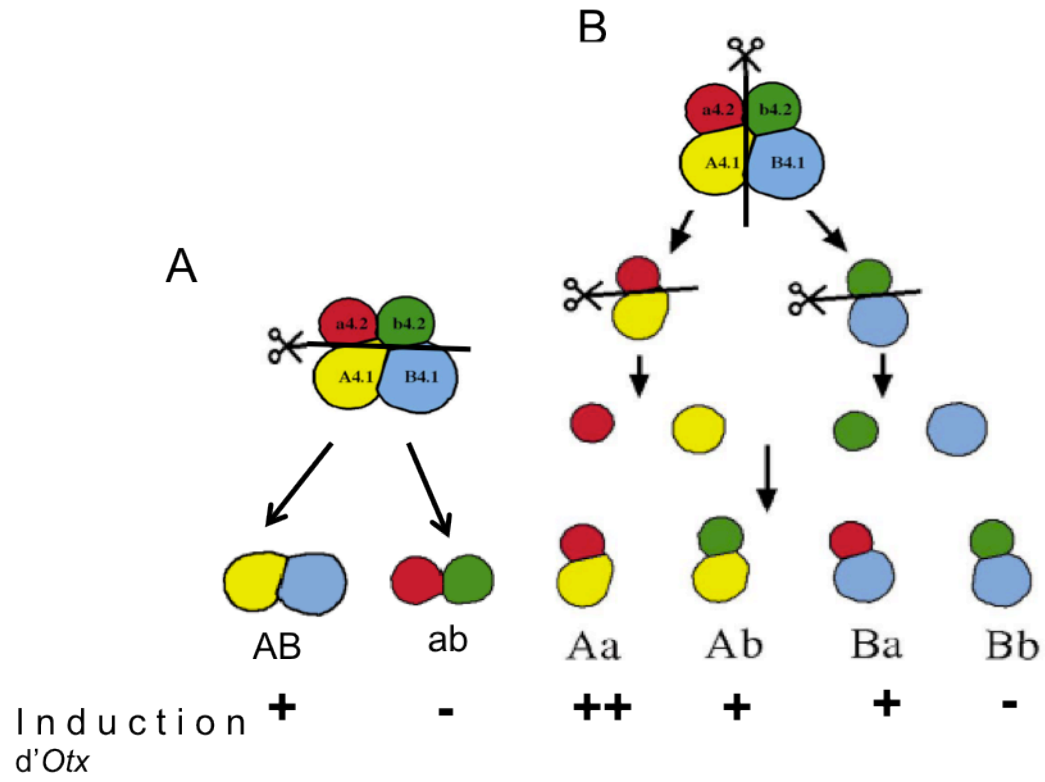


Figure II.13 : Dissection et recombinaisons de blastomères. A. Le contact avec les blastomères végétatifs (A4.1 et B4.1) est nécessaire pour induire *Otx* dans les blastomères animaux (a4.1 et b4.1). B. Les blastomères végétatifs n'ont pas la même capacité à induire *Otx*, les blastomères animaux n'ont pas la même aptitude à répondre à cette induction. L'induction d'*Otx* correspond au pourcentage d'embryons où l'expression d'*Otx* a pu être visualisée par hybridation *in situ* au stade du bourgeon caudal. (Adapté de Hudson 2001)

Il est intéressant de noter que les cellules a6.5 et b6.5 sont les blastomères animaux ayant les plus grandes surfaces de contact avec les blastomères végétatifs inducteurs, ce qui explique pourquoi seules ces deux paires de cellules ont un destin neural induit, alors que toutes les cellules animales sont également compétentes à répondre au FGF et exposées à FGF par leur contact avec les cellules végétatives inductrices (Hudson & Lemaire, 2001; Tassy, Daian,

Hudson, Bertrand, & Lemaire, 2006). De plus, *Otx* est parfois exprimé dans le blastomère animal a6.7 (Ohta & Satou, 2013). Cette cellule animale contribue également à la plaque neurale et a la troisième plus grande surface de contact avec l'hémisphère végétatif.

b) L'élément a, enhancer neural d'*Otx*

L'enhancer activant l'expression d'*Otx* dans les précurseurs neuraux au stade 32 cellules a été caractérisé par Vincent Bertrand, un ancien doctorant de l'équipe. L'isolement et le clonage de 3,5kb en amont d'*Otx* dans une construction rapporteur suivis de délétions successives ont permis d'identifier une petite région *cis*-régulatrice capable de récapituler qualitativement l'expression endogène d'*Otx* dans les blastomères a6.5 et b6.5 (Figure II.14). Cette région de 55pb, appelée **"l'élément a"**, contient trois sites de fixation pour Gata4/5/6 et deux pour Ets1/2 (Bertrand et al., 2003). L'analyse des sites de fixation et des expériences de perte de fonction ont permis de montrer qu'une synergie entre les déterminants maternels Ets1/2 et GATA4/5/6 est requise pour une réponse neurale spécifique des territoires animaux à la voie de signalisation FGF (Figure II.15A, B).

Rothbächer et ses collègues ont, eux, montré que les sites de fixation pour ETS ont deux rôles opposés : i) en l'absence de l'inducteur neural FGF9/16/20, ces sites répriment l'élément a. ii) A la suite de l'induction par FGF9/16/20, les sites de fixation pour Ets1/2 acquièrent un rôle activateur, et, en synergie avec les sites de fixation pour Gata4/5/6, activent l'expression d'*Otx* dans les précurseurs neuraux a6.5 et b6.5 (Rothbächer, Bertrand, Lamy, & Lemaire, 2007). Bertrand et ses collaborateurs proposèrent donc un modèle plus raffiné (Figure II.15 B) pour l'induction neurale chez *Ciona intestinalis* (Bertrand et al., 2003). A noter que la présence du second site de fixation pour GATA n'est pas nécessaire à l'activité de l'élément a (Figure II.8.b) (Khoueiry et al., 2010).

L'élément a semble être un enhancer de type "panneau d'affichage", car l'orientation et l'arrangement des sites de fixation pour ETS et GATA ne semble pas avoir d'effet sur son activité, tant qu'au moins 5pb séparent deux sites successifs (Figure II.10) (Khoueiry et al., 2010). Il est toutefois important de noter que cette flexibilité n'a pas été testée de façon systématique : seules les distances entre les trois premiers sites de fixation ont été individuellement modifiées. Il a été montré récemment que la distance entre les sites peut

influencer le niveau d'activité de l'enhancer (Farley et al., 2015). Cependant, certains résultats de ces deux études sur l'effet de la modulation des distances entre sites ETS paraissent contradictoires.

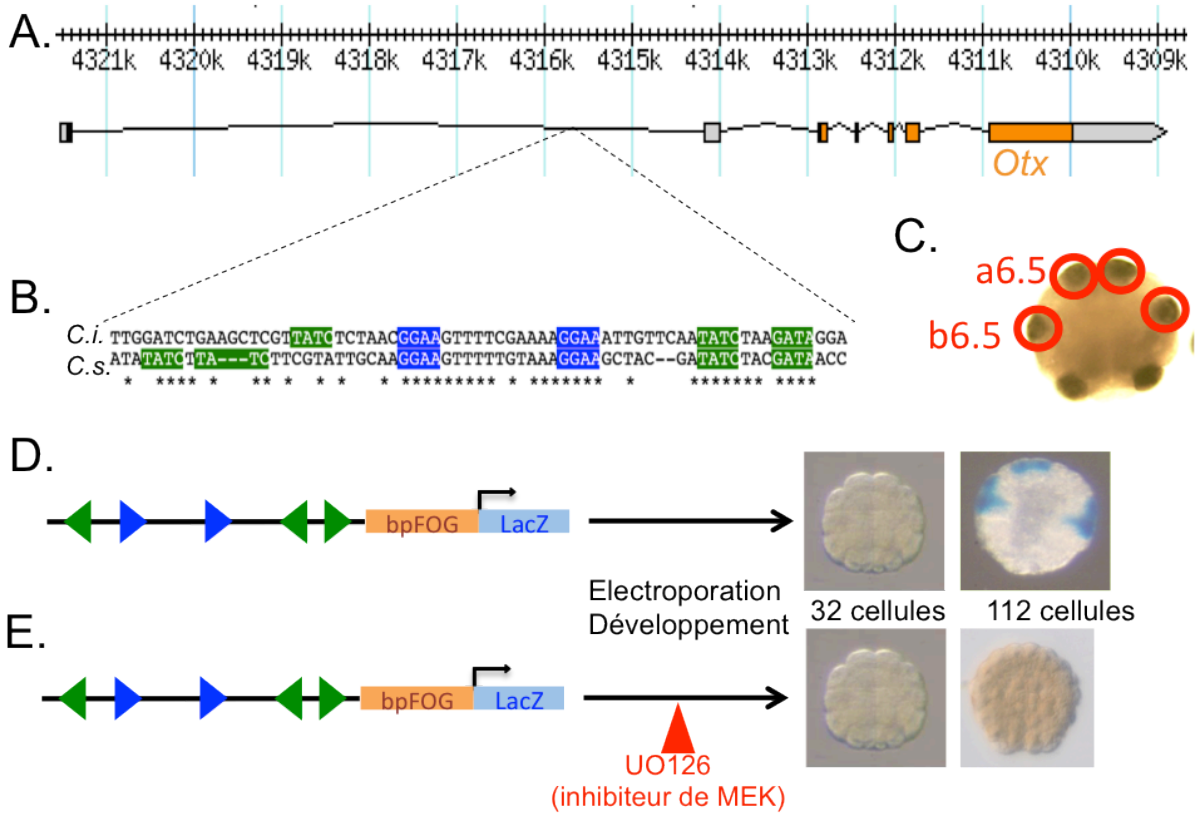


Figure II.14 L'élément a est l'enhancer neural précoce *d'Otx* **A.** L'élément a se situe dans le premier intron d'*Otx*. **B.** L'élément a est conservé entre *Ciona intestinalis (Ci)* et *Ciona savigniy (C.s)*. Les sites de fixation de Ets1/2 et Gata4/5/6 apparaissent en bleu et en vert respectivement. Les astérisques indiquent les nucléotides conservés entre les deux espèces. **C.** Au stade 32 cellules, *Otx* est exprimé dans les cellules neurales a6.5 et b6.5 (hybridation *in situ*) **D.** L'électroporation de l'élément a cloné dans une construction rapporteur (schématisée à gauche) permet de visualiser la présence de beta-galactosidase au stade 112 cellules dans les lignages des blastomères a6.5 et b6.5. Si l'ARN est détectable au stade 32 cellules tardif, la protéine dont la synthèse prend plus de temps et qui est stable, n'est détectée que plus tardivement. **E.** L'inhibition de la voie FGF au stade 16 cellules bloque l'activité de l'élément a. (D'après Bertrand 2003).

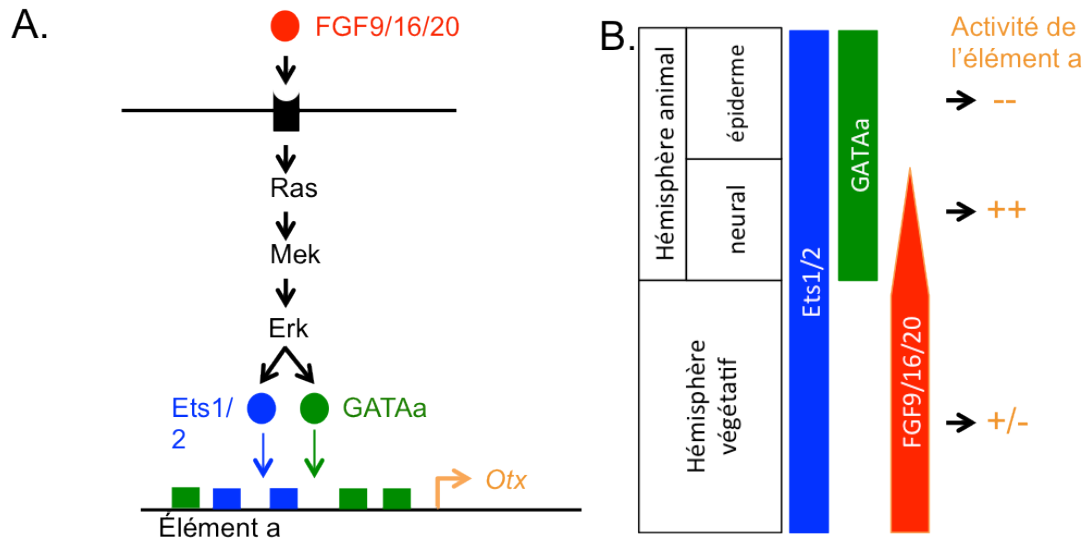


Figure II.15 : **Modèle pour l'induction neurale chez *Ciona intestinalis*. A**. Voie de signalisation pour l'activation de l'expression d'*Otx* dans les précurseurs neuraux par la voie de signalisation FGF. **B.** Activité des FGF16/9/20, Ets1/2 et Gata4/5/6 dans l'embryon. Les régions où Ets1/2 et Gata4/5/6 sont capables de répondre à la signalisation FGF sont représentées en bleu et en vert respectivement. Les régions où FGF16/20 est présent sont en rouge. L'expression d'*Otx* est la plus forte dans les cellules où Ets1/2 et Gata4/5/6 reçoivent de la signalisation FGF (adapté de Betrand, 2003)

c) À la recherche d'enhancers similaires à l'élément a

Dans le but de trouver des enhancers fonctionnellement similaires à l'élément a dans le génomes de *Ciona intestinalis*, Khoueiry et ses collègues (Khoueiry et al., 2010) conçurent un programme appelé SECOMOD (SEarch for Evolutionary COnserved MODules ; Recherche de modules conservés évolutivement), visant à identifier des séquences non codantes ayant une logique *cis*-régulatrice similaire à celle de l'élément a (Figure II.16). La structure de l'élément a n'étant apparemment pas contrainte, cet algorithme scanne les régions non codantes orthologues des génomes de *Ciona intestinalis* et *Ciona savignyi* à la recherche de séquences de 80 à 130pb contenant au moins deux sites de fixation pour ETS et GATA dans les deux espèces. Les sites de fixation pour GATA et ETS ont été prédits en utilisant des séquences consensus très permissives. Afin d'éviter un risque d'encombrement stérique entre protéines fixées sur des sites voisins, un filtre supplémentaire impose une distance d'au moins 5pb entre deux sites de fixation. Ce programme tolère un décalage entre les positions des sites de fixation dans des clusters orthologues chez *Ciona intestinalis* et *Ciona savigny*.

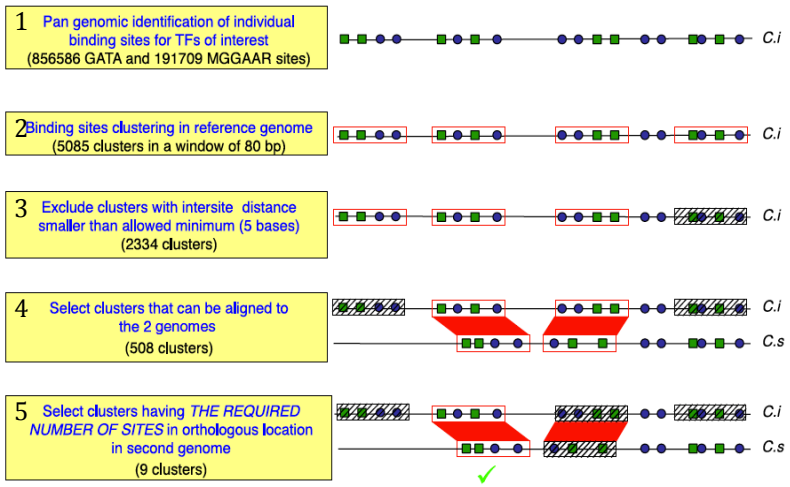


Figure II.16 : Stratégie utilisée par le programme SECOMOD. À droite, le logigramme (flow chart) du processus dans le cas de la recherche de clusters contenant deux sites de fixation pour ETS et deux pour GATA, dans une fenêtre de 80pb. À droite, la représentation schématique du type de cluster sélectionné à chaque étape (numérotées de 1 à 5). Pour chaque étape, seuls les clusters qui en satisferont les critères (en rouge) seront analysés à l'étape suivante. Les clusters éliminés sont hachurés en noir (Khoueiry, 2010).

Cette recherche *in silico* a permis d'identifier 55 regroupements de sites ETS et GATA conservés entre *Ciona intestinalis* et *Ciona savignyi*. L'activité de 19 d'entre eux a été testée *in vivo* ce qui a permis d'identifier quatre nouveaux enhancers actifs dans les précurseurs neuraux des lignages a et b et trois enhancers actifs dans d'autres lignages cellulaires. L'analyse des profils d'expression des gènes les plus proches des quatre nouveaux enhancers neuraux a montré qu'ils contrôlent l'expression d'*ELK*, *Prickle*, *ERF-a* et *Nodal* dans les progéniteurs neuraux (Figure II.17) (Khoueiry et al., 2010).

Le fait que seule une minorité de regroupements de sites aient une activité enhancer même lorsqu'ils sont isolés de leur contexte chromatinien endogène, et que les regroupements actifs soient localisés à proximité de gènes exprimés dans les cellules a6.5 et/ou b6.5, suggère que l'information *cis*-régulatrice réside dans leur courte séquence, et que la lecture de cette information est peu impactée par l'état chromatinien endogène.

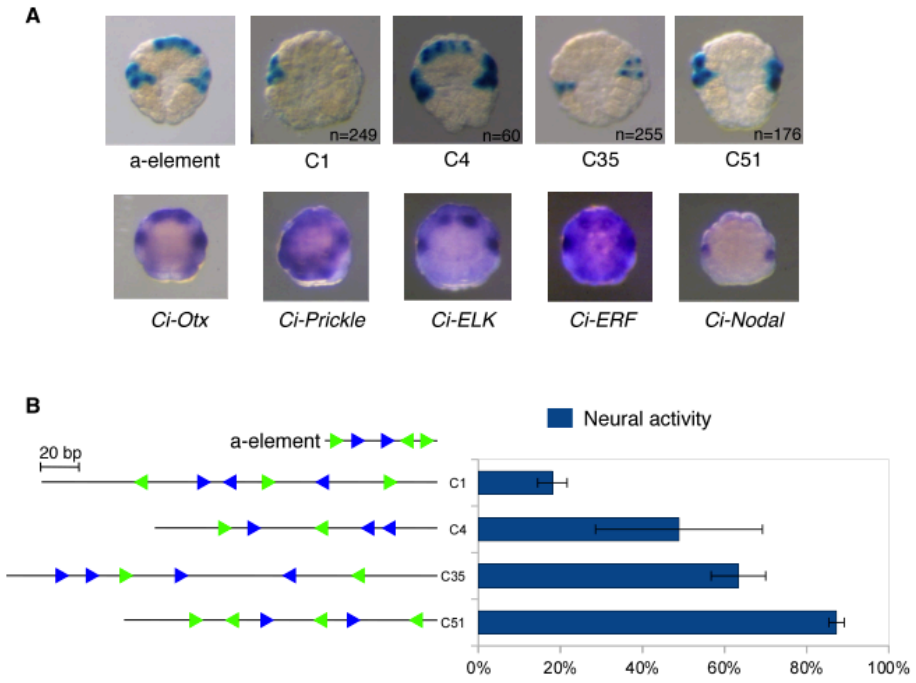


Figure II.17 Cinq clusters de sites pour les facteurs ETS et GATA sont des enhancers neuraux précoces. A. En haut : embryons au stade 112 cellules électroporés avec les différents clusters. En bas : hybridation *in situ* des gènes les plus proches de ces clusters. **B.** À gauche sont représentés les architectures des différents clusters contenant des sites de fixation d'ETS (en bleu) et de GATA (en vert) Les histogrammes représentent le pourcentage d'embryons où le cluster est actif dans les précurseurs neuraux. (Khoueiry 2010).

Du fait de la définition très permissive des sites des facteurs de transcription dans cette étude, une faible affinité pour leurs facteurs de transcription pourrait expliquer l'inactivité de certaines séquences testées. Pour tester cette hypothèse, les séquences des sites de 4 regroupements furent remplacés par celles, de forte affinité, du premier site GATA (AgataA) et du premier site ETS (ACggaAG) de l'élément a. Dans le cas de l'enhancer neural de *Prickle*, ces mutations ont effectivement permis d'augmenter significativement son activité. Mais ces modifications n'ont permis d'activer qu'un seul des trois regroupements inactifs (Figure II.18). Ainsi, l'affinité prédite des sites de fixation est un déterminant important mais insuffisant de l'activité d'un enhancer (Khoueiry et al., 2010).

Un autre déterminant crucial de l'activité d'un enhancer serait la structure locale de la chromatine, via la compétition thermodynamique entre la fixation des facteurs de transcription

et la présence de nucléosomes, les TFBSs fonctionnels étant souvent trouvés dans des régions dépourvues de nucléosomes (Segal & Widom, 2009) (voir partie I-A-1-c). L'élément a n'étant actif que dans 4 des 32 blastomères de l'embryon, il n'y a pas assez de matériel pour caractériser la position des nucléosomes et les marques épigénétiques *in vivo* dans ces cellules. Segal et ses collègues (2006) proposèrent que l'affinité des nucléosomes soit contrôlée par une signature dinucléotidique intrinsèque à la séquence d'ADN, et développèrent un modèle probabiliste pour prédire la position des nucléosomes à partir de la séquence d'ADN. Khoueiry et ses collègues utilisèrent cette approche pour tester l'hypothèse qu'une faible affinité pour les nucléosomes soit requise pour l'activité des regroupements de sites GATA et ETS.

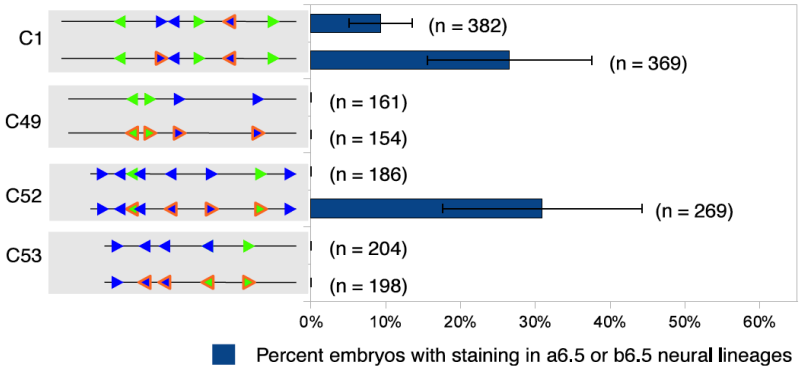


Figure II.18 : La modification de la séquence des TFBS affecte l'activité *cis*-régulatrice d'un cluster de TFBS. À gauche sont représentés les architectures de C1, l'enhancer neural de *Prickle,* et de trois clusters inactifs (C49, C52 et C53) avant (en haut) et après (en bas) le remplacement des sites de fixation d'ETS (en bleu) et GATA (en vert). Les TFBS modifiés ont un liseré orange. Les histogrammes représentent le pourcentage d'embryons où le cluster est actif dans les précurseurs neuraux. (Khoueiry 2010).

Les 20 regroupements cités précédemment et 79 enhancers actifs précédemment caractérisés chez *Ciona intestinalis* ont été analysés avec le programme de Segal, entraîné sur un jeu de données de poulet. Dans l'ensemble, les enhancers actifs présentent une signature dinucléotidique différente de celle des clusters inactifs et auraient peu de chance d'être occupés par un nucléosome, alors que les clusters inactifs ont de grandes chances de l'être.

Toutefois, de tels résultats n'ont pu être obtenus quand le programme était entraîné sur un jeu de données de levure (Khoueiry et al., 2010; Segal et al., 2006) ou en utilisant d'autres programmes de prédiction de positionnement de nucléosomes (Kaplan et al., 2009; Peckham et al., 2007). Il demeure donc incertain si la signature dinucléotidique associée à l'activité enhancer identifiée par Khoueiry et al, reflète effectivement la probabilité de fixation de nucléosomes, ou d'autres propriétés physiques de l'hélice d'ADN affectées par leur séquence, comme par exemple la largeur du petit sillon (Bishop et al., 2011; White, Myers, Corbo, & Cohen, 2013) ou la flexibilité de l'ADN (Geggier & Vologodskii, 2010).

La conservation évolutive de cette signature dinucléotidique entre *Ciona intestinalis* et *Ciona savignyi,* indépendamment du niveau de conservation des séquences nucléotidiques, suggère que les propriétés physiques de la molécule d'ADN sont aussi sous pression de sélection. Enfin, la corrélation entre activité *cis*-régulatrice et cette signature dinucléotidique est également observée chez la drosophile, mais pas chez l'Homme.

III. Buts de mon projet

Globalement, les découvertes présentées en introduction, pour encourageantes qu'elles soient, laissent encore de nombreuses zones d'ombres sur les déterminants des enhancers. Elles nous ont encouragés à continuer à rechercher les règles qui régissent l'activité des séquences régulatrices.

Au cours de ma thèse, j'ai utilisé le système ascidie et l'élément a comme modèles d'étude des déterminants de l'activité des enhancers développementaux. J'ai d'abord analysé l'effet de l'affinité des sites de fixation des facteurs de transcription ETS et GATA sur l'activité de l'enhancer *in vivo* et sur la fixation des facteurs de transcription *in vitro.* Je me suis ensuite intéressée aux séquences situées entre les sites de fixation, et montré qu'elles ont un effet majeur sur le niveau d'activité de l'élément.

De plus, dans le cadre d'une collaboration avec plusieurs laboratoires européens, j'ai testé chez *Ciona intestinalis* la conservation fonctionnelle de séquences non-codantes ultraconservées entre les vertébrés et les tuniciers.

RESULTATS

Article 1 : Spacer sequences set enhancer activity levels

Cet article – ou futur article – présente ma contribution à l'effort général de déchiffrage des séquences *cis*-régulatrices.

Nous avons choisi de simplifier la séquence d'un enhancer en distinguant les séquences reconnues et fixées par les facteurs de transcription des autres, les spacers, peu considérés en tant qu'acteurs potentiels. Nous avons étudié de façon indépendante leur rôle dans l'activité d'un enhancer, en utilisant *Ciona intestinalis* comme organisme modèle.

Nous avons d'abord testé *in vivo* l'impact sur l'activité de l'élément-a, l'enhancer neural du gène *Otx*, de mutations ponctuelles dans les sites de fixation d'ETS ou de GATA, et confirmé que l'affinité d'un facteur de transcription pour son site de fixation affecte quantitativement le niveau d'activité d'un enhancer. Nous avons également montré que des données d'affinité de fixation *in vitro* (SELEX-seq), ne permettent pas systématiquement de prédire le niveau d'activité des différents variants en fonction des séquences de leurs sites de fixation.

Peu d'études ayant cherché à caractériser les spacers, nous n'avions pas d'hypothèse à confirmer/infirmer. Pour que nos résultats soient le moins biaisés possible, nous avons randomisé dans l'élément-a les séquences situées entre les TFBS et n'affectant a priori pas l'affinité des facteurs de transcription. Un très faible nombre de regroupements génomiques de sites ETS et GATA étant actifs, nous avions supposé que la séquence des spacer était très contrainte, et nous attendions donc à ce que cette approche « à l'aveugle » génèrerait un très faible ratio d'éléments actifs. Pour cette raison, nous avions développé un système de codes barres transcrits uniques associés à chaque enhancer dégénéré, nous permettant de multiplier par 100 au moins le nombre de séquences dont l'activité pouvant être testée *in vivo* par les approches traditionnelles d'électroporation de constructions rapporteurs chez l'ascidie. L'idée était d'utiliser ensuite des approches informatiques de type apprentissage automatique pour essayer d'extraire des séquences testées une signature corrélant avec leur niveau d'activité.

D'abords testés *in vivo* individuellement, ces éléments synthétiques ont des niveaux d'activité très variés et, lorsqu'ils sont actifs, ils le sont dans les mêmes territoires que l'élément a. Et surtout, à notre grande surprise, la majorité de ces éléments sont actifs. Nous avons donc changé l'échelle de cette étude, perdant en résolution statistique, mais gagnant la résolution spatiale de l'activité de chaque élément testé. Nous avons reproduit ces résultats en randomisant les spacers d'un autre enhancer neural précoce, et deux regroupements de sites ETS et GATA inactifs.

Les avons montré que les spacers sont des acteurs à part entière de l'activité d'un enhancer, et qu'ils peuvent influencer la fixation des facteurs de transcription.

SPACER SEQUENCES SET ENHANCER ACTIVITY LEVELS

Marion Guéroult-Bellone[1], Rémy Beule-Dauzat[1], Willi Kari[2], Ute Rothbächer[2], Christelle Dantec[1], Jacques Piette[1#] and Patrick Lemaire[1#]

(1) Centre de Recherche de Biochimie Macromoléculaire, CNRS-Université de Montpellier, 1919 route de Mende, 34293 Montpellier, France
(2) Department of Evolution and Developmental Biology, Zoological Institute, University Innsbruck, Technikerstrasse 25, A-6020 Innsbruck, Austria.

# equal contribution and authors for correspondence

Enhancers are crucial regulators of gene expression during development. In spite of extensive research over the past decades, what confers enhancer properties to a stretch of genomic DNA remains poorly understood. The small (55pb) early neural enhancer of the *Ciona intestinalis Otx* gene is one of the best-studied animal enhancers. This module is activated in response to FGF signaling and its activity results from the binding of Gata4/5/6 and Ets1/2. Here, we independently mutated the bases contacted by the transcription factors or randomized the spacers separating binding sites. These experiments uncover an unexpected role for the spacers in setting the level of activity of the a-element, a finding which we extended to other active or inactive genomic clusters of ETS and GATA motifs. We show that randomization of the spacers sequences of the a-element can increase enhancer activity as a result of increased sensitivity to FGF signaling. These enhancers, however, become active in cells in which the wild-type a-element is normally only weakly active, suggesting that in ascidian embryos, enhancers are optimized for maximum specificity rather than for maximum activity. Previous work established that the interaction of transcription factors and DNA involves both the sequence of their binding site and structural features of the double helix. Using *in vitro* transcription factor binding assays, we show that randomization of spacer sequences can affect transcription factor binding to the a-element. In particular, base pairs located outside the binding motif set the affinity of Ets1/2 for the enhancer, most likely by modifying the shape of the DNA helix. Our results point to the critical role of spacer sequences in setting enhancer activity levels.

INTRODUCTION

Enhancers play a fundamental role in development, homeostasis, evolution and disease [1], [2]. They act as scaffolding platforms for transcription factors (TF) and are generally composed of clusters of several binding sites for at least 2 transcription factors. The degree of constraints on the spacing, order and orientation of transcription factor binding site (TFBS) is variable, with a majority of enhancers active during animal development showing little constraints [3]. In spite of this apparent flexibility, we do not understand the determinants of enhancer activity and it remains very difficult to rationally engineer synthetic enhancers from the knowledge of TF binding sites only [4].

The a-element of the ascidian *Ciona intestinalis* is one of the best-characterized chordate enhancers [5]–[8]. This short (55bp) enhancer drives the embryonic expression of the *Otx* gene from the late 32-cell stage in two animal neural lineages, a6.5 and b6.5 (Figure 1A), in response to the FGF9/16/20 neural inducer [5]. This element is also weakly active in the posterior muscle lineage (B6.4) and in the neural progeny of the a6.7 cell pair, the a7.13 lineage, two territories that also express *Otx* (Figure 1A). The *cis*-regulatory logic driving the activity of this element in neural lineages has been characterized in detail (Figure 1B). Two maternal transcription factors, Ets1/2 and Gata4/5/6, cooperate to achieve FGF inducibility and tissue specificity, respectively [5], [6]. FGF-induced binding of the ubiquitous Ets1/2 to two sites, E1 and E2, drives expression in FGF-responding cells across all germ layers, while binding of the animal determinant Gata4/5/6 to three sites, G1, 2 & 3, restricts the activation to the animal territories (Figure 1C). G2 only weakly binds Gata4/5/6 and is not required for enhancer activity. By contrast, mutational inactivation of G3 or E2 leads to a complete loss (G3) or a strong decrease (E2) of enhancer activity, respectively, indicating that binding of Gata4/5/6 and Ets1/2 to these two sites is crucial for a-element activity [7].

The spacing and orientation of ETS and GATA binding sites does not seem to play a major role in the element activity [7]. In spite of this apparent flexibility, only a minority of *Ciona* genomic clusters containing at least 2 ETS and 2 GATA binding motifs have enhancer activity [7]. Recent studies suggested that enhancer activity is, in part, determined by residues outside of the DNA sequence directly contacted by the transcription factors [7], [9]–[12]. A recent study of the a-element proposed that the major determinants of enhancer activity are included in an octamer composed of the core recognition tetramer for ETS and GATA and the adjacent nucleotides [8]. Here, we present the results of an independent analysis of the

sequence determinants of the activity of this enhancer, the a-element, and of other *Ciona* early neural enhancers responding to the same *cis*-regulatory logic.

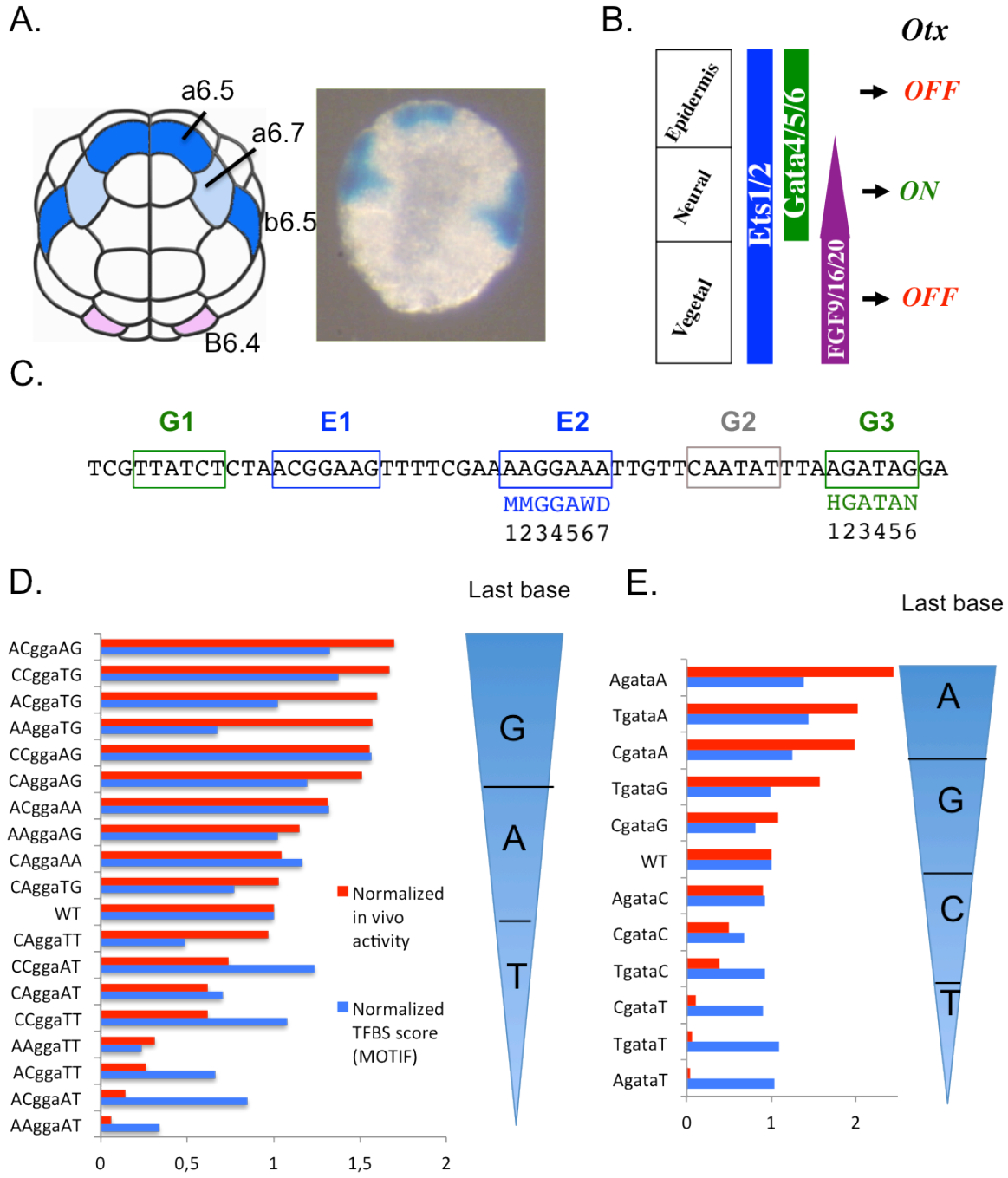


Figure 1. Influence of point mutations in the E2 and G3 sites on *in silico* TF binding scores and *in vivo* enhancer activity. A. Left: 32-cell stage embryo with coloured cells in which the a-element is strongly (blue) or weakly active (light blue and rose). Right: early gastrula with cells stained in blue for LacZ expression under the control of the a-element. **B.** Neural induction of the a-element by the combined activity of Gata4/5/6 and Ets1/2. **C.** a-element sequence describing ETS (green) and GATA (blue) site mutations tested *in vivo*. The inactivated G2 site is in grey. **D**. Comparison of *in vivo* enhancer activity and *in silico* predicted binding of ETS calculated for E2 (MOTIF score). **E.** Comparison of *in vivo* enhancer activity and *in silico* predicted binding of Gata4/5/6 calculated for G3.

RESULTS

Contribution of the bases contacted by Transcription Factors to enhancer activity

We first analyzed the influence of the sequence of the stretch of DNA directly contacted by Ets1/2 and Gata4/5/6 on the *in vivo* enhancer activity of the a-element. Published crystal structures for mammalian homologs bound to DNA (Supplementary Figure 1) suggest that Gata4/5/6 directly contacts 6 bases: a central "GATA" core flanked on either side by one nucleotide, in agreement with its recognition motif [13], [14]. The contacts established by Ets1/2 span 7 bases including a minor groove contact between positions 7 and 8 [15]. We therefore mutated combinations of bases flanking the "GGAA" core of E2 and the "GATA" core of G3 (Figure 1C) and compared the activity of mutant enhancers by scoring LacZ staining in the a6.5 and b6.5 lineages in 112-cell stage embryos electroporated with reporter constructs introduced by electroporation (Figure 1D, E).

As expected, changes in the sequences of E2 and G3 quantitatively affected the activity of the a-element in a6.5 and/or b6.5 lineages, while preserving the qualitative spatial pattern of activity (Not shown). In response to alteration of either binding site, the activity levels of variant enhancers ranged from a complete loss to stronger levels than WT (Supplementary Figure 2) (Figure 1D,E). For both factors, the base located in the most 3' position had the largest effect on enhancer activity.

Surprisingly, *in silico*-predicted affinities of E2 and G3 variant sites for their cognate protein (Supplementary Figure 3) does not necessarily reflect the *in vivo* activity of the variant enhancer (Figure 1D, E). For example, irrespective of the *in silico* affinity of E2 for Ets1/2, enhancers possessing a "G" in position 7 of E2 are more active than the WT, while those possessing a "T" are less active (Figure 1D and Supp. Figure 4). Correlation between G3 *in silico* affinity for Gata4/5/6 and *in vivo* activity was overall higher (R^2 of 0.68165, Supp. Figure 4). Yet, a strong divergence between the two measures was also observed in enhancers possessing a "T" in position 6 of G3, which lacked *in vivo* enhancer activity in spite of similar Gata4/5/6 predicted *in silico* binding as WT G3 (Figure 1E). Thus, a-element *in vivo* enhancer activity cannot be easily explained by the *in silico* affinity of the TFBS sequences.

Consistently, out of 14 genomic clusters of 2 ETS and 2 GATA sites with *in silico* affinity scores for their cognate factor at least as good as the a-element, only two, N83 and N26,

behaved as enhancers (Supplementary Table 1). The activity of these two elements was restricted to the early neural lineages. Conversely, out of 19 clusters tested by Khoueiry and colleagues [7], inactive cluster C39 has higher *in silico* scores for all its TFBS than those of the a-element, while active C35 has 3 weaker TFBS. Thus, high *in silico* TFBS affinity is neither sufficient nor required for neural enhancer activity of genomic clusters of ETS and GATA sites.

Role of spacer sequences in enhancer activity

As neither the arrangement [7] nor the sequences of TF recognition sites fully explain enhancer activity, we next tested whether the stretches of nucleotides located between TFBS, subsequently called spacers, affect enhancer activity.

We constructed a library of synthetic enhancers, which were cloned upstream of the *Fog* basal promoter and the *LacZ* reporter gene. Each randomized variant shared with the a-element the six bases centered on the central "GATA" and "GGAA" core sequences of each of the 4 GATA- and ETS-binding site, respectively, as well as the orientation and spacing of these sites. All spacer sequences were, however, fully randomized. The *in vivo* enhancer activity of 34 randomized a-elements (Supplementary Figure 5; Supplementary Table 2) was determined by electroporation as above.

While a large majority of genomic clusters of putative ETS and GATA sites are inactive, 25 out of 34 (74%) randomized a-element variants had an activity higher or equal to 10% of the wild-type activity, and were thus considered active (Figure 2A). These enhancers displayed a wide range of activity levels ; half of the variants are at least as active as the original a-element. The activity of these variants was mostly restricted to a6.5 and b6.5 lineages, but we note an increased activity in other cells for most variants with higher activity than the WT enhancer, but also in some with lower activity (Supplementary Figure 7A).

Farley and coworkers [8] independently quantified the *in vivo* activity of 163000 randomized a-element variants. Reanalysis of this dataset supports our conclusions. Out of 325 variants with 2 ETS and 2 GATA sites with similar or higher *in silico* scores than the original a-element, 217 (67%) have an activity level greater than 10% of the a-element. Out of 51

variants with near-optimal sites according to their analysis, only 37 (73%) are considered active.

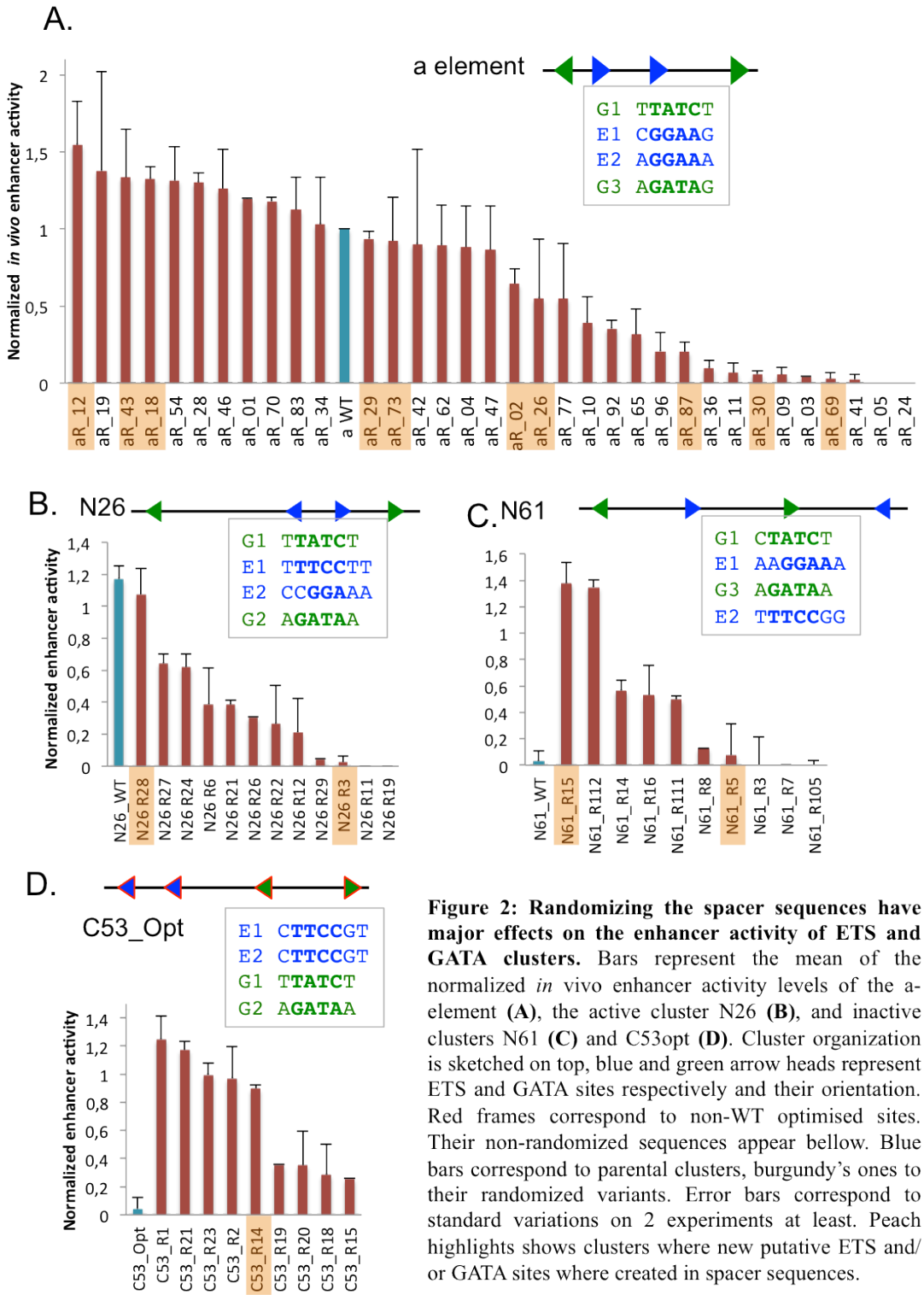


Figure 2: Randomizing the spacer sequences have major effects on the enhancer activity of ETS and GATA clusters. Bars represent the mean of the normalized *in* vivo enhancer activity levels of the a-element **(A)**, the active cluster N26 **(B)**, and inactive clusters N61 **(C)** and C53opt **(D)**. Cluster organization is sketched on top, blue and green arrow heads represent ETS and GATA sites respectively and their orientation. Red frames correspond to non-WT optimised sites. Their non-randomized sequences appear bellow. Blue bars correspond to parental clusters, burgundy's ones to their randomized variants. Error bars correspond to standard variations on 2 experiments at least. Peach highlights shows clusters where new putative ETS and/ or GATA sites where created in spacer sequences.

To test whether this major effect of spacers sequences on enhancer activity levels was a specific feature of the very compact a-element, we randomized the spacer sequences of another active genomic ETS and GATA cluster, N26 (Supplementary Table 3), which sites

spacing and orientation differ from the a-element (Supplementary Figure 5B). Similar results (8/12 active variants in the same lineages) were obtained with N26, suggesting that the specific organization of the a-element is not involved (Figure 2B).

In some clusters, spacer sequence randomization created additional ETS and GATA binding sites (Figure 2A, B, highlighted in salmon, see also Supplementary Figure 8A). These novel sites cannot, however, provide a simple explanation for increased enhancer activity levels. For example, an additional putative recognition sequence for ETS was created in the spacers of both strongly active (aR_12, aR_43) and inactive (aR_87) a-element variants (Supplementary Figure 8A, B). Similarly, an additional GATA site appeared in the spacer of both active (aR_18, aR_73) and inactive (aR_30, aR_69) randomized variants (Supplementary Figure 8A,B). Potential high affinity recognition motifs for additional transcription factors present in the embryo before the onset of *Otx* activation were also detected in some of the variants Supplementary Figure 9)[16]. Yet, the spatial activity patterns of the variants were not consistent with the activity profiles of these additional factors.

We next tested the effect of randomizing the spacers of genomic inactive clusters of high-scoring *in silico* recognition sequences for ETS and GATA. Surprisingly, spacer randomization in two such clusters, N61, and a TFBS-optimized version of C53 [7] conferred early neural enhancer activity to some of their variants (5 out of 10 and all 9 variants respectively) (Figure 2C,D). Here also, creation of new ETS and GATA binding sites in the spacers is not sufficient to explain higher activity levels of variants compared to the original cluster (Figure 2C,D). Consistent with the proposed architectural flexibility of the a-element, all 4 randomized clusters have different TFBS organization.

Taken together, these experiments point to an unexpected and crucial role of spacer sequences in enhancer activity: depending on their sequences, they can not only modulate enhancer activity levels, but also inactivate active enhancers or activate inactive genomic clusters of ETS and GATA binding sites.

Randomized enhancers responding to lower levels of FGF signaling are more active in muscle and a-neural plate than the WT

Farley et al. showed that GATA and ETS BS were not optimal, and that optimization of the sites led to ectopic expression [8]. Nevertheless, we noted ectopic expression also in some variants with lower activity than the WT. This suggests that ectopic expression is not directly correlated to overall activity levels, but could be more specifically linked to one of the regulators of the a-element. Therefore, we decided to analyze in more detail the pattern of ectopic expression and the response of the variants to FGF signaling.

In addition to its main a6.5/b6.5 activity, the wild-type a-element also drives expression in muscle and a6.7-derived a-neural plate cells in a small percentage of embryos, reflecting endogenous expression of *Otx* [17]. Increasing the affinity of the four ETS and GATA binding sites by point mutations in the a-element (a_Opt construct) led to an overall increase in LacZ-expressing cells in particular the progeny of a6.5/b6.5 (dark blue), a6.7 (red) and muscle cells (orange) (Figure 3A). The activity pattern of four randomized variants was similar to that of the optimized a-element with activity in territories not expressing *Otx* in less than 5% of the stained embryos (Figure 3A). We note, however, that the fraction of embryos showing additional LacZ staining in muscle and/or a7.13 neural lineages, is higher in randomized variants than in the WT a-element. This broader spatial activity pattern is not entirely correlated to the activity levels of the variants as already noted. For instance aR__73 is more active in muscle and a7.13 neural lineages than the WT, although its overall activity levels are similar (Supplementary Figure 7A).

The activity of the a-element is induced by FGF signaling between the 16- and 32-cell stages [5] and increased FGF signaling leads to broader expression in animal territories including a7.13 lineages [17]. An increased sensibility of randomized variants to FGF signaling could thus explain their higher activity in territories where the parental enhancer is weakly active.
To compare the sensitivity to FGF signaling of variants, this pathway was partially or completely inhibited by treatment of electroporated embryos with the MEK inhibitor U0126. All variants were inactive upon complete FGF signaling inhibition when the drug was added at the 16-cell stage (Figure 3B). Most randomized variants were however less sensitive to a partial inhibition of FGF signaling than the a-element when the drug was added at the 32-cell stage. This ability to respond to lower levels of FGF may explain the reduced selectivity of

the variants. As strong non-a6.5-b6.5 expression does not always correlate with high enhancer activity levels, the spatial specificity of the enhancers may be – at least partially- encoded by its spacers.

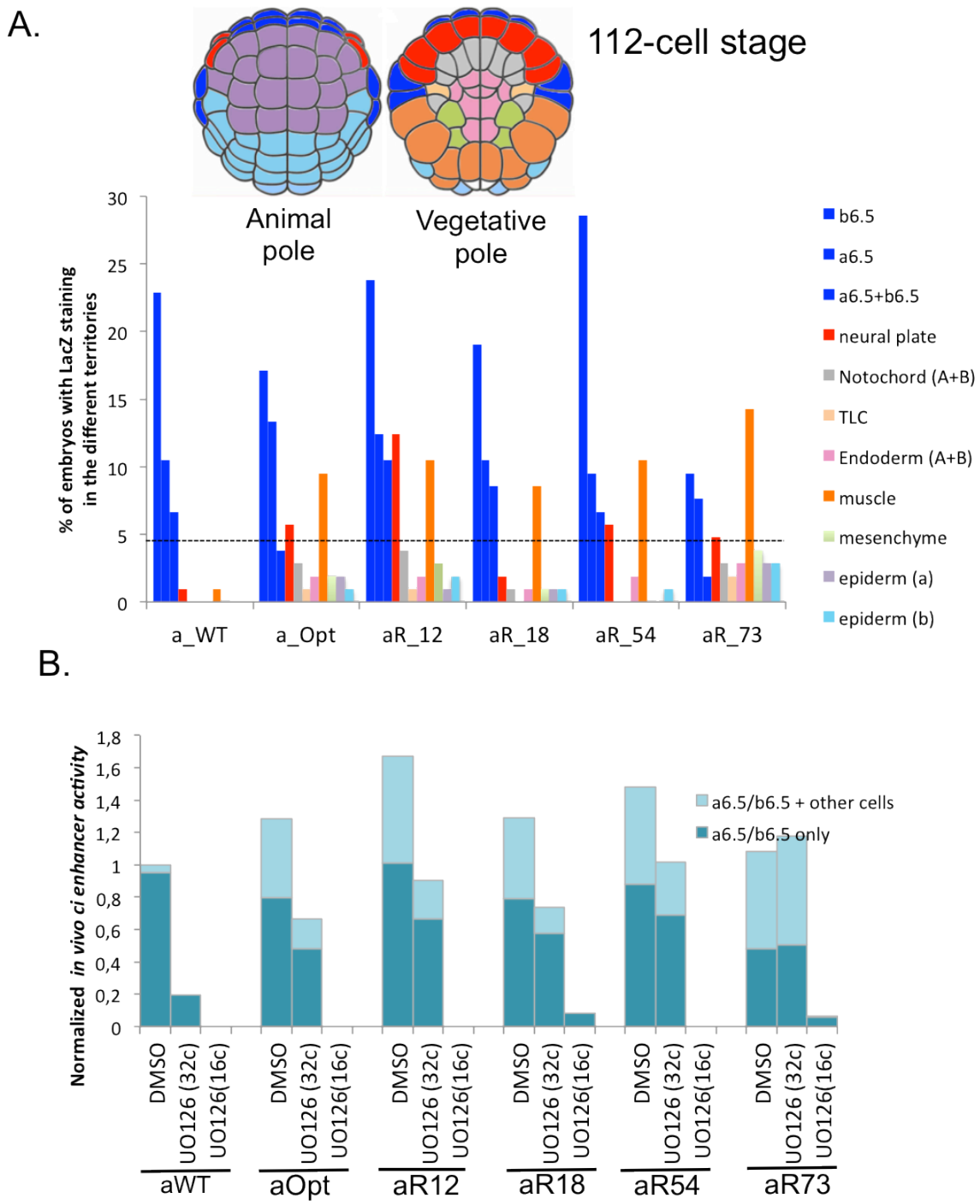


Figure 3. Ectopic expression and response to FGF. **A.** % of electroporated embryos with enhancer activity in the indicated territories. **B.** Relative enhancer activity of indicated variants with respect to the a-element in control conditions (DMSO) and in cases of complete and partial FGF signaling inhibition by the addition of the chemical inhibitor UO126 added at the 16- or 32 cell stage.

a-element randomized variants with weaker *in vivo* activity have decreased *in vitro* affinity for Ets and GATA

Transcription factors recognize their target sequences by reading both the DNA sequence and structural features of the double helix [12]. The failure of *in silico* prediction of TF binding affinity to reliably account for enhancer activity suggests that the spacers may affect transcription factor binding through a change in structural features of the helix.

We selected 9 randomized variants with equal or higher activity levels than the a-element and 9 variants with undetectable or very low activity (Figure 4A), and very similar *in silico* scores for ETS and GATA binding sites (Supplementary Figure 11). Using the quantitative multiple fluorescence relative affinity assay (QuMFRA) [18], we determined the relative *in vitro* binding affinities of the transcription factors for the whole enhancer (Supplementary Figure 12). Active variants showed significantly higher *in vitro* binding for a combination of Ets1/2 and Gata4/5/6 proteins than low activity variants (paired t-test, p=0.001) (Figure 3B). Both Ets1/2 and Gata4/5/6 proteins individually contributed to the binding selectivity to active enhancers, with a significantly higher contribution for Ets1/2 (Supplementary Figure 11B).

We conclude that spacer sequences affect the *in vitro* binding of Ets1/2 and Gata4/5/6 transcription factors on the enhancers. This could partly explain the wide range of enhancer activity levels obtained with the variants. The activity of the enhancers cannot, however, always be inferred from their *in vitro* affinity for the two transcription factors, as examplified by the comparison between *in vitro* binding to active aR_19 or aR_46 and inactive aR_24 (Figure 4B). One possibility is that spacer sequences of these variants may affect binding of nucleosomes or other transcriptional regulators.

Spacer sequences set the affinity of Ets1/2 to their binding sites

Next, we analyzed the contribution of the individual sites to the observed binding of Ets1/2 and Gata4/5/6 to the a-element. Farley and colleagues proposed that the activity of the enhancers is primarily determined by the two bases flanking either the “GGAA” or “GATA” core motif sequences [8]. We note, however, that in our experiments active variant aR_30 and inactive variant aR_70 share very similar 8 bp-extended ETS and GATA sites, suggesting that additional bases of the spacers contribute to the spacer effect (Supplementary Figure 13).

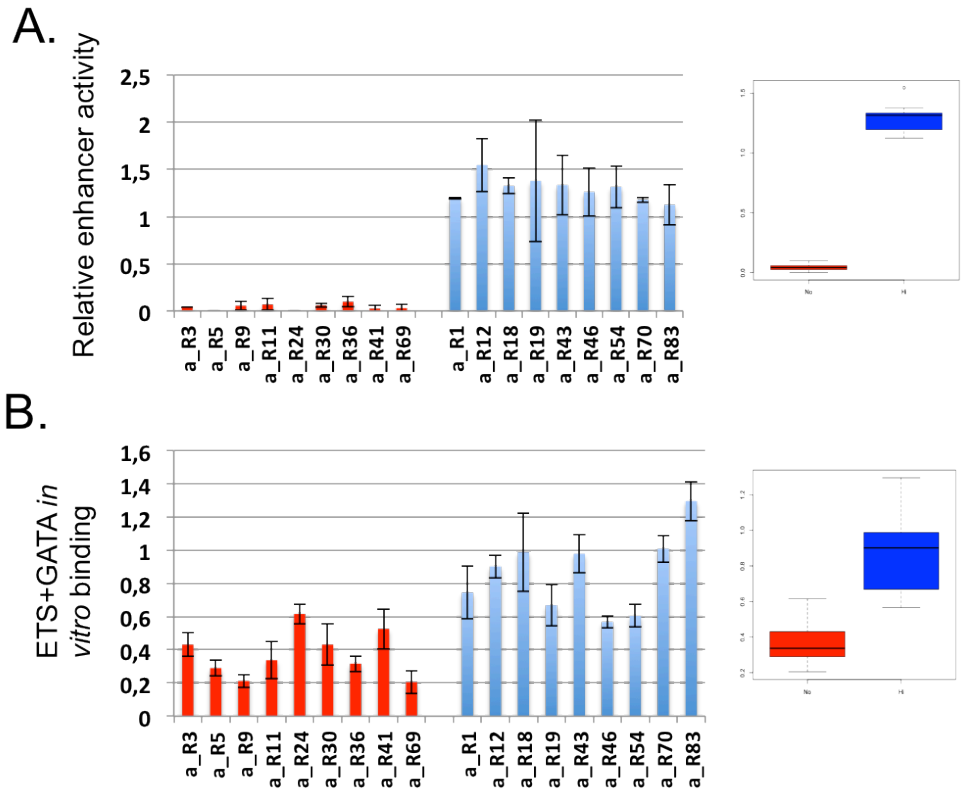


Figure 4. Active randomized enhancer variants have a higher *in vitro* affinity for Ets1/2 and Gata4/5/6. A. Relative *in vivo* enhancer activity compared to the WT a_element of inactive (red) and active variants (blue). The two populations are different (paired t-test, p=2.825e-09). **B.** Relative *in vitro* binding of Ets1/2 and Gata4/5/6 compared to the WT a_element of the same revertants as in (A). The two populations are different (paired t-test, p=0.001).

We first decided to keep 10 bp of the randomized variants centered on the 6bp unmodified binding sites and to complete the fragment used for the gel shift experiments with wild type neighboring sequences to obtain a 30mer (Figure 5C). These fragments were analyzed in gelshift experiments using the corresponding a_WT fragment as internal control. There is a slight, yet significantly better binding of Gata4/5/6 to their BS in the tested active enhancers (Figure 5B). Surprisingly, very different affinities were observed for Ets1/2 (Figure 5A). Importantly, sites with similar core octamers may have different affinities, for instance the core octamer "TAGGAAAT" is present in the active aR9_E2 and the inactive aR30_E2 sites, and the core "GCGGAAGG" is present in the active aR19_E1 and the inactive aR5_E1 and aR43_E1 sites. Interestingly, the identity of the base pairs at both ends of the decamer can influence the structure of the central octamer as suggested by modeling of the DNA helix shape (Supplementary Figure 14). Thus, immediate flanking sequences can affect binding of Ets1/2 to the octamer core, most likely by modifying the secondary structure of DNA double

helix. Further addition of spacer sequences at both sites of this decamer further modify the affinity of the ETS binding site (Supplementary Figure 15).

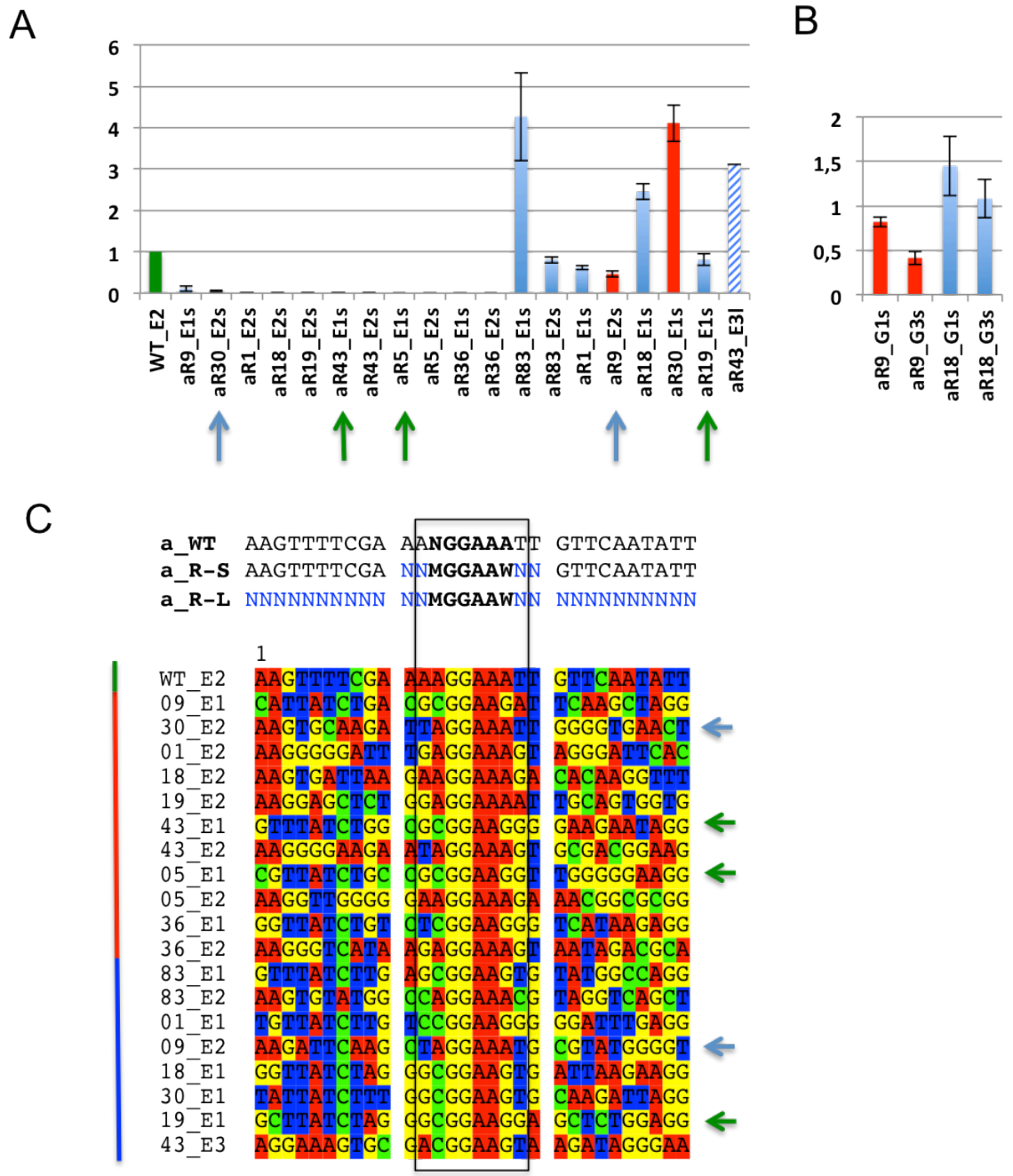


Figure 5 Sequences flanking the core octamer set the affinity for Ets1/2. **A.** Relative affinities of Ets1/2 for the BS represented in **(C)** compared to WT_E2. The new created BS is striped. BS with a similar octamer core are indicated with blue or green arrows. **B.** Relative affinities of Gata4/5/6 for the indicated BS compared to WT_G3. **C.** Upper part: fragments used in the gel shift experiments. Sequences from the WT a-element and randomized variants are in black and blue respectively. Lower part : 30bp windows centered on ETS BS. Variants with a similar octamer core are indicated with blue or green arrows. The octamer core is boxed. Color code used throughout the figure: Green, blue and red bars correspond to WT a-element, active and inactive variants respectively.

It is difficult to compare the affinity of individual sites to that of the whole enhancer fragment since the same internal control cannot be used. Nevertheless, when only one of the 2 isolated ETS sites is functional, only one band is shifted with the complete enhancer suggesting that the inactive site is not bound in the context of the whole enhancer neither (Supplementary Figure 12). Consistently, for aR18, where the two ETS sites are functional, two bands are observed in the context of the whole enhancer. Also, in aR43, E1 and E2 sites are not functional and a new ETS site E3 is created in the randomized spacer (Supplementary Figure 9) : only one band is observed in the gel shift with the complete enhancer (Figure 5A).

We conclude that although the two flanking base pairs can modify substantially the affinity of Ets1/2 for its core octamer BS, in most tested cases they cannot explain the observed affinity for the whole enhancer. Addition of more flanking sequences of the variants can further modify the affinity, demonstrating that even distant spacer sequences can affect binding of the TF to their binding site.

DISCUSSION

The low information content of eukaryotic TFBS leads to their overrepresentation in the genome. Clustering has been proposed as a solution for the resulting specificity issue [19]. Nevertheless, all clusters are not active [7], implicating additional constraints. These could be an inappropriate chromosomal environment [20] or highly local sequence features independent of the binding motifs [21]. Indeed, the affinity of TF for specific sites in DNA is controlled both by sequence readout and shape readout, the latter being less dependent on a very specific nucleotide sequence [22]. Since most predictive algorithms are based on sequence specificity, improvements in prediction have been reported when the local DNA structure is taken into consideration [23]. Here, we bring experimental evidence for the crucial role of spacer sequences in setting the activity of TFBS clusters, and suggest that this could involve optimal shaping of the DNA helix.

Using TFBS prediction from SELEX-seq data, we could confirm the findings of Khoueiry et al. that most clusters of predicted ETS and GATA binding sites have no early neural enhancer activity. It is unlikely that this is due to the chromatin environment since the short 55bp sequences were inserted in the same reporter constructs. Inappropriate spacing or orientation of the TFBS in the inactive clusters is also unlikely, since the organization of TFBS in the a-element of the *Otx* gene was shown to be flexible [7]. Two other not mutually exclusive possibilities remain to be considered: the TFBS were not optimal and/or are influenced by surrounding, so-called spacer sequences.

We analyzed the role of spacer sequences of the a-element of the *Otx* gene by keeping the 6 bp-core nucleotides of ETS and GATA BS and randomizing the spacer sequences. Surprisingly, we obtained a large range of enhancer activity levels. This could be reproduced on N26, one of the rare active clusters. The fact that the genomic cluster C53 remained inactive even when its binding site sequences were optimized was a strong argument to incriminate either the architecture of the TFBS or the spacer sequences. Strikingly, we could activate to variable extent not only the optimized C53 cluster but also another inactive cluster we attempted to.

There are a few reports in the literature of systematic large-scale analysis of enhancer variants, although in most cases single point mutations were tested, and the resulting effects are expected to be less severe. Nevertheless, the study of Kwasnieski et al. [24] on the 52 bp

rhodopsin enhancer is particularly interesting, since they noticed dramatic effects of base pair substitutions, even when occurring in spacer sequences. Less important effects of mutations were reported in the studies of Melnikov et al. [25] and Patwardhan et al. [26], mostly affecting the TFBS. The enhancers tested in the latter studies are larger than the a element and the *rhodopsin* enhancer, what could attenuate the impact of the point mutations on their activity.

Very similar to our work is the recent report of Farley et al. of the activity of randomized bar coded variants of the same *Otx* neural enhancer that we analyzed in *Ciona* [8]. The main difference in their approach being that only a four base pair core was maintained fixed for the 2 ETS and 3 GATA sites. They conclude that the essential information for enhancer activity is included in the octamer centered on the four bp core. Nevertheless, a closer analysis of their data indicates that sub-selections of their synthetic enhancers containing binding sites whose predicted *in vitro* activity is above a certain threshold, always contain inactive clusters and a wide range of activity levels. Thus, their data do not infirm our claim that spacer sequences set enhancer activity levels, even if they did not consider this option.

The most interesting conclusion of Farley et al. is that the TFBS are not optimized in order to assure cell specific expression. Accordingly, most of our tested randomized variants are more expressed in cell lineages were the WT a-element is only weakly expressed, and are also expressed in other lineages of the ectoderm. However, this extended activity pattern cannot only be explained by the global activity levels of the randomized variants in our experiments. One hint to a possible mechanism is suggested by the higher responsiveness of the spacer variants to FGF signaling: these enhancers could thus be activated by lower levels of active Ets1/2 in cells expressing Gata4/5/6. Noteworthy, most variants display one ETS BS with increased affinity, which should bind lower levels of Ets1/2.

One of the most surprising results of our work is the ease with which we can activate ETS and GATA clusters, which seems in apparent contradiction with the small proportion of active clusters in the *Ciona* genome. This raises the very intriguing possibility that negative selection is operating to avoid inappropriate activation of nearby genes, which should be detectable by micro-evolutionary analysis.

Spurious TFBS have been proposed to be non-functional decoy sites in eukaryotic cells [19].

Nevertheless, White et al. (2013) could show that binding of Crx to clustered motifs was dependent on highly local sequence features such as high GC content. Similarly, Parker et al. [27] provided evidence that the molecular shape of DNA is under selection in regulatory regions.
Our gel shift experiments further provide experimental evidence suggesting that neighboring base pairs can modify the shape of the DNA helix in the core octamer of ETS BS and thus influence their specific interaction with ETS protein. A similar observation was done for two yeast TF by Levo et al. [28]. GATA binding in contrast, seems to be less sensitive to local variations, which could indicate a greater flexibility of its zinc fingers with respect to the alpha-helices present in the helix-turn-helix motif of ETS.

Altered *in vitro* binding of Ets1/2 and Gata4/5/6 to their binding sites cannot provide an exclusive explanation for the variability in enhancer activity of the randomized variants. In at least one case we could show that a newly created ETS site is functional. Also, some inactive variants are still binding Ets1/2 and Gata4/5/6 as well as active ones. Binding sites for other factors could also be created in the randomized spacer sequences, including factors competing for Ets1/2 or Gata4/5/6 binding or repressors of transcription.

Surprisingly, most of the active enhancers we tested have only one ETS site bound *in vitro* and one variant binds Ets1/2 only very weakly. Although we did not find good evidence for cooperativity *in vitro*, it cannot be excluded that the pioneer factor Gata4/5/6 facilitates binding of Ets1/2 and/or binding of one Ets1/2 facilitates binding of the second molecule. This should be tested by appropriate mutagenesis experiments. An intriguing way spacer sequences could also influence enhancer activity is by facilitating allostery through DNA, whereby fixation of one TF can facilitate binding of a TF at distance [29].

To conclude, the presence of a minimal cluster of TFBS is not sufficient to make an active enhancer, the nature of the spacer sequences also has to be considered. One way through which these could act is by setting the affinity of TF for their cognate sites by optimizing DNA-protein contacts through shaping the DNA helix. Progress in modeling DNA shape should greatly help in enhancer prediction.

MATERIALS AND METHODS

Embryo experiments and scoring

Mature *Ciona intestinalis* (type B) were provided by the Roscoff Marine Biological station and maintained in natural sea water at 16°C under constant illumination. Eggs were collected, fertilized and dechorionated as previously described [5].

Electroporation was performed as previously described [5] using the following parameters: 50µg DNA in 50µl H20 + 200µl D-Mannitol 0,96M ; 50V-16ms pulse, using a Electro Square Porator machine (BTX T820; Harvard Apparatus). Embryos were grown in 0.1% gentamycin ASWH (Artificial Sea Water with Hepes) until their harvest at the 112-cell stage collection. Fixation and LacZ staining were performed as described in [5].

Where indicated, embryos were treated with a final concentration of 10µM U0126 from the early 16-cell stage or the early 32-cell stage [30]. Control embryos were treated with the same amount of DMSO, added at the same time point (3µl DMSO in 15ml ASWH per plate). All experiments presented were at least repeated once.

At least 100 electroporated embryos where scored for each experiment by counting the % of embryos stained with LacZ in each territory, as the level of activity in a given cell lineage is considered to be a function of the % of embryos in which X-gal staining is detected in this cell lineage [31]. For each embryo, the following information was retrieved: staining in a6.5 and/or b6.5 lineages, staining in other *Otx* expressing lineages (muscle, a6.7 cell lineage), activity in territories not expressing *Otx* was only detailed in experiments shown in Figure 3. Enhancer variants driving detectable LacZ expression in less than 5% of stained embryos in all experiments was considered inactive. all values were normalized to the WT a-element activity electroporated in each experiment.

To simplify the system, these experiments were carried out using a modified version of the minimal wild-type element, in which the G2 binding site (Khoueiry et al., 2010) is innactivated (X-->X). This mutation has neither qualitative nor quantitative impact on the activity of the element.

Gene IDs

Ciona intestinalis Otx : Gene model ID KH.C4.84 (Unique gene ID: Ciinte.g00006940)

Ciona intestinalis Ets1/2: Gene model ID KH.C10.113 (Unique gene ID: Ciinte.g00001309)

Ciona intestinalis Gata4/5/6: Gene model ID KH.L20.1 (Unique gene ID: Ciinte.g00012060)

Construct design and molecular cloning

All these experiments were carried out using a modified version of the minimal wild-type element, in which a weak GATA binding site (G2, Khoueiry et al., 2010) is mutated without quantitative or qualitative impact on the activity of the element.

Point mutations in ETS and GATA binding sites motifs

The family of GATA TFs preferentially binds the consensus "HGATAR" (H = A, C or T) (Merika M and Orkin SH, 1993; SELEX data from the lab). Therefore, 12 variants of the a-element harboring all variants of HGATAN at the third GATA position were designed to test both consensus (HGATAR) and non-consensus (HGATAY, Y = C or T) binding site motifs. The ETS family of TFs preferentially binds the consensus site "MMGGAWR" (M = A or C; W = A or T; R = G or A), though with a higher affinity for "CCGGAWR" (Boros et al.; 2009; Macleod et al., 1992; Wasylyk et al., 1992), which is consistent with *Ciona* SELEX data (Nitta et al., in preparation). "T" at the seventh position was tested as negative control. 21 variants of the a-element were tested harboring the different combinations MMggaWD (D = A, G or T) at the second ETS site, with the exception of the MMggaTA combinations as they create an additional overlapping GATA site that could interfere with the E2 activity and make the interpretation of the results not straightforward.

Oligos were synthesized containing part of Gateway attB1 and attB2 recombination sites in 5' and 3' respectively of the different elements we tested. These oligos were amplified by PCR using attB1F and attB2R primers, then inserted by successive BP and LR reactions in pDONOR221_P1-P2 and pDEST-L1-RFA-L2-bpFOG-LacZ (Roure 2007).

Randomized variants

Only six bases per TFBS were kept from the WT sequences, centered on "GGAA" and "GATA" for ETS and GATA binding sites respectively for the a-element variants. 7 bases were kept in randomized N26, N61 and C53, as it appeared the 1st base of ETS site is important for *in vitro* binding.

A unique transcribed barcode was added to the other studied constructs between bpFOG and LacZ in order to be able to do a large scale screening if needed. We showed that the sequence of this barcode has no effect on LacZ reporter activity (data not shown). Two nested PCRs were done to amplify the 5' end of the insert containing an attB1 site, the sequence studied for its enhancer activity and the 5' end of bpFOG using primers attB1F and P5 R first, then attB1 and P4 R. Three nested PCRs were performed to amplify the 3' end of the insert containing the other half of bpFOG, the barcode and an attB2 site, using primers P1 F and attB2 R first,

then P2 F and attB2 R then P3 F and attB2 R. Both fragments were assembled in a last PCR using attB1 F and attB2 R. They were then inserted in pDEST-L1-RFB-L2-LacZ thanks to a one step BP-LR reaction (Roure 2007).

in silico TFBS prediction using MOTIF

6-mer enrichment measures form SELEX-seq data

In a random set of oligos, k-mer frequencies will be distributed uniformly. If the oligos are not random, as with the HT-SELEX method, the k-mer frequency distribution will become skewed. The enriched k-mers with frequencies greater than expected can be used to determine the efficacy of the SELEX process. K-mer frequencies are determined by counting the occurrences of each k-mer in the set of unique oligos. 6-mers are used for quality control since 4,096 k-mers provides a sufficient number of k-mers without becoming sparse and unwieldy. The observed count of k-mers in the sequenced oligos, *obs,* are normalized using the expected count, *exp*, of each k-mer based on the number of sequenced oligos, *n*, with a random base size of *d*, as shown in equation 1.

$$exp = \frac{n * (d - k)}{4^k} \quad (1)$$

The enrichment score, *e,* was calculated as shown in equation 2.

$$e = \log_{10}\left(\frac{obs}{exp}\right) \quad (2).$$

The synthesis method used to produce the original random pool of oligos is often biased, enriching certain k-mers over others. To correct for this bias, the enrichments are adjusted by the enrichment in the background set, shown in equation 3.

$$e_{adj} = e_{raw} - e_{background} \quad (3)$$

Calculation of TFBS scores by MOTIF

For any transcription factor whose SELEX-seq data is available, MOTIF can associate to each base of the analyzed DNA sequence a score reflecting the *in vitro* binding of this TF on the sequence. This score is calculated for all consecutive 8-mers, and allocated to their first base (Supp. Figure 3). It corresponds to the sum of the 6-mer enrichments measures of the three 6-mers contained in each 8-mers.

14 tested ETS/GATA genomic clusters

101 clusters containing at least 2 sites ETS and 2 sites GATA were identified in *Ciona intestinalis* genome, using SECOMOD, and a very relaxed consensus for the TFBS sequences (as described in Khoueiry 2010). We looked for clusters of maximum 140pb, non-conserved in *Ciona savignyi* genome, with at least 5bp between 2 TFBS .55 conserved such clusters were identified by Khoueiry et al.. . We tested the activit of 8 non conserved and 6 conserved clusters

The 14 tested clusters contain 2 ETS and GATA sites predicted to be "good" according to their high MOTIF scores. Their sequences are in Supp. Table 1.

2-colour Fluorescent EMSA

The DNA binding domains of Ets1/2 (Ensembl ID: ENSCINT00000011848), i.e. aa 581-708, and Gata4/5/6 (Ensembl ID: ENSCINP00000009159), i.e. aa 291-415, were determined by homology to domains of orthologous human proteins used in the cristalographic 3D structure determination (Ets1, MMDB ID: 62790 and GATA1, MMDB ID: 106606). The corresponding DNA sequences were amplified by PCR from the cDNAs and cloned in the expression vector pETG20A (EMBL Protein Expression and Purification Facility) by Gateway technology (Life Sciences). N-terminally poly-His-thioredoxin tagged recombinant proteins were produced in Rosetta-pLys-R strain and purified on Nickel Agarose columns as described in [32].

Enhancer DNA fragments were produced by PCR from the plasmids used for the ß-galactosidase reporter assays using Cy5 or Alexia 488-5' labeled 19 nt primers (MWG Eurofins) flanking the enhancer sequences, i.e. TTGTACAAAAAAGCAGGCT for the forward and GGTACAATACACGAAGCTT for the reverse primer. DNA fragments containing unique TFBS were synthesized directly (MWG Eurofins) and 5'-terminally labeled with Cy5 or Cy3 for the internal control on one strand.

The reaction conditions for the GS experiments were adapted from [33]. Labeled DNA was incubated at 0.015 µM with recombinant Ets1/2 at 0.2 µM or GATAa at 0.1 µM during 15 minutes at room temperature in 25mM Hepes pH7.9, 50mM KCl, 0.5 mM EDTA, 10% glycerol, 0.5mM di-thiothreitol and 100 µg/ml poly(dI-dC) and loaded on a 6% polyacrylamide gel in 0.5 % TAE, which was run at 10V/cm. The fluorescence was registered with an Amersham Imager 600 (General Electric) and quantified with the software provided by the supplier.

To have a better control over the experimental conditions we included an internal control: the randomized DNA fragments are fluorescently labeled with Cy5 and mixed with an equimolar amount of control DNA fragment labeled with Alexia 488 or Cy3. Relative affinities Y are quantified by reporting the fraction of shifted randomized DNA fragments to that of control fragment [18] (suppl Figure 12).

REFERENCES

[1] I. Miguel-Escalada, L. Pasquali, and J. Ferrer, "Transcriptional enhancers: functional insights and role in human disease," *Curr. Opin. Genet. Dev.*, vol. 33, pp. 71–76, Sep. 2015.

[2] A. T. Douglas and R. D. Hill, "Variation in vertebrate cis-regulatory elements in evolution and disease," *Transcription*, vol. 5, no. 3, p. e28848, 2014.

[3] M. J. Borok, D. A. Tran, M. C. W. Ho, and R. A. Drewell, "Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila.," *Development*, vol. 137, no. 1, pp. 5–13, Jan. 2010.

[4] R. P. Smith, L. Taher, R. P. Patwardhan, M. J. Kim, F. Inoue, J. Shendure, I. Ovcharenko, and N. Ahituv, "Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model," *Nat. Genet.*, vol. 45, no. 9, pp. 1021–1028, Sep. 2013.

[5] V. Bertrand, C. Hudson, D. Caillol, C. Popovici, and P. Lemaire, "Neural tissue in ascidian embryos is induced by FGF9/16/20, acting via a combination of maternal GATA and Ets transcription factors.," *Cell*, vol. 115, no. 5, pp. 615–27, Nov. 2003.

[6] U. Rothbächer, V. Bertrand, C. Lamy, and P. Lemaire, "A combinatorial code of maternal GATA, Ets and beta-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm.," *Development*, vol. 134, no. 22, pp. 4023–32, Nov. 2007.

[7] P. Khoueiry, U. Rothbächer, Y. Ohtsuka, F. Daian, E. Frangulian, A. Roure, I. Dubchak, and P. Lemaire, "A cis-regulatory signature in ascidians and flies, independent of transcription factor binding sites," *Curr. Biol. CB*, vol. 20, no. 9, pp. 792–802, May 2010.

[8] E. K. Farley, K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar, and M. S. Levine, "Suboptimization of developmental enhancers," *Science*, vol. 350, no. 6258, pp. 325–328, Oct. 2015.

[9] J. O. Yáñez-Cuna, C. D. Arnold, G. Stampfel, L. M. Boryń, D. Gerlach, M. Rath, and A. Stark, "Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features.," *Genome Res.*, vol. 24, no. 7, pp. 1147–56, Jul. 2014.

[10] I. Dror, T. Golan, C. Levy, R. Rohs, and Y. Mandel-Gutfreund, "A widespread role of the motif environment on transcription factor binding across diverse protein families.," *Genome Res.*, vol. 25, no. 9, pp. 1268–80, Jul. 2015.

[11] R. Gordân, N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs, and M. L. Bulyk, "Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape," *Cell Rep.*, vol. 3, no. 4, pp. 1093–1104, Apr. 2013.

[12] M. Slattery, T. Zhou, L. Yang, A. C. Dantas Machado, R. Gordân, and R. Rohs, "Absence of a simple code: how transcription factors read the genome," *Trends Biochem. Sci.*, vol. 39, no. 9, pp. 381–399, Aug. 2014.

[13] Y. Chen, D. L. Bates, R. Dey, P.-H. Chen, A. C. D. Machado, I. A. Laird-Offringa, R. Rohs, and L. Chen, "DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation," *Cell Rep.*, vol. 2, no. 5, pp. 1197–1206, Nov. 2012.

[14] D. L. Bates, Y. Chen, G. Kim, L. Guo, and L. Chen, "Crystal structures of multiple GATA zinc fingers bound to DNA reveal new insights into DNA recognition and self-association by GATA.," *J. Mol. Biol.*, vol. 381, no. 5, pp. 1292–306, Sep. 2008.

[15] M. H. Werner, G. M. Clore, C. L. Fisher, R. J. Fisher, L. Trinh, J. Shiloach, and A. M. Gronenborn, "Correction of the NMR structure of the ETS1/DNA complex.," *J. Biomol. NMR*, vol. 10, no. 4, pp. 317–28, Dec. 1997.

[16] K. S. Imai, M. Levine, N. Satoh, and Y. Satou, "Regulatory blueprint for a chordate embryo.," *Science*, vol. 312, no. 5777, pp. 1183–1187, 2006.

[17] N. Ohta and Y. Satou, "Multiple signaling pathways coordinate to induce a threshold response in a chordate embryo.," *PLoS Genet.*, vol. 9, no. 10, p. e1003818, Oct. 2013.

[18] T. K. Man and G. D. Stormo, "Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay.," *Nucleic Acids Res.*, vol. 29, no. 12, pp. 2471–2478, 2001.

[19] Z. Wunderlich and A. M. Leonid, "Different gene regulation strategies revealed by analysis of binding motifs," *Trends Genet.*, vol. 25, no. 10, pp. 429–34, Oct. 2009.

[20] J. Yan, M. Enge, T. Whitington, K. Dave, J. Liu, I. Sur, B. Schmierer, A. Jolma, T. Kivioja, M. Taipale, and J. Taipale, "Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites.," *Cell*, vol. 154, no. 4, pp. 801–13, Aug. 2013.

[21] M. a. White, C. a. Myers, J. C. Corbo, and B. a. Cohen, "Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks," *Proc. Natl. Acad. Sci.*, pp. 1–6, Jul. 2013.

[22] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann, "Origins of specificity in protein-DNA recognition.," *Annu. Rev. Biochem.*, vol. 79, pp. 233–69, Jan. 2010.

[23] T. Zhou, N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordân, and R. Rohs, "Quantitative modeling of transcription factor binding specificities using DNA shape," *Proc. Natl. Acad. Sci.*, p. 201422023, 2015.

[24] J. C. Kwasnieski, I. Mogno, C. a Myers, J. C. Corbo, and B. a Cohen, "Complex effects of nucleotide variants in a mammalian cis-regulatory element.," *Proc. Natl. Acad. Sci. U. S. A.*, pp. 2–7, Nov. 2012.

[25] A. Melnikov, A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. Callan, J. B. Kinney, M. Kellis, E. S. Lander, and T. S. Mikkelsen, "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.," *Nat. Biotechnol.*, vol. 30, no. 3, pp. 271–277, Feb. 2012.

[26] R. P. Patwardhan, J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S.-I. Lee, G. M. Cooper, N. Ahituv, L. a Pennacchio, and J. Shendure, "Massively parallel functional dissection of mammalian enhancers in vivo.," *Nat. Biotechnol.*, vol. 30, no. 3, pp. 265–270, Feb. 2012.

[27] S. C. J. Parker, L. Hansen, H. O. Abaan, T. D. Tullius, and E. H. Margulies, "Local DNA topography correlates with functional noncoding regions of the human genome.," *Science*, vol. 324, no. 5925, pp. 389–392, 2009.

[28] M. Levo, E. Zalckvar, E. Sharon, A. Carolina, D. Machado, Y. Kalma, M. Lotam-pompan, A. Weinberger, Z. Yakhini, R. Rohs, and E. Segal, "Unraveling determinants of transcription factor binding outside the core binding site," pp. 1–12, 2015.

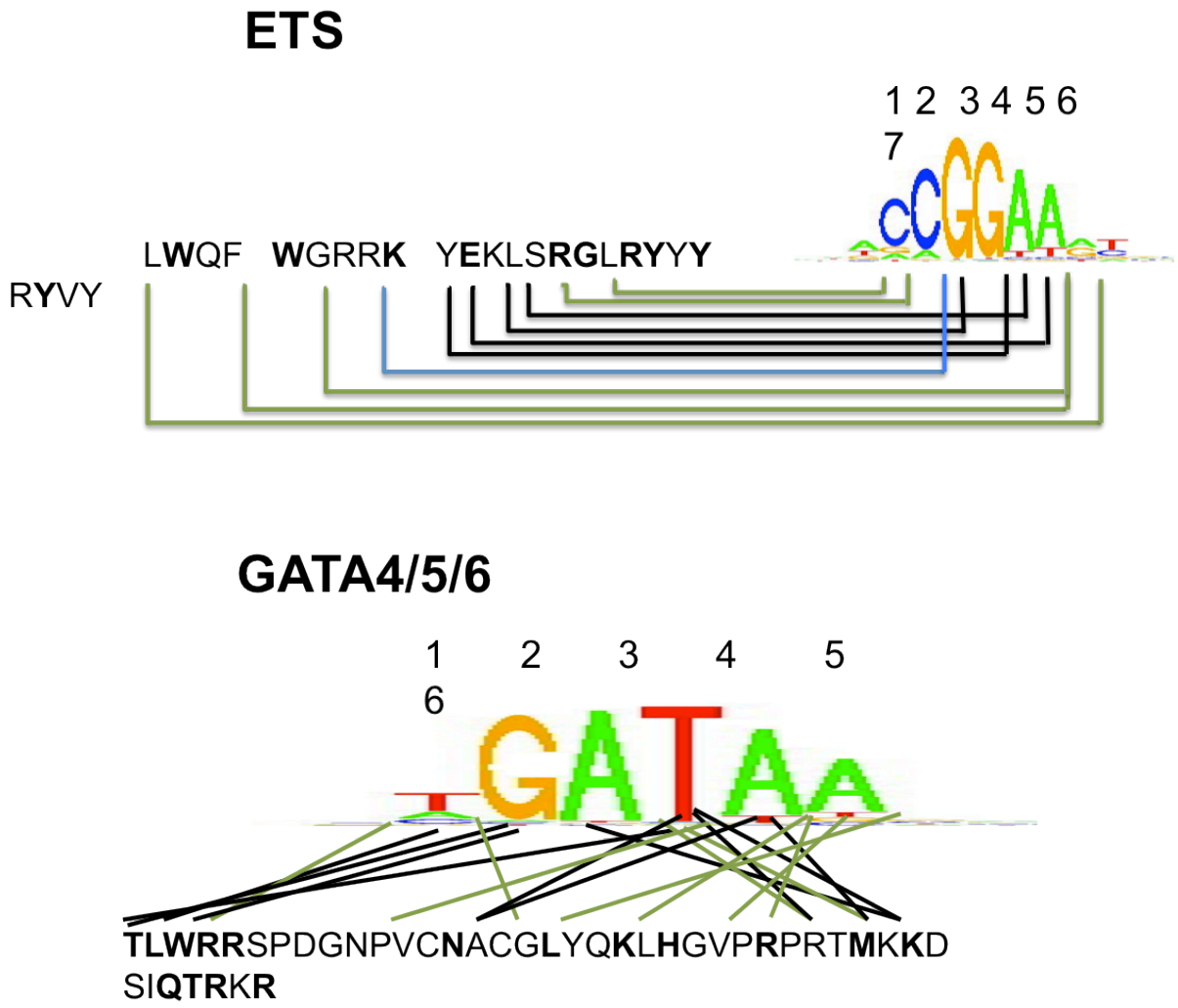
[29] S. Kim, E. Broströmer, D. Xing, J. Jin, S. Chong, H. Ge, S. Wang, C. Gu, L. Yang, Y. Q. Gao, X. Su, Y. Sun, and X. S. Xie, "Probing allostery through DNA.," *Science*, vol. 339, no. 6121, pp. 816–9, 2013.

[30] C. Hudson, S. Darras, D. Caillol, H. Yasuo, and P. Lemaire, “A conserved role for the MEK signalling pathway in neural tissue specification and posteriorisation in the invertebrate chordate, the ascidian Ciona intestinalis,” *Development*, vol. 130, no. 1, pp. 147–59, Jan. 2003.

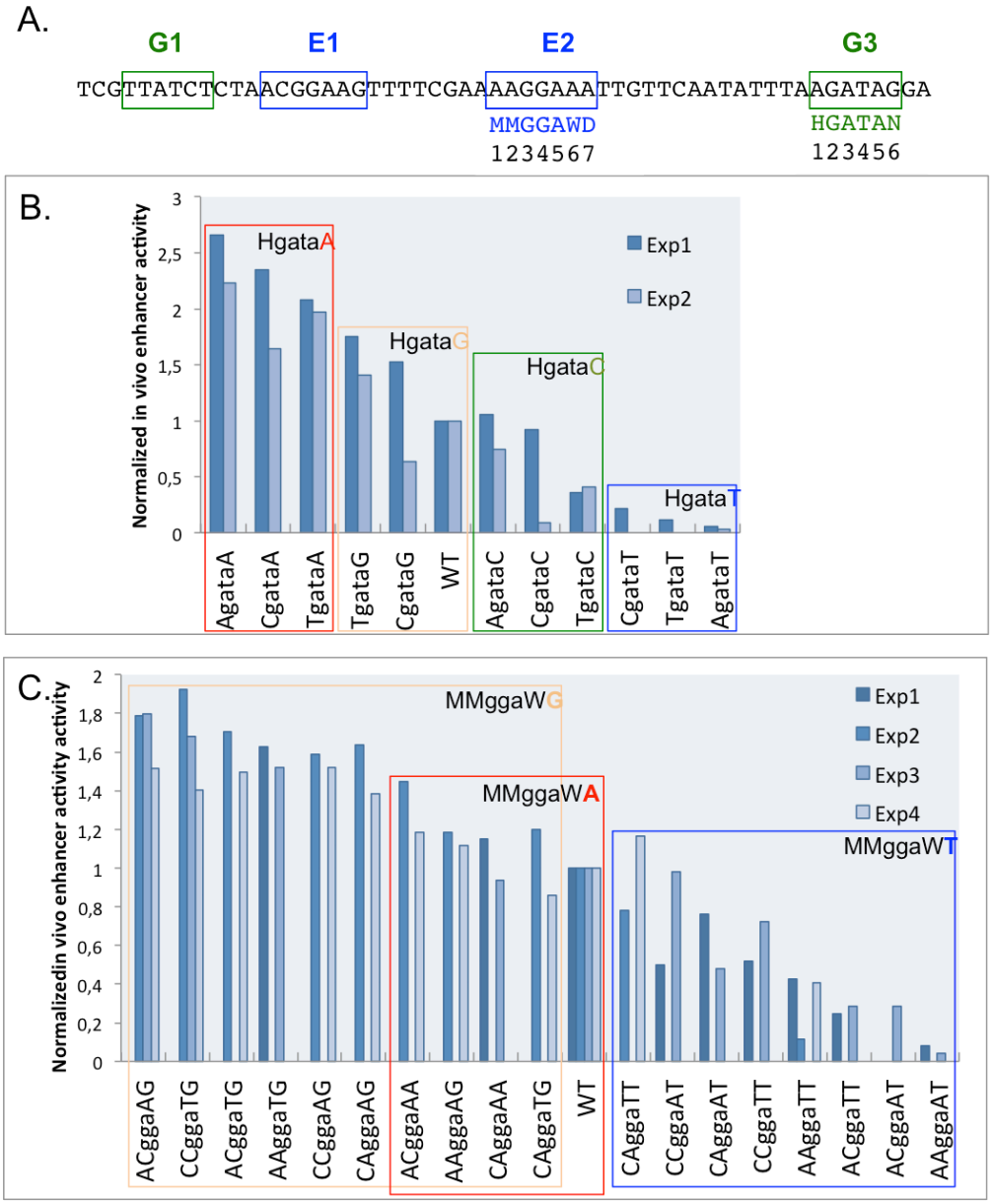
[31] C. D. Brown, D. S. Johnson, and A. Sidow, “Functional architecture and evolution of transcriptional elements that drive gene coexpression.,” *Science*, vol. 317, no. 5844, pp. 1557–60, Sep. 2007.

[32] R. Vincentelli, A. Cimino, A. Geerlof, A. Kubo, Y. Satou, and C. Cambillau, “High-throughput protein expression screening and purification in Escherichia coli,” *Methods*, vol. 55, no. 1, pp. 65–72, 2011.

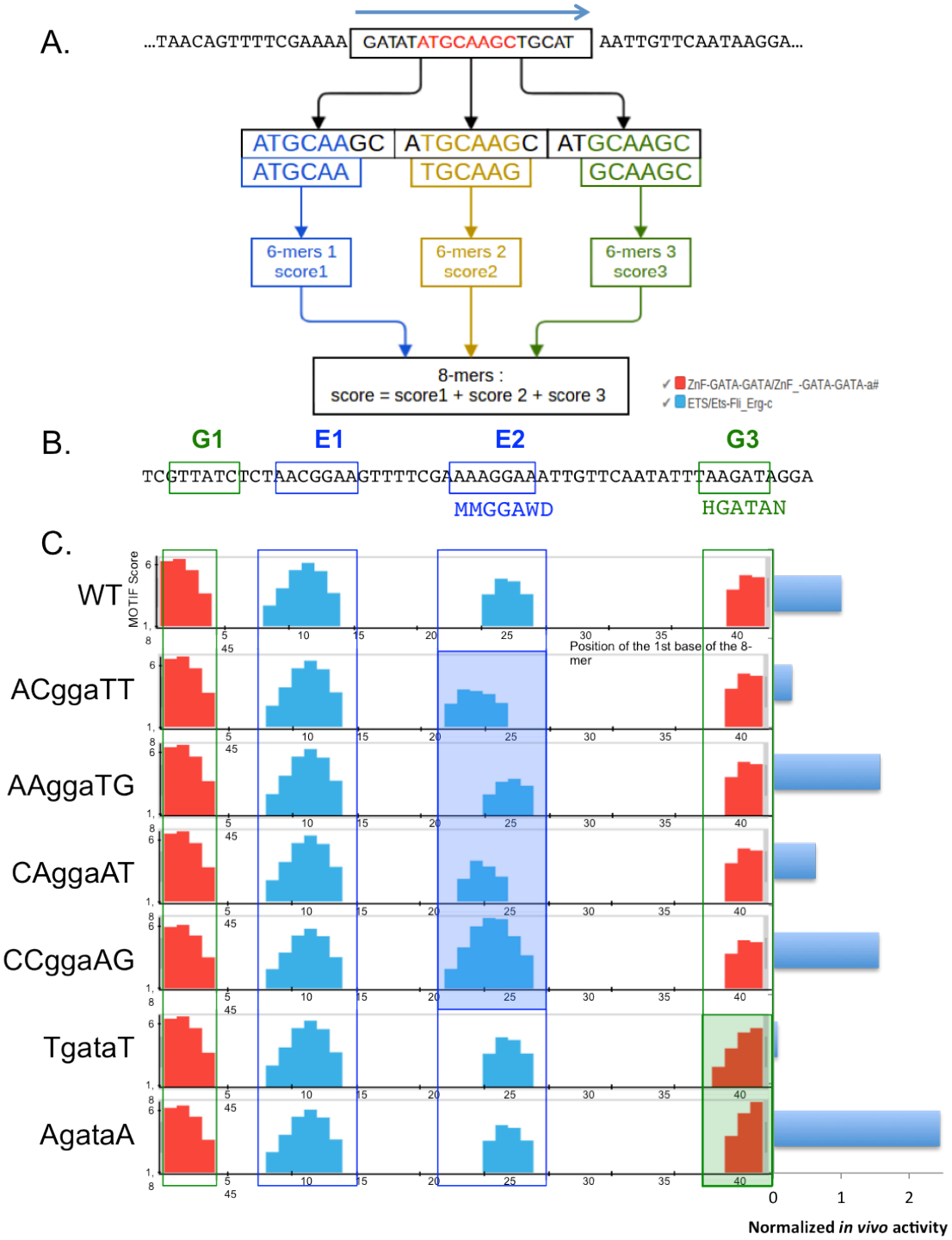
[33] Y. Hashimoto and J. Ware, “Identification of essential GATA and Ets binding motifs within the promoter of the platelet glycoprotein Ibα gene,” *J. Biol. Chem.*, vol. 270, no. 41, pp. 24532–24539, 1995.



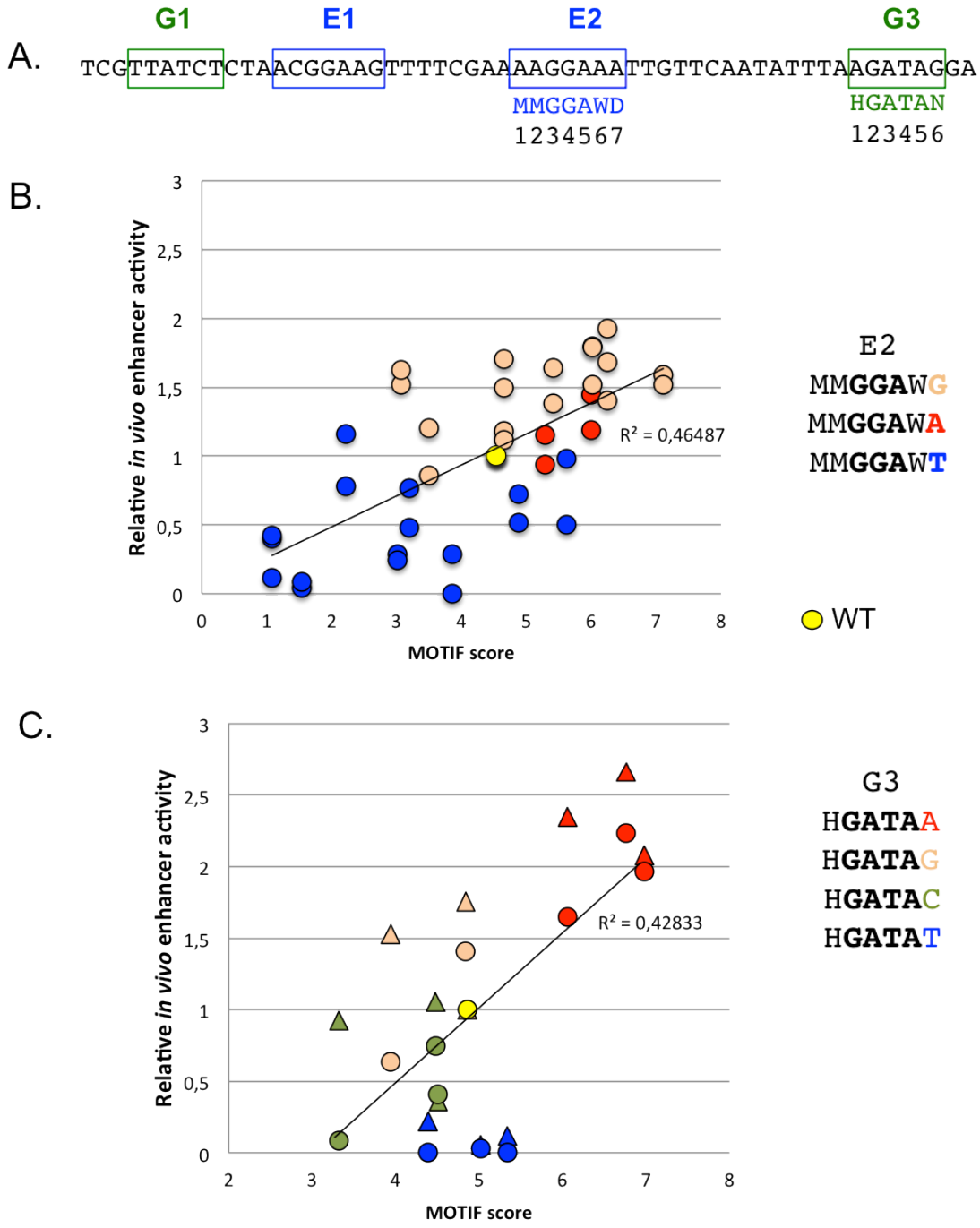
Supplementary Figure 1: DNA residues contacted by Ets and Gata4/5/6. The PWM is deduced from the SELEX-seq data. Contacts are deduced from the crystal structure-data compiled by Wei et al. on mammalian Ets1 and the crystalstructure of human GATA3-DNA complex [13]. Conserved amino-acids in contact with DNA are in bold, contacts with the base pairs are in black, with water molecules in blue and with the sugar-phosphate backbone in green.



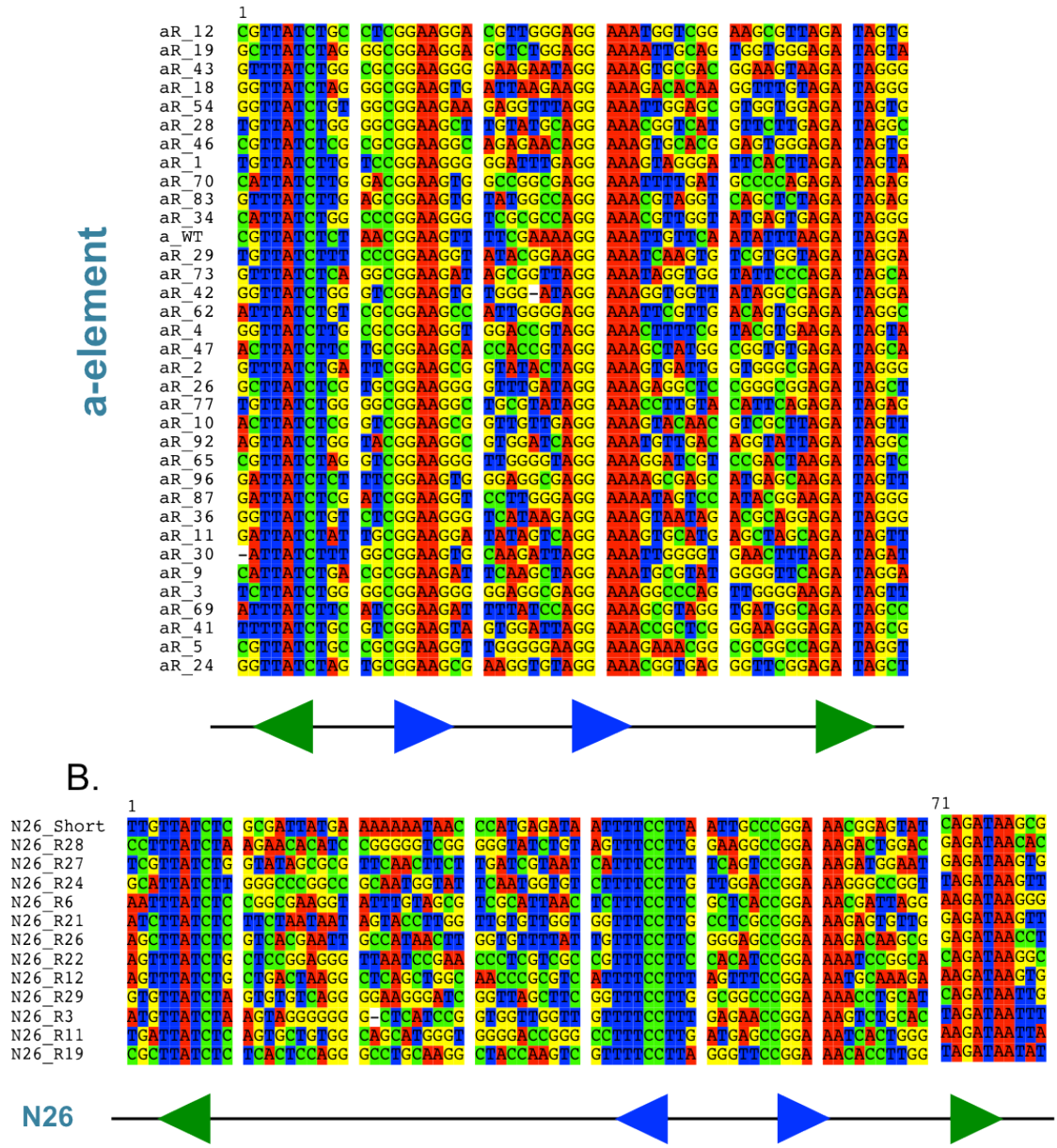
Supplementary Figure 2: **Individual experiments of binding sites point mutations effect on *in vivo* enhancer activity**. **A**. a-element sequence describing ETS and GATA site mutations tested *in vivo*. **B.** All GATA site variants were tested in two individual experiments. **C.** All ETS site variants were tested at least twice in four individual experiments. Each individual experiment was normalized to the activity level of the WT. The order of the tested constructs according to their activity levels slightly varies between individual experiments, but the last base of the sites globally appears to be a good indicator of the activity range (coloured boxes).



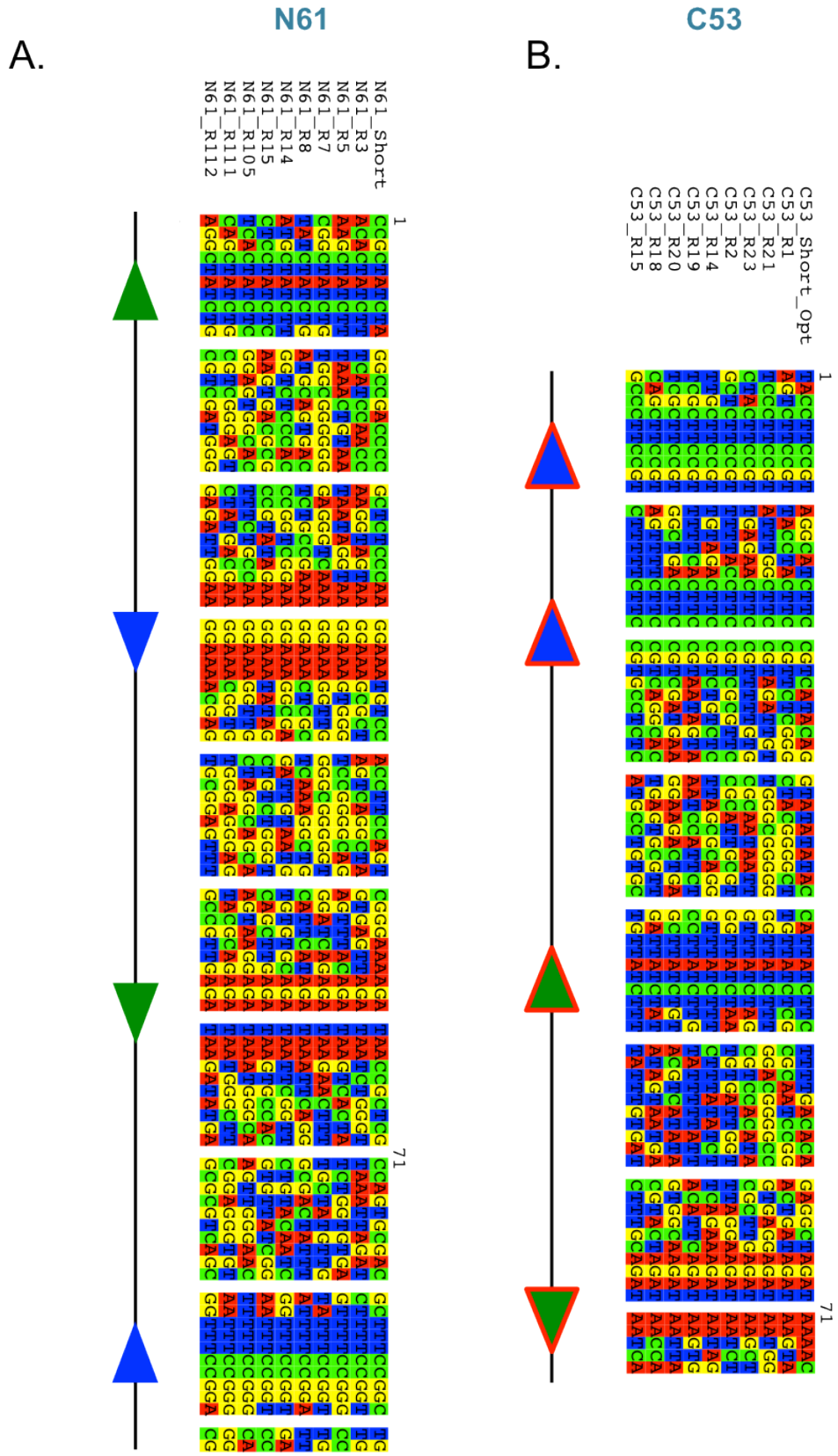
Supplementary Figure 3: *In silico* binding affinity calculations with MOTIF and exemples of scoring profiles. **A.** MOTIF scans a DNA sequence, and associates to each 8-mers the sum of the three 6-mers enrichment scores it contains. This score is associated to the first base of each 8-mer. **C.** Exemples of MOTIF scoring profiles and *in vivo* enhancer activity of ETS and GATA sites in the WT a-element and variants with mutated E2 or G3 (described in **B**.). Interestingly, enhancer activity levels cannot be predicted from their *in vitro* predicted binding.



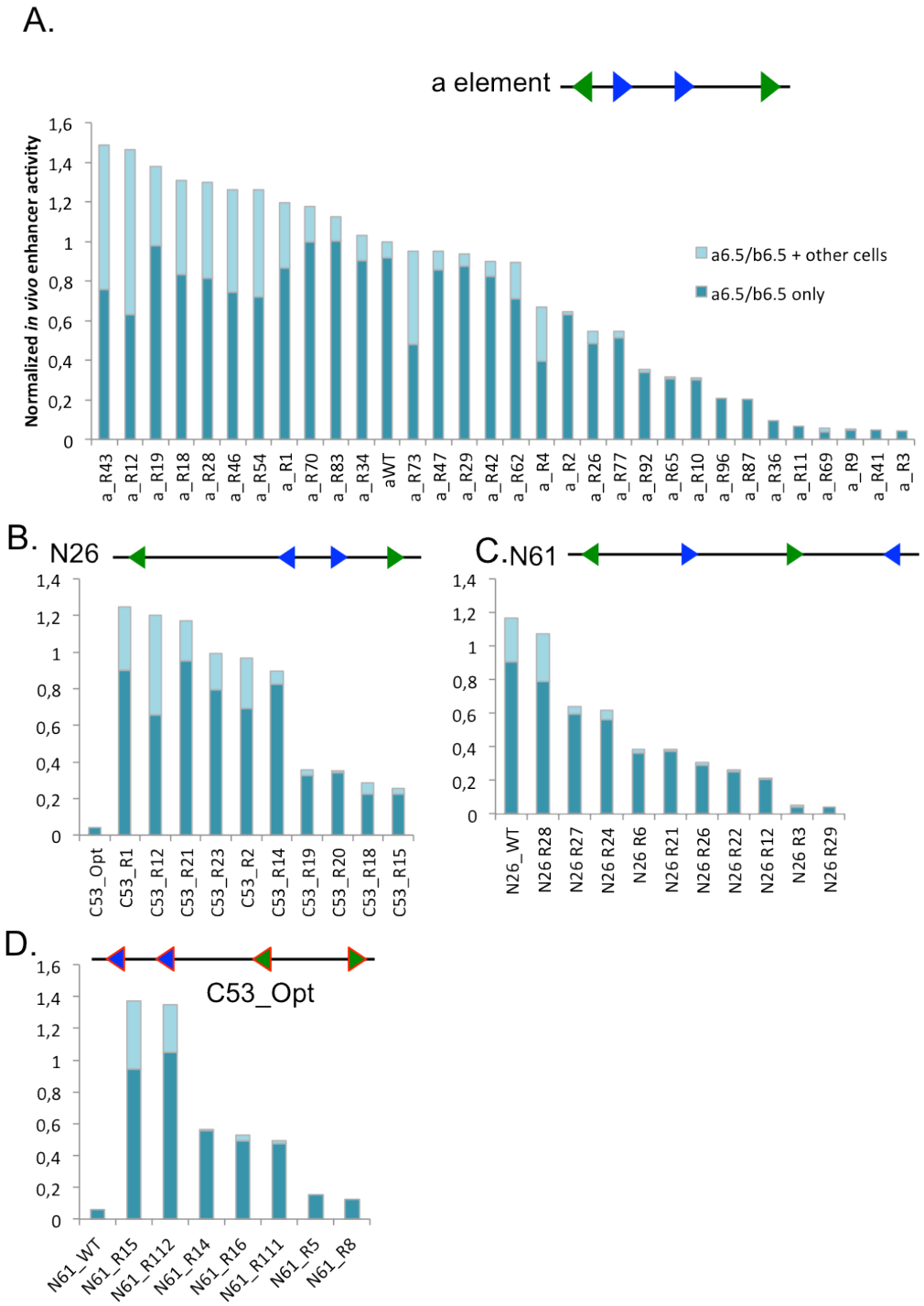
Supplementary Figure 4: Poor correlation between *in silico* binding and *in vivo* activity. **A**. a-element sequence describing ETS (blue) and GATA (green) site mutations tested *in vivo*. **B.** Comparison of *in vivo* enhancer activity and *in silico* predicted binding calculated for E2 (MOTIF score). Each point corresponds to an ETS site variant individual *in vivo* experiment. **C.** Comparison of *in vivo* enhancer activity and *in silico* predicted binding calculated for G3. Each point corresponds to a GATA site variant individual *in vivo* experiment (triangles and circles respectively correspond to experiments 1 and 2 from Supp. Figure 2). Colours correspond to the last base and yellow circles correspond to the WT a element.



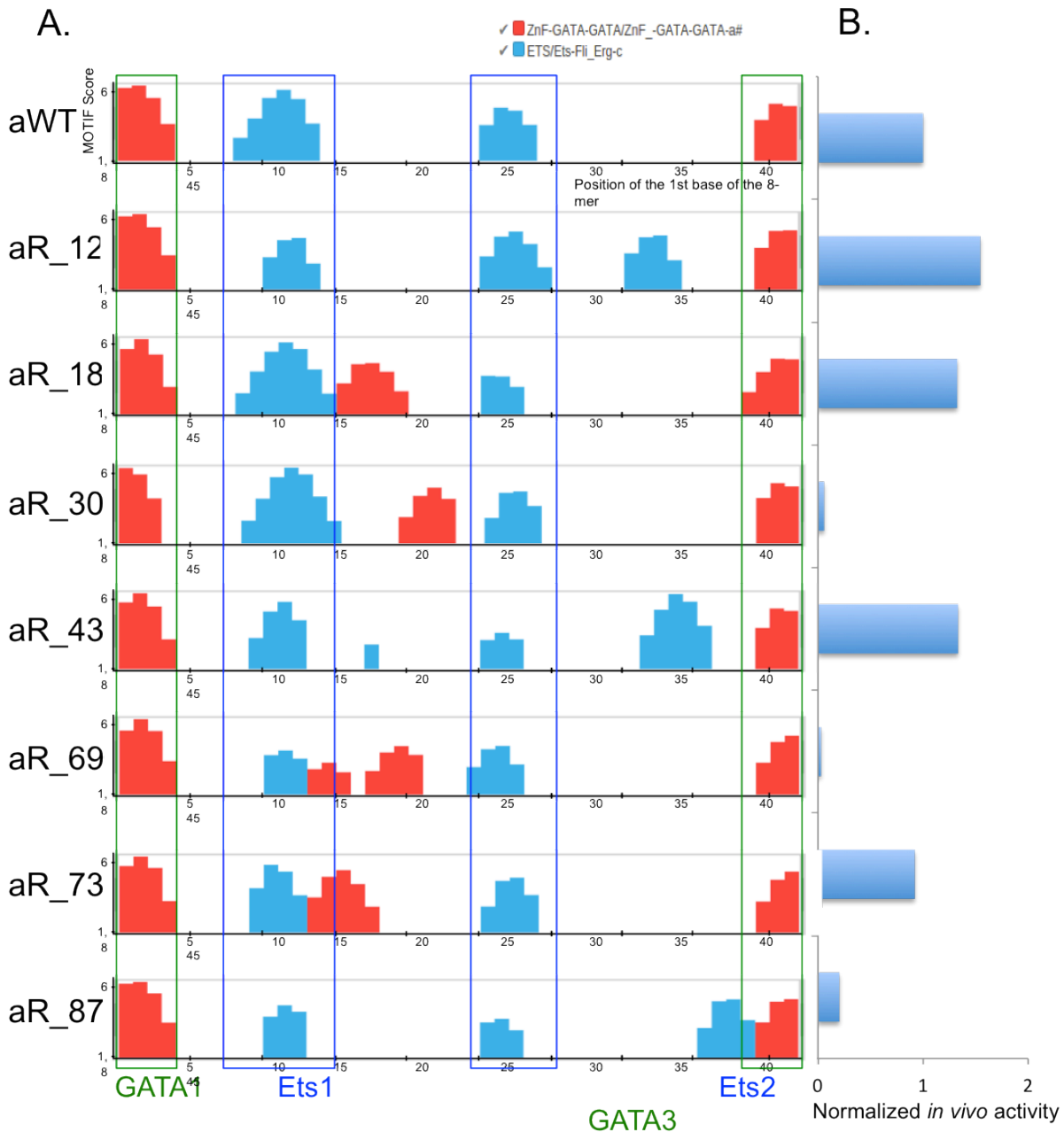
Supplementary Figure 5: DNA sequence alignment of randomized variants for the a-element and the active N26 cluster. ETS and GATA binding sites are represented by blue, respectively green, arrowheads.



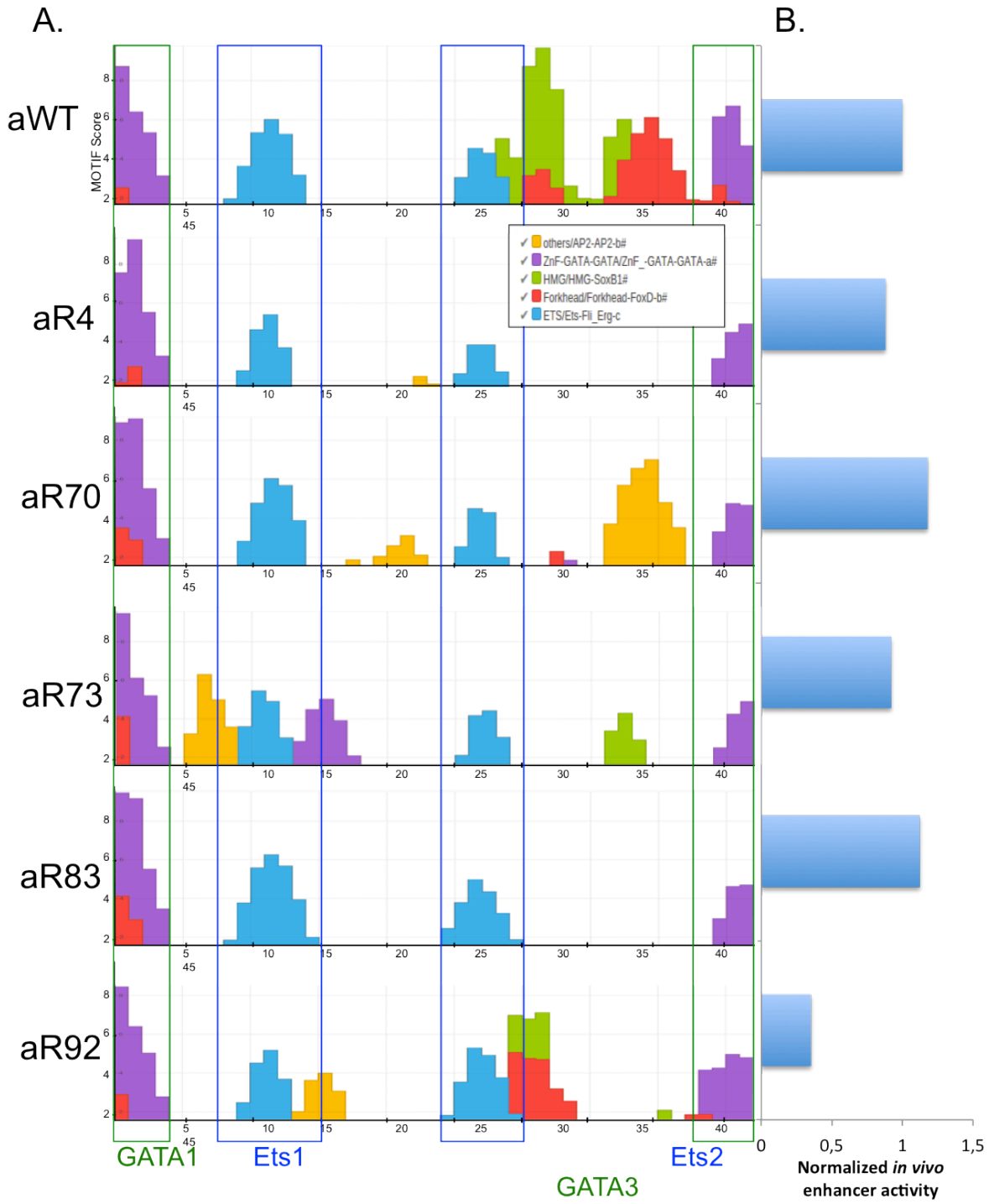
Supplementary Figure 6: DNA sequence alignment of randomized variants for the inactive N61 and C53 clusters. ETS and GATA binding sites are represented by blue, respectively green, arrowheads. Optimized sites are framed in red.



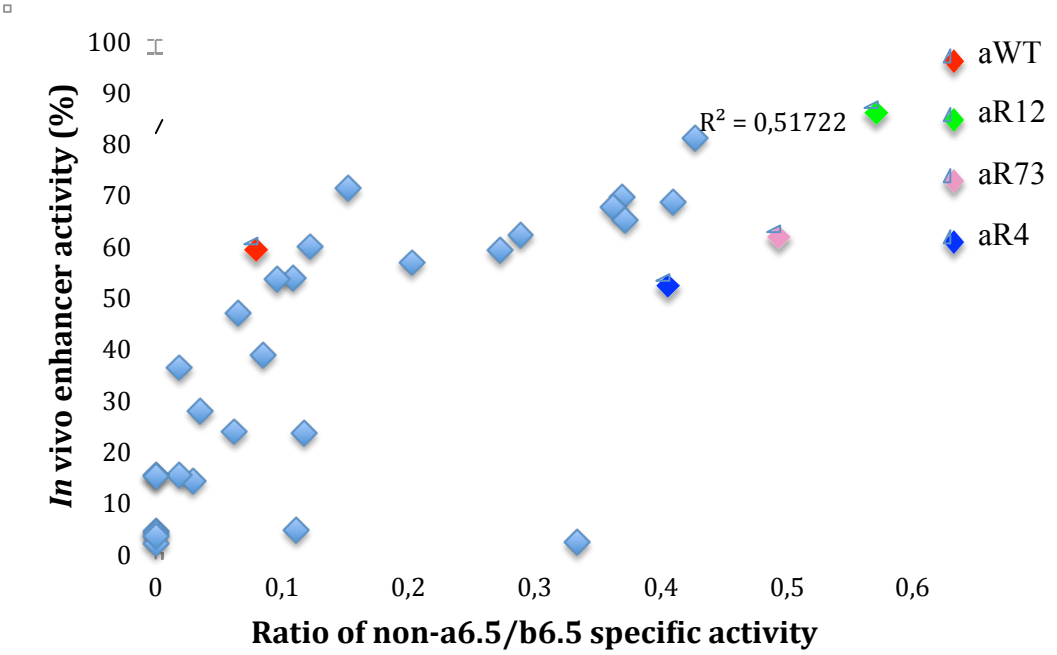
Supplementary Figure 7: Some randomized variants have a broader activity pattern than the a-element. Normalized *in vivo* enhancer activity in indicated cells for WT and randomized variants of **(A)** the a-element, **(B)** N26, **(C)** N61 et **(D)** C53_Opt. The proportion of embryos where LacZ staining was only detected in a6.5 and b6.5 progeny is shown in green. Embryos where activity was detected in other cells appear in light green. Interestingly, this "ectopic" activity is always associated with specific activity in a6.5 and/or b6.5 progeny.



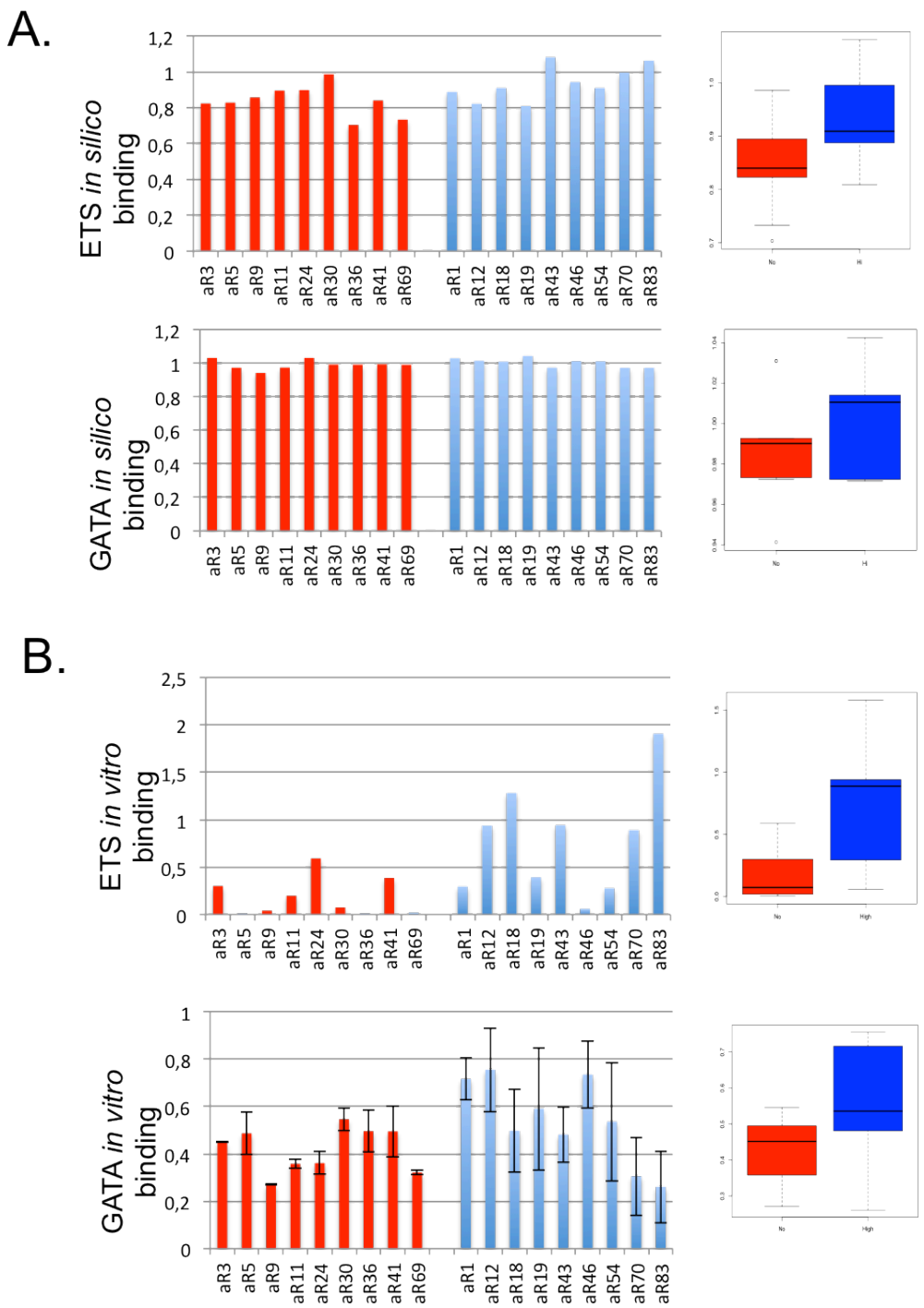
Supplementary Figure 8: Creation of new ETS and GATA sites in randomized spacers cannot explain *in vivo* activity levels of randomized variants. A. ETS and GATA binding sites predicted in the WT a-element and a few randomized variants. For both ETS (Blue) and GATA (red), the predicted affinity binding score calculated by MOTIF on each 8-mer is plotted (X axis). **B**. *In vivo* enhancer activity levels of the variants, normalized to the activity of the WT.



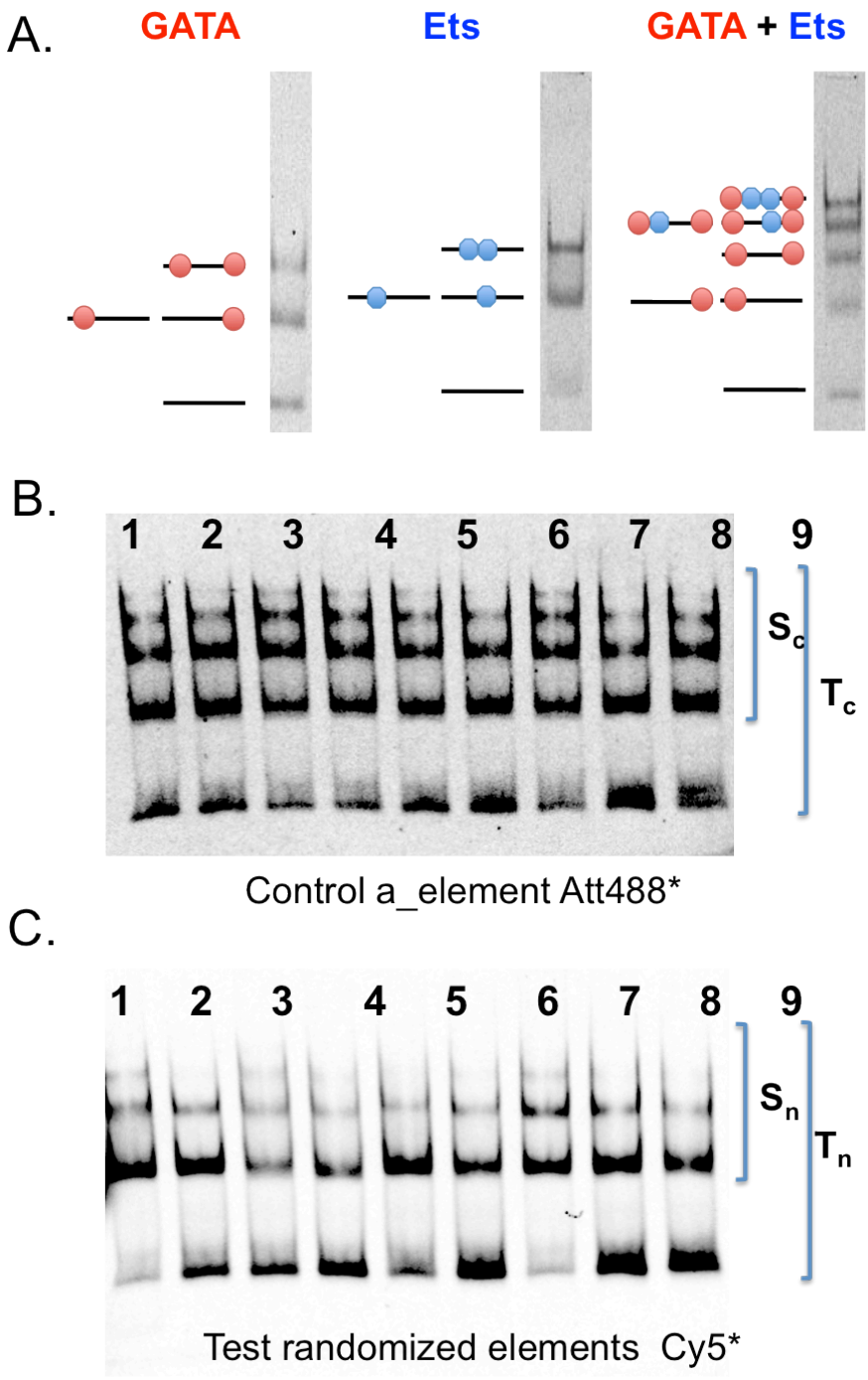
Supplementary Figure 9: New transcription factor binding sites can be created in randomized spacers. A. MOTIF scores for ETS and GATA binding sites (blue and purple), and sites for FoxD (red), SoxB1 (green) and AP2 (yellow), transcription factors present in the embryo at the 16 and 32-cell stages, calculated for the WT a element and a few randomized variants. **B**. *In vivo* enhancer activity levels of the variants, normalized to the activity of the WT.



Supplementary Figure 10 : Non-a6.5/b6.5 enhancer activity is not necessarily cannot be explained by enhancer activity level only. Y axis : for each a-element variant, percentage of embryos harboring LacZ staining. X axis : for each a-element variant, percentage of embryos with staining in a6.5, b6.5 lineages and other cells divided by the percentage of stained embryos. A few elements are identified, as described in the right pannel.



Supplementary Figure 11: Comparison of *in vitro* and *in silico* binding affinities of active *vs.* inactive a-element randomized variants. **A.** Upper=Relative sum of the MOTIF scores for the ETS BS compared to WT of the variants of Figure 3 (paired t-test for the inactive versus active variants, p=0.06651). Lower=Relative sum of the MOTIF score for the GATA BS compared to WT of the variants of Figure 3 (paired t-test for the inactive versus active variants, p=0.3789). **B.** Upper= Relative *in vitro* binding of Ets1/2 compared to the WT a-element of the same revertants as in (A) (paired t-test, p=0.01737). Lower=Relative *in vitro* binding of Ets1/2 compared to the WT a-element of the same revertants as in (A). (paired t-test, p=0.05347)



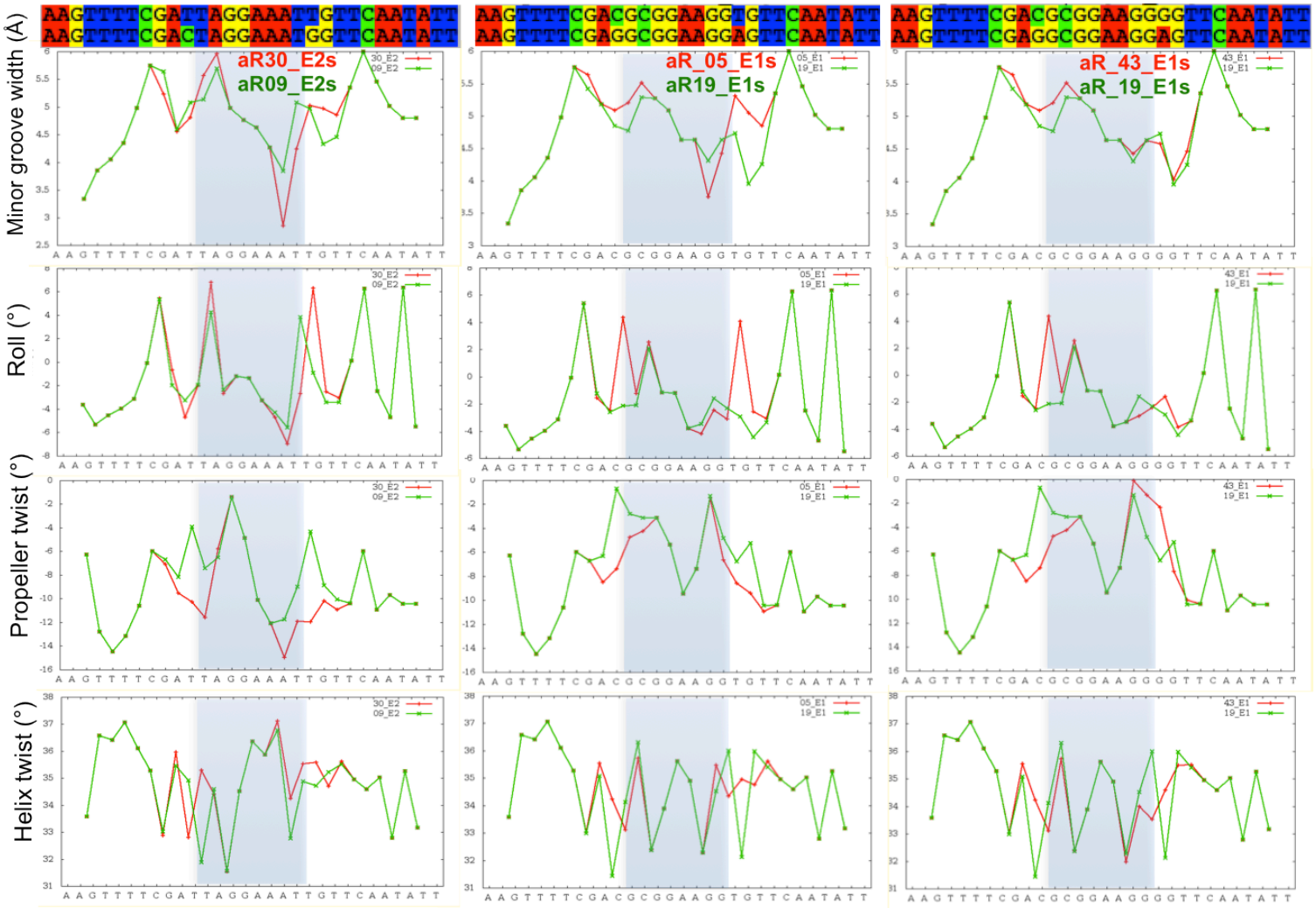
Supplementary Figure 12: Relative affinities are quantified by QuMFRA.

A. Illustration of gel shift results for *in vivo* binding of Ets1/2 (blue circles) and/or Gata4/5/6 (red circles) DNA binding domains on the a-element. Equimolar amounts of Att488 labeled aWT (**B**) and Cy5 labeled randomized variants (**C**) were incubated with recombinant Ets1/2 and Gata4/5/6 DNA binding domains and loaded on a 6% PAGE as explained in materials and methods. The fluorescence of the shifted bands S and the total fluorescence T was quantified with a Amersham imager 600 and the relative affinity Y was calculated as $(S_n/T_n)/(S_c/T_c)$.

#●●# aR30, 70,

```
aR_30 _ATTATCTTTGGCGGAAGTGCAAGATTAGGAAATTGGGGTGAACTTTAGATAGAT  *
aR_70 CATTATCTTGGACGGAAGTGGCCGGCGAGGAAATTTTGATGCCCCAGAGATAGAG  *
```

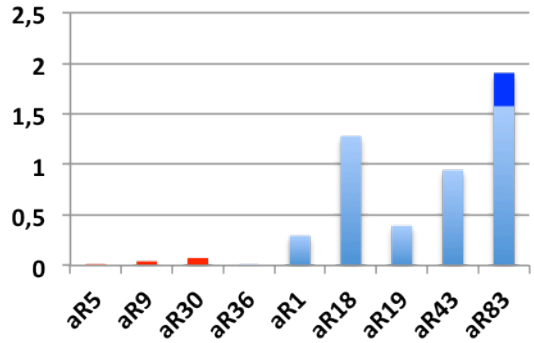
Supplementary Figure 13: Inactive and strongly active randomized variants share similar "good" binding sites. The randomized variants aR30 and aR70 have similar 8bp-binding sites matching the logo sequences for ETS and GATA binding sites extracted by Farley et al. from their experiments [8].



Supplementary figure 14. Flanking sequences modify the conserved core octamer

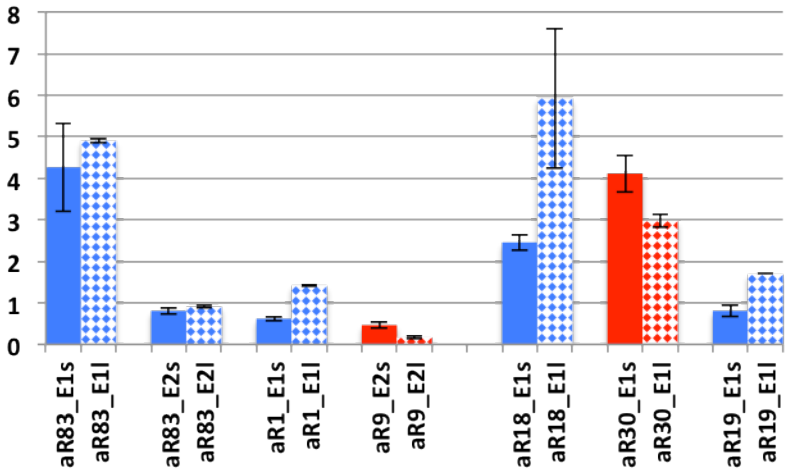
Active and inactive ETS BS with a similar core are compared for the indicated DNA structure parameter using high-throughput DNA shape prediction (Zhou et al., 2013). Inactive BS is in red, active one in green.

A



B

```
a_WT   AAGTTTTCGA AANGGAAATT GTTCAATATT
a_R-S  AAGTTTTCGA NNMGGAARNN GTTCAATATT
a_R-L  NNNNNNNNNN NNMGGAARNN NNNNNNNNNN
```



Supplementary Figure 15: Extension of flanking sequences further modifies the binding affinity of Ets1/2

A. *In vitro* relative binding of Ets1/2 to the whole randomized elements compared to the WT enhancer. Dark blue represents the ratio of the upper shifted band to the lower.

B. *In vitro* relative affinities of Ets1/2 for their indicated binding sites compared to E2 of the WT. BS of active enhancers are in blue, BS of inactive enhancers are in red. The a_R-s fragments are in plain colors, the a_R-L fragments in patched colors.

Name	Sequence	Genome coordinates
C10	TATTGTACGGTATCTTCGTAAGGCAATATCGTTCGTCACGCAGTTTCAG GAAACAACTTGTTACAGTATTCCAGTCAAGACGCATAGTTGCTGCTATT CTCACGGCTTAACCATCCAAGGGGCTTCCGCTTTAGTCTA	Scaffold_1:454236,454373
C13	AAATACAACAGATAGCAGATCGACGCAAACGATAAATCAATCAGCGTTT CCTTTGGCGTGGGAAGATGAATTCTGATGTACTGCTTCCTGTTAGCTTG GTTGCCACAGTAAGGATACAGTCTGTCAGAATCAGGAAGCTGAGCATGG	Scaffold_110:215425,215571
C20	ACAGTCGCGCAGGAAAGGCAAAGCGGACGCCGCAGATTTATCATAGAAA TCAAATATCTTTTTCGTGGCCAATTTATCCGGTTACCTCGAAGAC	Scaffold_178:21070,21201
C24	AGGAGGCTCTTATCTGCCAAGTTTAAACACGGAAATCTCCGTGAAGTCT ACATTTTAAAACGCGTTTTGCGCAGATGCGCTCATGCGGAAATTT	Scaffold_23:188334,188457
C32	GTCTCGTCGCTATCTTATAACGTATTTTTAGTTTTTCTTGTGAACATCC CCTTATATAACCCCTACTTCGCGTCGTTGTACTTTTCCTAAAGTA	Scaffold_40:408605,408753
C39	TGCGTATTTGCGGAAGTAGCTTTCGATAAAATGTGTTATAAAGTCACGA AACCTATGTTTTAAACTTTTTATGATTTTTATCGTTGCCTAATTC	Scaffold_93:252247,252360
N26	TTGTTATCTCGCGATTATGAAAAAAATAACCCATGAGATAATTTTCCTT AATTGCCCGGAAACGGAGTATCAGATAAGCG	caffold_225:76559..76680
N36	CTATTGTCTCTTTAGAATTCCAATCAGGAAATGAGAGGAAATCAAAGCA CTGCCATATGGGGTTATCTCCACACTCCGTCGTATCTGCGTTTAG	Scaffold_111:103345..103473
N39	CCAAGATTACAGGCAATGACATCATTATCTTTTCCAATGCACTCTGTGA CATCATAATTACTTTTCCTGCATTCACGGTATCTTGCAATGACAC	Scaffold_353:16201..16327
N61	CCGCTATCTAGGCCGACCCCGCTCTCCCAAGGAAATGTCCACCTTCCAG TCGGGAAAAGATAACCGCTCGCCAGTGCGACGCTTTCCGGCTG	Scaffold_1:462124..462257
N76	GCTGCATATCCTGTATTAGAATAACAGGAAATCAATATTGTGAGTTACG TAGAAATTTCCGTTCCCATGTGTATTGTTATCTTTGACATTACTG	Scaffold_38:209989..210112
N83	TTCAGATAAGAAATTTAGAGGAAGCCGAAACCCTAAGGAAAAAAATATA TTTATCTCGG	Scaffold_356:28870..28972
N88	GGGGTTGGTATTCTATGGAATGCACAGATATTTAATCATTATCTTTTTA TAGCGGGGCTATTTCCTGTCGACATTGCCAGCGTCAGCGGGGACC	Scaffold_213:73074..73202
N96	GTAGCTAAAGCTACGTCCTTGTCTAAGATATAGCTAGGAGTTGTCGTTT CCTGTTTCCTAGTCTGGGTTATCTAGGGTTTTGACTTGCAGGAAA	Scaffold_231:78191..78309

Supplementary Table 1: Sequences and genome coordinates of genomic ETS and GATA clusters conserved (C) and non-conserved (N) in *Ciona savignyi* genome

Name	Sequence
a_WT	CGTTATCTCTAACGGAAGTTTTCGAAAAGGAAATTGTTCAATATTTAAGATAGGA
aR	NNTTATCTNNNNNGGAAGNNNNNNNNNAGGAAANNNNNNNNNNNNNAAGATAGNN
aR_01	TGTTATCTTGTCCGGAAGGGGGATTTGAGGAAAGTAGGGATTCACTTAGATAGTA
aR_02	GTTTATCTGATTCGGAAGCGGTATACTAGGAAAGTGATTGGTGGGCGAGATAGGG
aR_03	TCTTATCTGGGGCGGAAGGGGGAGGCGAGGAAAGGCCCAGTTGGGGAAGATAGTT
aR_04	GGTTATCTTGCGCGGAAGGTGGACCGTAGGAAACTTTTCGTACGTGAAGATAGTA
aR_05	CGTTATCTGCCGCGGAAGGTTGGGGGAAGGAAAGAAACGGCGCGGCCAGATAGGT
aR_09	CATTATCTGACGCGGAAGATTCAAGCTAGGAAATGCGTATGGGGTTCAGATAGGA
aR_10	ACTTATCTCGGTCGGAAGCGGTTGTTGAGGAAAGTACAACGTCGCTTAGATAGTT
aR_11	GATTATCTATTGCGGAAGGATATAGTCAGGAAAGTGCATGAGCTAGCAGATAGTT
aR_12	CGTTATCTGCCTCGGAAGGACGTTGGGAGGAAATGGTCGGAAGCGTTAGATAGTG
aR_18	GGTTATCTAGGGCGGAAGTGATTAAGAAGGAAAGACACAAGGTTTGTAGATAGGG
aR_19	GCTTATCTAGGGCGGAAGGAGCTCTGGAGGAAAATTGCAGTGGTGGGAGATAGTA
aR_24	GGTTATCTAGTGCGGAAGCGAAGGTGTAGGAAACGGTGAGGGTTCGGAGATAGCT
aR_26	GCTTATCTCGTGCGGAAGGGGTTTGATAGGAAAGAGGCTCCGGGCGGAGATAGCT
aR_28	TGTTATCTGGGGCGGAAGCTTGTATGCAGGAAACGGTCATGTTCTTGAGATAGGC
aR_29	TGTTATCTTTCCCGGAAGGTATACGGAAGGAAATCAAGTGTCGTGGTAGATAGGA
aR_30	_ATTATCTTTGGCGGAAGTGCAAGATTAGGAAATTGGGGTGAACTTTAGATAGAT
aR_34	CATTATCTGGCCCGGAAGGGTCGCGCCAGGAAACGTTGGTATGAGTGAGATAGGG
aR_36	GGTTATCTGTCTCGGAAGGGTCATAAGAGGAAAGTAATAGACGCAGGAGATAGGG
aR_41	TTTTATCTGCGTCGGAAGTAGTGGATTAGGAAACCGCTCGGGAAGGGAGATAGCG
aR_42	GGTTATCTGGGTCGGAAGTGTGGG-ATAGGAAAGGTGGTTATAGGCGAGATAGGA
aR_43	GTTTATCTGGCGCGGAAGGGGAAGAATAGGAAAGTGCGACGGAAGTAAGATAGGG
aR_46	CGTTATCTCGCGCGGAAGGCAGAGAACAGGAAAGTGCACGGAGTGGGAGATAGTG
aR_47	ACTTATCTTCTGCGGAAGCACCACCGTAGGAAAGCTATGGCGGTGTGAGATAGCA
aR_54	GGTTATCTGTGGCGGAAGAAGAGGTTTAGGAAATTGGAGCGTGGTGGAGATAGTG
aR_62	ATTTATCTGTCGCGGAAGCCATTGGGGAGGAAATTCGTTGACAGTGGAGATAGGC
aR_65	CGTTATCTAGGTCGGAAGGGTTGGGGTAGGAAAGGATCGTCCGACTAAGATAGTC
aR_69	ATTTATCTTCATCGGAAGATTTTATCCAGGAAAGCGTAGGTGATGGCAGATAGCC
aR_70	CATTATCTTGGACGGAAGTGGCCGGCGAGGAAATTTTGATGCCCCAGAGATAGAG
aR_73	GTTTATCTCAGGCGGAAGATAGCGGTTAGGAAATAGGTGGTATTCCCAGATAGCA
aR_77	TGTTATCTGGGGCGGAAGGCTGCGTATAGGAAACCTTGTACATTCAGAGATAGAG
aR_83	GTTTATCTTGAGCGGAAGTGTATGGCCAGGAAACGTAGGTCAGCTCTAGATAGAG
aR_87	GATTATCTCGATCGGAAGGTCCTTGGGAGGAAAATAGTCCATACGGAAGATAGGG
aR_92	AGTTATCTGGTACGGAAGGCGTGGATCAGGAAATGTTGACAGGTATTAGATAGGC
aR_96	GATTATCTCTTTCGGAAGTGGGAGGCGAGGAAAAGCGAGCATGAGCAAGATAGTT

Supplementary Table 2: Sequences of the a-element and its randomized variants

Name	Sequence
N26	TTGTTATCTCGCGATTATGAAAAAAATAACCCATGAGATAATTTTCCTTAATTGCCCGGAAACGGAGTATCAGATAAGCG
N26_R	NNNTTATCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTTCCTTNNNNNNCCGGAAANNNNNNNNNAGATAANNN
N26_R3	ATGTTATCTAAGTAG-GGGGGGCTCATCCGGTGGTTGGTTGTTTTCCTTTGAGAACCGGAAAGTCTGCACTAGATAATTT
N26_R6	AATTTATCTCCGGCGAAGGTATTTGTAGCGTCGCATTAACTCTTTCCTTCGCTCACCGGAAACGATTAGGAAGATAAGGG
N26_R11	TGATTATCTCAGTGCTGTGGCAGCATGGGTGGGGACCGGGCCTTTCCTTGATGAGCCGGAAATCACTGGGAAGATAATTA
N26_R12	AGTTTATCTGCTGACTAAGGCTCAGCTGGCAACCCGCGTCATTTTCCTTTAGTTTCCGGAAATGCAAAGAAAGATAAGTG
N26_R19	CGCTTATCTCTCACTCCAGGGCCTGCAAGGCTACCAAGTCGTTTTCCTTAGGGTTCCGGAAACACCTTGGTAGATAATAT
N26_R21	ATCTTATCTCTTCTAATAATAGTACCTTGGTTGTGTTGGTGGTTTCCTTGCCTCGCCGGAAAGAGTGTTGGAGATAAGTT
N26_R22	AGTTTATCTGCTCCGGAGGGTTAATCCGAACCCTCGTCGCCGTTTCCTTCCACATCCGGAAAATCCGGCACAGATAAGGC
N26_R24	GCATTATCTTGGGCCCGGCCGCAATGGTATTCAATGGTGTCTTTTCCTTGTTGGACCGGAAAGGGCCGGTTAGATAAGTT
N26_R26	AGCTTATCTCGTCACGAATTGCCATAACTTGGTGTTTTATTGTTTCCTTCGGGAGCCGGAAAGACAAGCGGAGATAACCT
N26_R27	TCGTTATCTGGTATAGCGCGTTCAACTTCTTGATCGTAATCATTTCCTTTTCAGTCCGGAAAGATGGAATGAGATAAGTG
N26_R28	CCTTTATCTAAGAACACATCCGGGGGTCGGGGGTATCTGTAGTTTCCTTGGAAGGCCGGAAAGACTGGACGAGATAACAC
N26_R29	GTGTTATCTAGTGTGTCAGGGGAAGGGATCGGTTAGCTTCGGTTTCCTTGGCGGCCCGGAAAACCTGCATCAGATAATTG
N61	CCGCTATCTAGGCCGACCCCGCTCTCCCAAGGAAATGTCCACCTTCCAGTCGGGAAAAGATAACCGCTCGCCAGTGCGACGCTTTCCGGCTG
N61_R	NNNCTATCTNNNNNNNNNNNNNNNNNNNAAGGAAANNNNNNNNNNNNNNNNNNNNNNNAGATAACNNNNNNNNNNNNNNNNNTTTCCGGNNN
N61_R3	ACACTATCTTTCATCCAACCAAGGTAGTAAGGAAAGCGCTAGTCAGGCTAGTGAGTTAGATAATCTGGGTTAAATTAGATCTTTTCCGGTTG
N61_R5	AAGCTATCTTTAAACTGTAATAATAGGTAAGGAAAGTGGGTCCGGGGCAGAGTTTACAGATAAGTCACTATCTGGTGTGAGTTTTCCGGGCC
N61_R7	CGGCTATCTGTGGCGGGGGGGAGGGTCAAAGGAAAGGTTGGGGCGGGGGTGGATCAGAGATAAGAACTTTTGCTATGTTGTATTTCCGGTTG
N61_R8	TATCTATCTGATGCAGTGACTCTGCCGAAAGGAAACTCGCTCAAAGGCTGCATTCATAGATAAATTCAGGGCGACTATTTATTTTCCGGATT
N61_R14	ATGCTATCTTGGTCTCCCACCCGGTCGGAAGGAAAGGCGAGATTGTAATGTGTGTGCAGATAATTCGGTTCTGTACAATCGGTTTCCGGTGA
N61_R15	CTCCTATCTCAAGTGACGCGCCGTATAGAAGGAAATATAGCTGTCTGGGTCAGCTTGAGATAAGTGCCACGTGTTATCGGAATTTCCGGTCC
N61_R105	TCACTATCTCGGAGTGGCACTTTCTGCCAAGGAAAGGTTGCTAGGCAGCAAGTAAGGAGATAAATGGGCAAGTTGGGAACTTTTTCCGGGCA
N61_R111	CAGCTATCTGCGTCGGGAGTCTATGACGAAGGAAACGGTGTGGGAGGGAGTACGCAGAGATAATGGGCTTCGGAGGGTGTAATTTCCGGGGG
N61_R112	AGGCTATCTGCGTCGATGGGGAGATTGGAAGGAAAACGAGTGCGGAGTTTGCCGTTGAGATAAGATATGAGCGCGTCAGCGGTTTCCGGACG
C53_Opt	TACCTTCCGTAGGCATCTTCCGTCATACAGGTATATATACCATTATCTTCTTTTGACACAGAGGCTAGATAAAAC
C53_R	NNNCTTCCGNNNNNNNCTTCCGTCNNNNNNNNNNNNNNNNNNTTATCTNNNNNNNNNNNNNNNNNNAGATAANNN
C53_R1	AGTCTTCCGTTACCTACTTCCGTTCTCGGGCTACGGGGCTTGTTATCTCGCGCAATCCGGACTGATAGATAAGTA
C53_R2	GCTCTTCCGTTTTTACCTTCCGTCGCGTTCCGCATTGCGTGGTTATCTAATGTGATTGGTTTAGGAAGATAATCT
C53_R14	TTGCTTCCGTTGTAGACTTCCGTCTGTCTCTTAGCGCAGGGTTTATCTTTCTTTATATCTTCAGAAAGATAATAG
C53_R15	GCCCTTCCGTCTTTTTCTTCCGTGCCTCTCATGCCTGTGCTGTTATCTTTTATTTGTGATCTTTGCAGATAATCA
C53_R18	CAGCTTCCGTAGTTTTCTTCCGTCAGGCACTGAGTGCGTTGATTATCTATATAGTAATGACGTACTAGATAACCA
C53_R19	TCGCTTCCGTTTTTCACTTCCGTAATAGTAAATCCATTCTCCTTATCTTGTATTTTTATTACATTCAGATAAGTG
C53_R20	TCGCTTCCGTGGCTGACTTCCGTCGATGAAGGAAGACTGAGCTTATCTGTACGTCATATAGTGGCAAGATAATTA
C53_R21	TCCCTTCCGTATTTGGCTTCCGTAGATTGTTGGGCGGGGGGGTTATCTTTGGACGGGGCCGTGAGGAGATAATGG
C53_R23	CTACTTCCGTTGAGAACTTCCGTTTTTGTGCGCAATAATTGTTTATCTAGCGTCCACATACGCTTGAGATAAGCT

Supplementary Table 3: Sequences of N26, N62, C53_Opt and their randomized variants

Article 2 : Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development

Au sein de différents taxons de métazoaires et de plantes (ex: Vertébrès ou Drosophilidae), des séquences non-codantes d'espèces, distantes parfois de plusieurs centaines de millions d'années (ex: mammifères-poissons téléostéens), sont autant -voire plus- conservées que des séquences codantes : il s'agit de séquences non-codantes très- et ultra- conservées, suivant le % d'identité qu 'elles partagent ((H/U)CNE : (Higly/Ultra) Conserved Non-coding Elements). Selon une première définition, on qualifie de CNE tout fragment de plus de 350 paires de bases conservé à plus de 70% (Ovcharenko 2004). Pour Bejerano et ses collègues, des éléments ultra conservés correspondent à des séquences d'au moins 200 paires de bases parfaitement conservées entre l'Homme et la souris ou à des fragments d'au moins 50 paires de bases ayant plus de 95% d'identité entre des espèces plus distantes (Bejerano 2004, Lee&Venkatesh 2013).

Comme nous l'avons vu dans l'introduction, cette conservation de séquence reflète généralement la présence d'éléments fonctionnels sous pression de sélection purificatrice/négative, et sont souvent associés à la transcription de gènes du développement (Vavouri 2009, Nelson&Wardle 2013).

Le séquençage des premiers génomes de mammifères a permis d'identifier 481 UCEs dans les génomes de la souris, du rat et de l'homme (Bejerano et al., 2004). Ces éléments sont également présents (avec 95% d'identité) dans les génomes de mammifères et d'oiseaux, et de poissons téléostéens. Les trois quarts d'entre eux ont été identifiés dans le génome du requin-baleine (Venkatesh et al., 2007), et un petit nombre dans le génome de la lamproie (Ishibashi et al., 2012). Très peu (5%) de ces éléments peuvent être partiellement identifiés dans les génomes d'invertébrés (Bejerano et al., 2004), suggérant leur origine évolutive à la base des vertébrés (Lee & Venkatesh, 2013).

Alors que les ascidies sont les invertébrés les plus proches phylogénétiquement des vertébrés (Delsuc et al., 2006), la première étude cherchant des séquences non-codantes communes à tous les génomes de chordés n'a pu identifier des CNEs communes que chez l'amphioxus (céphalochordé) et les vertébrés (Hufton et al., 2009).

L'article ci-dessous présente une nouvelle analyse des génomes d'olfactores (vertébrés et tuniciers) qui a permis d'identifier 183 UCEs. Ces éléments sont situés dans des régions homologues et synthéniques chez les vertébrés, mais pas chez les tuniciers. J'ai caractérisé l'activité *in vivo* de trois de ces régions issues des génomes de la cione et du poisson-zèbre chez *Ciona intestinalis* et montré qu'elles activent la transcription pendant le développement embryonnaire. Les éléments des deux espèces agissent également en tant qu'enhancers lorsqu'ils sont testés chez l'autre espèce, mais leurs profils d'activité divergent, suggérant que l'activité enhancer de ses séquences a été conservée, en dépit d'une divergence de la logique régulatrice, possiblement en *trans.* J'ai également analysée une large région située de part et d'autre le désert génique où se situent ces trois UCEs dans le génome de *Ciona*, cherchant - en vain - à identifier des gènes dont le profil d'expression peut être associé au profil d'activité d'un de ces éléments.

3600–3618 *Nucleic Acids Research, 2013, Vol. 41, No. 6* *Published online 7 February 2013*
doi:10.1093/nar/gkt030

Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development

Remo Sanges[1,*], Yavor Hadzhiev[2], Marion Gueroult-Bellone[3,4], Agnes Roure[3], Marco Ferg[5], Nicola Meola[6], Gabriele Amore[1], Swaraj Basu[1], Euan R. Brown[1,7], Marco De Simone[8], Francesca Petrera[8], Danilo Licastro[8], Uwe Strähle[5], Sandro Banfi[6,9], Patrick Lemaire[3,4], Ewan Birney[10], Ferenc Müller[2] and Elia Stupka[11,*]

[1]Laboratory of Animal Physiology and Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy, [2]Centre for Rare Diseases and Personalised Medicine, School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK, [3]Institut de Biologie du Développement de Marseille Luminy, UMR 6216 CNRS/Université de la Méditerranée, F-13288 Marseille cedex 9, France, [4]Centre de Recherche de Biochimie Macromoléculaire (CRBM), UMR5237 CNRS/Universités Montpellier 1, 2, 1919 route de Mende, F-34293 Montpellier cedex 5, France, [5]Karlsruhe Institute of Technology (KIT), Institute of Toxicology and Genetics and University of Heidelberg, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany, [6]Telethon Institute of Genetics and Medicine, 80131 Naples, Italy, [7]School of Engineering and Physical Sciences, Heriot Watt University, Edinburgh EH14 4AS, UK, [8]CBM Scrl, AREA Science Park, Basovizza, 34149 Trieste, Italy, [9]Medical Genetics, Department of Biochemistry, Biophysics and General Pathology, Second University of Naples, 80138 Naples, Italy, [10]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK and [11]Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute, Via Olgettina 58, 20132 Milano, Italy

Received September 29, 2012; Revised December 21, 2012; Accepted January 3, 2013

ABSTRACT

Co-option of cis-regulatory modules has been suggested as a mechanism for the evolution of expression sites during development. However, the extent and mechanisms involved in mobilization of cis-regulatory modules remains elusive. To trace the history of non-coding elements, which may represent candidate ancestral cis-regulatory modules affirmed during chordate evolution, we have searched for conserved elements in tunicate and vertebrate (Olfactores) genomes. We identified, for the first time, 183 non-coding sequences that are highly conserved between the two groups. Our results show that all but one element are conserved in non-syntenic regions between vertebrate and tunicate genomes, while being syntenic among vertebrates. Nevertheless, in all the groups, they are significantly associated with transcription factors showing specific functions fundamental to animal development, such as multicellular organism development and sequence-specific DNA binding. The majority of these regions map onto ultraconserved elements and we demonstrate that they can act as functional enhancers within the organism of origin, as well as in cross-transgenesis experiments, and that they are transcribed in extant species of Olfactores. We refer to the elements as 'Olfactores conserved non-coding elements'.

INTRODUCTION

The sequencing of a large number of vertebrate genomes has enabled the identification of conserved non-coding

*To whom correspondence should be addressed. Tel: +39 081 5833428; Fax: +39 081 7641355; Email: remo.sanges@szn.it
Correspondence may also be addressed to Elia Stupka. Tel: +39 022 6439137; Email: stupka.elia@hsr.it
Present addresses:
Agnes Roure, Laboratoire de Biologie Intégrative des Organismes Marins (BIOM), Observatoire Océanologique de Banyuls, 1 Avenue du Fontaule, 66650 Banyuls-sur-mer, France.
Gabriele Amore, Istituto Regionale Vini e Oli, via Libertà 66, 90143 Palermo, Italy.

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

elements (CNEs) that are constrained during evolution. They were shown to act as tissue-specific enhancers mostly associated with transcription factors that are active during development (1–4). Owing to this role, CNEs are thought to play an important role in gene regulatory networks that specify body plans (5). Genes associated with CNEs require complex spatial and temporal cis-regulation, and indeed key developmental genes contain arrays of CNEs in their intergenic and intronic regions (3,4,6). While CNEs are present in several groups of metazoans such as vertebrates, flies and nematodes and are the most conserved sequences within these groups, they have diverged beyond recognition (if they were originally related) among these groups (6). Such enhancers can work in a modular and autonomous way. They can be active and maintain their specificities regardless of the genetic background, such as associated promoters or genes, and can work in combination with other enhancers from different genomic contexts (7–9). Finally, they can also be functional when transfected from one species to another even if their specificity is not always retained (10–12). Limited evidence also supports the potential activity of enhancers across organisms belonging to different groups. This was shown in the Hox locus for some amphioxus enhancers tested in chicken, mouse and *Ciona* experiments, as well as *Ciona* enhancers tested in chicken (13,14).

Most comparative studies to date have focused on mammalian and vertebrate genomes. Two studies (10,15) have been able to find functional CNEs between vertebrates and lancelets (amphioxus), which has been suggested to be the most basal chordate group (16). Recently, an interesting article showed the conservation of both sequence and function of a neural-specific enhancer conserved among human, zebrafish, amphioxus, *Saccoglossus*, sea urchin and *Nematostella* (12). However, the sequence of the enhancer is not conserved in tunicate genomes (data not shown; Salvatore D'Aniello, personal communication). Therefore, no CNEs between vertebrates and tunicates have been reported so far, and several studies propose that no such regions exist (2,15). The CNEs discovered in amphioxus act as developmental enhancers and are conserved in syntenic vertebrate regions, which makes the lack of CNEs in tunicates an interesting subject for further studies, given that they are the proposed sister group of vertebrates (17). Regulatory elements may have diversified too much to be recognizable using common comparative genomic strategies, which mainly rely on the identification of collinearly conserved elements found within orthologous loci. It is important to point out that tunicate larvae present the classical chordate body plan but they have greatly diverged at the molecular level leading to the paradox that divergent gene expression programs can lead to similar body plans (18). In addition, in tunicates, many developmental genes and signaling pathways are co-opted differentially as compared with vertebrates [reviewed in (19)], making it difficult to understand where this divergence is embedded.

According to this reasoning, the existence of evolutionary phenomena termed 'cis-regulatory rewiring' and 'enhancer shuffling' were proposed to account for such differences several times (5,9,20–22). Limited evidence so far relies on individual instances, which have been verified in yeast, insects, sea urchins and tunicates (23–27). Such evidence mostly involve the shift of individual binding sites, a variation that could arise by non-conservation–based mechanisms such as mutations followed by stabilizing selection. Another mechanism for the spreading and shuffling of regulatory regions relies on these regions evolving from transposable elements, which has been observed in mammalian genomes (28). Several studies have previously shown that *cis-regulatory* elements can shuffle during evolution within the same gene context, i.e. changing location with respect to the gene structure, but maintaining the association to the same gene (26,29,30). A study in plants has also reported an event of promoter shuffling generated by inter-chromosome and subsequent intra-chromosome recombination (31). Kent *et al*. (32) noticed an unexpected number of small fragments conserved between non-syntenic regions analyzing mammalian genomes, and similarly, in the ENCODE pilot project, the presence of small non-syntenic conserved regions were reported (33). Therefore, non-syntenic rearrangements of conserved (hence potentially functional) sequences did happen during evolution, and they are unlikely to be the mere result of assembly errors, but no further elucidation of their evolution and function has been undertaken so far.

By using an approach that allows the identification of shuffled elements, we have previously demonstrated that the number of functional vertebrate CNEs is significantly higher than reported by using BLAST-like approaches. We identified syntenic rearrangements of regulatory sequences that occurred in vertebrate conserved non-coding regions (29). Our approach has also been successfully adopted in the discovery of elements conserved between vertebrates and the basal chordate amphioxus (10). Now, to evaluate the existence of conserved non-coding sequences between vertebrates and tunicates, we improved and extended our methodology, using progressive alignments, randomizations and a strict false discovery rate (FDR) filtering. We were able to explore the conservation of putative regulatory regions with unprecedented sensitivity and developed a pipeline that led, for the first time, to the discovery of 183 non-coding sequences conserved within Olfactores.

MATERIALS AND METHODS

Data selection

Local installations of the MySQL Ensembl databases version 49 and the relative API (34) were used to extract sequences and annotations from the next Olfactores genomes: *Mus musculus*, *Homo sapiens*, *Canis familiaris*, *Danio rerio*, *Takifugu rubripes*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Ciona intestinalis*, *Ciona savignyi* (see Supplementary Table S1 for information about the used Ensembl core databases and the relative genome builds). Species were divided into three groups according to their phylogenetic classification: mammals, fishes and tunicates. For each group, a representative organism was chosen and its sequences used as the reference genome in

VISTA analysis as well as in inter-group analysis. The representative organisms are *M. musculus*, *T. rubripes* and *C. intestinalis* for mammals, fishes and tunicates, respectively. The selection of genes was conducted in each group independently. Basically, using the annotation from Ensembl Compara (35) we selected all the genes containing an homolog classified as ortholog_one2one inside all the species of the group, which led us to collect 14 201, 12 896, 5786 groups of orthologous genes loci for mammals, fishes and tunicates, respectively. For each gene, we extracted the whole genomic repeat-masked sequence containing the transcriptional unit and the flanking sequences up to the preceding and following gene. If there were nested genes present in the locus, they were not taken into consideration to determine the extent of sequence to analyze. The regions were extracted from Ensembl and the 5′–3′ sequence of the locus was stored in a custom database having always the determining gene on the positive strand. The pipeline makes use of custom perl scripts and the Bioperl API (36).

Identification of rCNEs

To define regionally conserved non-coding elements (rCNEs), analyses were conducted independently in each collection of orthologous genes for each group. Global multiple alignments in each group were performed on each collection of homologous genes using MLAGAN (37) with default parameters. The multiple alignments thus obtained were parsed using VISTA (38) and perl scripts with the next parameters according to the following groups:

Mammals and fishes: sliding 20 bp; minimum length 100 bp; minimum identity 80%; minimum length after three species overlap 100 bp.

Tunicate: sliding 20 bp; minimum length 100 bp; minimum identity 60%.

Resulting conserved regions were then filtered stringently to distinguish 'known genic' (having evidence of transcription) from 'non-genic' (not having evidence of transcription) into Ensembl and to discard redundant sequences. Basically, all the conserved sequences were screened against the annotations of overlapping complementary DNAs (cDNAs), proteins, ESTs and predictions from the Ensembl core, other features and eventually cDNAs databases of the reference organisms. In the case of overlap, the elements were considered 'genic' and excluded from the remaining analysis. Finally, in cases in which the upstream region of an analyzed gene coincided with the downstream region of another analyzed gene, rCNEs were counted only once and associated to the locus showing the highest score or the longest transcript in case of equal scores.

Identification of vCNEs

To identify vertebrate conserved non-coding elements (vCNEs), a combined local and multiple alignment strategy was used. This procedure does not look necessarily for collinear elements as the previous one. In the first step, we selected conserved elements between mammals and fish. Each representative mammalian rCNE was aligned against the entire set of representative fish rCNEs by using CHAOS (39) with the next parameters: b = 1, ext = 1, v = 1, co = 10, rsc = 1500, wl = 10, nd = 1 and selecting for segments conserved at least 50 bp sharing an identity of at least 70%. Resulting pairwise alignments were used to extract the corresponding slice from the original rCNE multiple alignments that subsequently were aligned between them using PROLAGAN (37) with default parameters. The cutoff to define the significance of these alignments was determined by randomization analysis. The alignment columns in each rCNE were shuffled so that we maintained the same base composition and identity scores inside the elements but creating not-biologically meaningful sequences. The CHAOS anchoring and PROLAGAN alignment steps were then performed on the randomized rCNEs as reported above and the results used as false positives in the determination of the cutoff for the selection of true vCNEs. The cutoff was calculated on the basis of the overall percentage identity of the multiple alignments to consider significant a percentage of false positives $<0.5\%$ (FDR < 0.005) when the same filter is applied to the randomized data.

Identification of oCNEs

The same CHAOS, PROLAGAN, randomization and FDR filtering procedures were used as reported above, aligning the vCNEs with the tunicate rCNEs. The CHAOS analysis was executed between the *T. rubripes* and *C. intestinalis* sequences with the next parameters: b = 1, ext = 1, v = 1, co = 10, rsc = 1500, wl = 10, nd = 2 and selecting for segments conserved at least 40 bp sharing an identity of 60% minimum. Finally, the fugu sequences of the resulting 204 mammal/fish/tunicate conserved elements were searched against the repeat-masked sequences of the zebrafish genome by using WU-BLAST with the following parameters: E = 1, W = 5, B = 100, M = 1, N = −1, Q = 2, R = 1, filter = none, hspsepSmax = 10 hspsepQmax = 10, hspmax = 0. All the resulting hits were filtered to present: percentage identity $\geq 80\%$ and query coverage $\geq 80\%$. For each fugu sequence, the top hit was manually chosen, curated and classified as Olfactores conserved non-coding element (oCNE) according to the following criteria in the given order: smaller e-value, presence of the hit on a chromosome, bigger length, higher identity, fugu/zebrafish collinearity, longer contig containing the sequence. We were able to retain 183 of the 204 conserved elements and we focused on this set of conserved regions. It is important to mention that zebrafish was excluded from the initial analysis because it retained many more duplicated loci in respect to other teleosts, and this made the initial 1-to-1 homologous group definition poorly efficient.

Homology analysis

We collected all the Ensembl genes mapped in intervals up to 2 Mb (1 Mb upstream and downstream) around each element in every representative genome per group. For each gene, we collected the evolutionary relationship from

Ensembl Compara. We took into consideration the following relationship from the database: ortholog_one2one, ortholog_one2many, ortholog_many2many, between_species_paralog and apparent_ortholog_one2one. For each gene in the interval, we also calculated the number of bystander genes as the number of genes intervening between the gene and the conserved element. We verified within the 1 Mb intervals upstream and downstream of each element, in each pair of species, the presence of evolutionarily related genes as opposed to unrelated genes, taking into account as syntenic conserved oCNEs only those showing a maximum of four bystander genes between the conserved fragment and the closest pair of orthologous genes. For each element, we also measured the number of evolutionarily related pair of genes in the analyzed genomic interval. We also searched for duplicated oCNEs in the genome of *M. musculus*, *D. rerio* and *C. intestinalis* using Blastn with default parameters and selecting only hits showing at least 95% coverage at 95% identity. Results have been manually checked on the Ensembl genome browser in the searched species and in *H. sapiens*, *T. rubripes* and *C. savignyi*.

Aniseed annotation integrations

The next transcript models and annotations were downloaded from the ANISEED database (40): JGI version 1, KYOTOGRAIL2005, KH and ENSEMBL. Data were downloaded from the webpage http://bit.ly/12oO1NL that redirects to the respective archives. Transcript models and annotations were parsed and joined together to form a unique collection, and a MySQL database containing all the information downloaded and generated was used to collect and manage the data using custom perl scripts. Annotations were attached to the data in the pipeline by using the Ensembl transcript ID that represented the ID common to the two sets of data (Ensembl and Aniseed).

oCNEs search in *Oikopleura* and amphioxus

Oikopleura genome assembly version 3 was downloaded at JGI from http://bit.ly/VPjaD7 with relative annotations from http://bit.ly/TYKrTY and proteome from http://bit.ly/V5Q8yC. Amphioxus genome version 2 was downloaded at JGI from http://bit.ly/12oNxXJ with relative annotations from http://bit.ly/UcbKOg and proteome from http://bit.ly/ZievBL. oCNE multiple alignments containing the sequences of all the analyzed organisms were used to build Hidden Markov Models (HMMs) using the program HMMB from the HMMER tool version 1.8.5 (41). The program HMMFS was then used to search the HMMs against the entire *Oikopleura* and the amphioxus genomes on both strands. A cutoff score of 20 bits was used to determine whether an oCNE was conserved. This score indicates that a selected match is 2^{20}-fold more likely to represent an authentic match than to occur by chance. Putative target genes were considered the genes flanking and overlapping (if any) the regions where HMMs matched the analyzed genomes.

Blast2GO annotation

Protein sequences of the putative target genes in amphioxus and *Oikopleura* were functionally annotated using the Blast2GO (42) tool with default parameters. We executed the following analysis step: Blastp against NR proteins, Gene Ontology (GO)-mapping, annotation, annotation augmentation, InterProScan. Finally we run the analysis 'make combined graph' to count the frequencies and evaluate the scores of the GO classes occurrences in the annotated sequences. We took into consideration the top 10 scoring GO classes with a score higher than 5 at a level higher than 2.

Amphioxus gene pair analysis

We used the proteomes of mouse and sea squirt from Ensembl 49 and the amphioxus proteome version 1 downloaded at JGI from http://bit.ly/ZievBL. We classified all the putative homology relationships between *Ciona*/amphioxus and mouse/amphioxus proteomes by executing Blastp searches with default parameters but a maximum e-value of 0.001. We took only the best hit (or the best ones in case of equal e-values) showing a minimal coverage of 50% to build a table of putative homologies. We then analyzed the locations of the genes flanking or overlapping oCNEs in mammals and tunicates in the amphioxus genome. The same analysis was repeated 1000 times randomizing the homology associations between *Ciona*/amphioxus and mouse/amphioxus. Positive elements were considered only those showing at least one pair of associated genes on the same amphioxus scaffold.

Ciona enhancer validations

The *Ciona* oCNE test fragments were designed cloning the corresponding entire *C. intestinalis*/*C. savignyi* conserved block as taken manually from the Ensembl browser (34). Genomic fragments containing the *Ciona* sequences of the three selected oCNEs were cloned in Gateway constructs (43) upstream of the pFOG basal promoter and the LacZ reporter gene. Each construct has been tested twice and two constructs have been prepared for each element. Ciona electroporated embryos were developed until the early tailbud stage and fixed for Xgal staining. About 100 embryos were inspected for each fragment. The sequences of the primers used are listed in Supplementary Table S2. Each clone was verified by Sanger sequencing.

Zebrafish enhancer validations

The zebrafish oCNE test fragments were designed cloning the corresponding entire zebrafish/mammals conserved block as taken manually from the Vista browser (44). The fragments were amplified from zebrafish genomic DNA and cloned in reporter construct containing zebrafish *hsp70* minimal promoter and *venus* reporter gene (mCherry in transgenesis experiments), using Gateway system (43,45). The Gateway destination vector has been previously modified by introducing medaka Tol2 transposase recognition sequences flanking both sites of

the reporter cassette to allow more efficient integration of the transgene into the genome (46). A 570 bp fugu genomic fragment named *EK*, previously reported (45) to lack enhancer activity, was used as enhancer control. The reporter construct for each element was injected into fertilized zebrafish eggs. The composition of the injection solution was as follows: 15 ng/μl plasmid DNA (reporter construct), 10 ng/μl tol2 *in vitro* synthesized transposase messenger RNA supplemented with 0.1% Phenol red. Approximately 150–200 (100 in transgenesis experiments) embryos were injected for each reporter construct. The injected zebrafish embryos were analyzed for reporter gene activity between 24 and 28 hpf using Nikon SMZ1500 fluorescent microscope (Olympus ScanR automated microscope in transgenesis experiments). The expression was quantified as percentage of the embryos showing a specific pattern of reporter expression from the total number of normal developing embryos. Oligo sequences used to amplify the zebrafish oCNEs are listed in Supplementary Table S2.

Ultraconserved elements and enhancer browser data overlap

The genomic coordinates of the set of extended ultraconserved elements (UCEs) by Stephen *et al.* (47) were kindly provided by John S. Mattick. Overlap analyses were performed between the human coordinates of oCNEs and the coordinates of the 5404 vertebrate UCEs. oCNEs overlapping a UCE for at least 50% of their length were considered to be derived from this family of conserved elements. To analyze the overlap of oCNEs with validated UCEs, we downloaded all the elements found in the enhancer browser database (48), together with the functional validation results on 14 November 2012 from http://enhancer.lbl.gov/. The database was composed of 1756 elements. Mouse oCNEs sequences were searched in the downloaded sequences by using Blastn. Fisher exact test was used to test significance for the positive/negative ratio of the complete set of validated conserved sequences and the oCNE overlapping set.

eRNAs overlap

The genomic coordinates related to intergenic transcribed enhancers (eRNAs) were extracted by the supplementary material associated to the article by Kim *et al.* (49). The dataset contained 5117 single nucleotide positions related to intergenic enhancers of which 2052 were classified as transcribed and 3065 as non-transcribed. oCNEs and vCNEs were considered to overlap eRNAs if their genomic coordinates were overlapping within a 1.5-Kb interval upstream/downstream of the eRNA single-nucleotide position.

Expression analysis

Reverse transcriptase-polymerase chain reaction (RT-PCR) were executed on cDNAs and RNAs extracted by different developmental stages of *M. musculus* (embryonic day 8.5, 12.5 and adults), *D. rerio* (dome, shield, 24 hpf and 5 dpf) and *C. intestinalis* (10 hph). Primers were designed to amplify a fragment of ~100 bp around each element using Primer3 (50). As positive control, we used primers designed to show the transcription of the following coding transcripts: *bActin* (*D. rerio* and *M. musculus*), *otx2* (*M. musculus*), *Ci-ATBF* (*C. intestinalis*). We used the following non-coding transcripts: *Ci-Pans* (*C. intestinalis*) (51) and *Pans* (*Mm.221244*, the murine homolog of *Ci-Pans*) (52). All the used controls are known to be expressed at the time of the sampling. As negative control, we used DNAseI-digested RNA that was also used as template for cDNA synthesis. In addition, in *C. intestinalis*, we also used different combinations of the validated oCNEs forward/reverse primers. The primers used and schemas with the protocols for the reactions can be found in Supplementary Tables S3 and S4.

ENCODE/CSHL Long and Short RNA-seq overlap analysis

Human oCNEs sequences were mapped on the hg19 version of the human genome by using the liftOver tool. A custom perl script was then used to query the public instance of the UCSC MySQL database of the human genome version hg19 at the host genome-mysql.cse.ucsc.edu (53). The script queried all the tables pertaining to the ENCODE/CSHL Long and Short RNA-seq data (54,55) publicly available at the moment of the analysis to test for the overlap of the 183 oCNEs human sequences in every sample. The tables analyzed correspond to 182 samples.

Domain analysis

We collected all the Ensembl genes mapped in intervals up to 2 Mb upstream and downstream around each element in every representative genome per group. For each gene, we collected the domains composition from the Ensembl core annotations and looked for domains common to all the three groups. The same analysis was executed on a set of randomly sampled genomic regions from the three groups of the same dimension as the oCNEs dataset. To avoid methodological biases, the randomly sampled genomic regions were selected in the next way: vCNEs conserved between mammals and fishes were randomly selected and associated to randomly selected Tunicate rCNEs. We considered three different intervals around the conserved elements (500, 1000 and 2000 Kb) and performed the Fisher exact text comparing the proportion of common domains between the real oCNEs set and the random vCNEs/rCNEs associations for each interval and for each domain. *P*-values were corrected using the Benjamini and Hochberg method.

TFBS analysis

oCNEs sequences from *M. musculus*, *T. rubripes* and *C. intestinalis* were analyzed for their composition in transcription factor binding sites. We used the FAMILY collection of matrices from the 2008 version of the Jaspar4 database (56) together with the TFBS perl API (57). A threshold of 80% was used to map binding sites on sequences. To avoid methodological biases, the same

TFBS scanning procedure was performed, for each species, on a set of randomly selected regions of the same length and number of oCNEs extracted from the set of vCNEs (mouse ad fugu) or rCNEs (*Ciona*) and the results compared using the Fisher exact text. *P*-values were corrected using the Benjamini and Hochberg method. Only binding sites significantly enriched in all the three organisms tested were kept into consideration.

GO analysis

We considered the association of the conserved elements to the genes determining the genomic regions selected in the data-selection phase. GO functional enrichment analyses were performed on the set of mouse and *Ciona* genes associated to oCNEs. For the mouse set, we compared the following: (i) functional annotation of the vCNEs with the mammalian rCNEs; and (ii) functional annotation of the vCNEs with the oCNEs by using DAVID (58) and FATIGO (59). The *Ciona* GO annotations are not present in the above tools nor in the official GO release; therefore, we extracted them from the Aniseed database (40). We then compared the set of genes retaining tunicate rCNEs with the set of genes retaining oCNEs. For each class, we used the Fisher exact text and the *P*-values were corrected using the Benjamini and Hochberg method.

RESULTS

Discovery of mammalian, fish and tunicate conserved non-coding elements (Olfactores CNEs)

CNEs between vertebrates and tunicates have never been reported so far (15). Therefore, we developed a novel highly sensitive, highly stringent, progressive pipeline to be able to identify the presence of vertebrate CNEs in *Ciona* (see 'Material and Methods' section and Figure 1A). The pipeline used the following combinations of three groups of genomes: (i) Mammals: *M. musculus*, *H. sapiens*, *C. familiaris*; (ii) Fishes: *T. rubripes*, *G. aculeatus*, *O. latipes*; and (iii) Tunicates: *C. intestinalis*, *C. savignyi*. We took into consideration all genes for which there were predicted 1-to-1 orthologs within the Ensembl database (34) in all the genomes within each group, which led to the analysis of 14 201, 12 896 and 5786 groups of orthologous loci for mammals, fishes and tunicates, respectively. We began by selecting elements conserved in a collinear manner within each single group using MLAGAN and VISTA (38,60). The results were filtered stringently to discard 'known genic' regions (potentially transcribed regions overlapping annotated genes, proteins, EST and predictions) and eliminating redundancies from 'non-genic' regions. The analysis produced 92 435, 27 145 and 4525 non-redundant, non-genic, collinearly conserved elements in mammals, fish and tunicates, respectively. We will refer to these initial dataset as rCNEs because they are conserved collinearly within each group. We then proceeded to use these data in a multi-step local and multiple alignment strategy to progressively search for conserved elements among groups, allowing non-collinear conservation and using randomization steps to exclusively select significantly conserved regions with a FDR <0.5% (see 'Material and Methods' section) (61). To get sequences conserved in vertebrates, we aligned mammalian rCNEs against fish rCNEs, generating what we will refer to as vertebrate CNEs (vCNEs). We obtained 900 vCNEs, the majority of which overlap the set of CNEs by Woolfe *et al.* (3). Then, to select sequences significantly conserved across Olfactores, we aligned tunicate rCNEs against the 900 vCNEs, performing a further step of randomizations and FDR filtering (see 'Material and Methods' section). The analysis resulted in a final set of 183 oCNEs associated to 91, 93 and 121 genes in mammals, fishes and tunicates, respectively. Table 1 can be used to get a clearer understanding of all the abbreviations most commonly used to refer to conserved elements in this and other articles.

The pipeline herein presented progressively joins together groups of species to extract group-specific conserved elements. We focused on the 183 elements conserved among all the Olfactores species considered (see Figure 1B–E for an exemplar element and descriptive charts). Their overall percentage identity (number of identical columns divided by the length of the alignment) spans from 52 to 67% (average 55%). The average length of the elements is 45 bp, and the majority of them are found in intergenic regions in all the analyzed species. Supplementary Table S5 contains all the information about the elements discovered and their sequences.

oCNEs are non-syntenic between vertebrates and tunicates

The 183 oCNEs identified in this study are syntenic among vertebrate loci, but are found in non-syntenic locations in tunicates (i.e. surrounding genes for which the orthologous genes are not found in the corresponding vertebrate locus). It is well known that conserved enhancers can be functional over long distances and that bystander genes can be found between enhancers and their target genes (62,63). To check if oCNEs could be located far from their target genes, we searched for orthologous and/or paralogous genes in regions up to 2 Mb in mouse/fugu for vertebrates and mouse/sea squirt for Olfactores. We verified within 1 Mb intervals upstream and downstream of species, the presence of evolutionarily related genes as opposed to unrelated genes, taking into account as syntenic conserved oCNEs only those showing a maximum of four unrelated genes between the conserved fragment and the orthologous genes. In vertebrates, the majority of the oCNEs are found directly flanking or overlapping orthologous genes and they are also found in prevalence in large syntenic blocks. Only seven elements in mouse show the presence of one unrelated gene and three are separated by two unrelated genes from their putative target gene. In fugu, three elements show one unrelated gene. Overall, >85% of the elements analyzed contain >1 pair of evolutionarily related genes in the analyzed interval, >50% of the elements contain >5 pairs and ~20% contain >10 pairs. We performed the same check comparing the mouse and the *Ciona* genomes, and no element could be classified as syntenically conserved. To further verify this finding, we

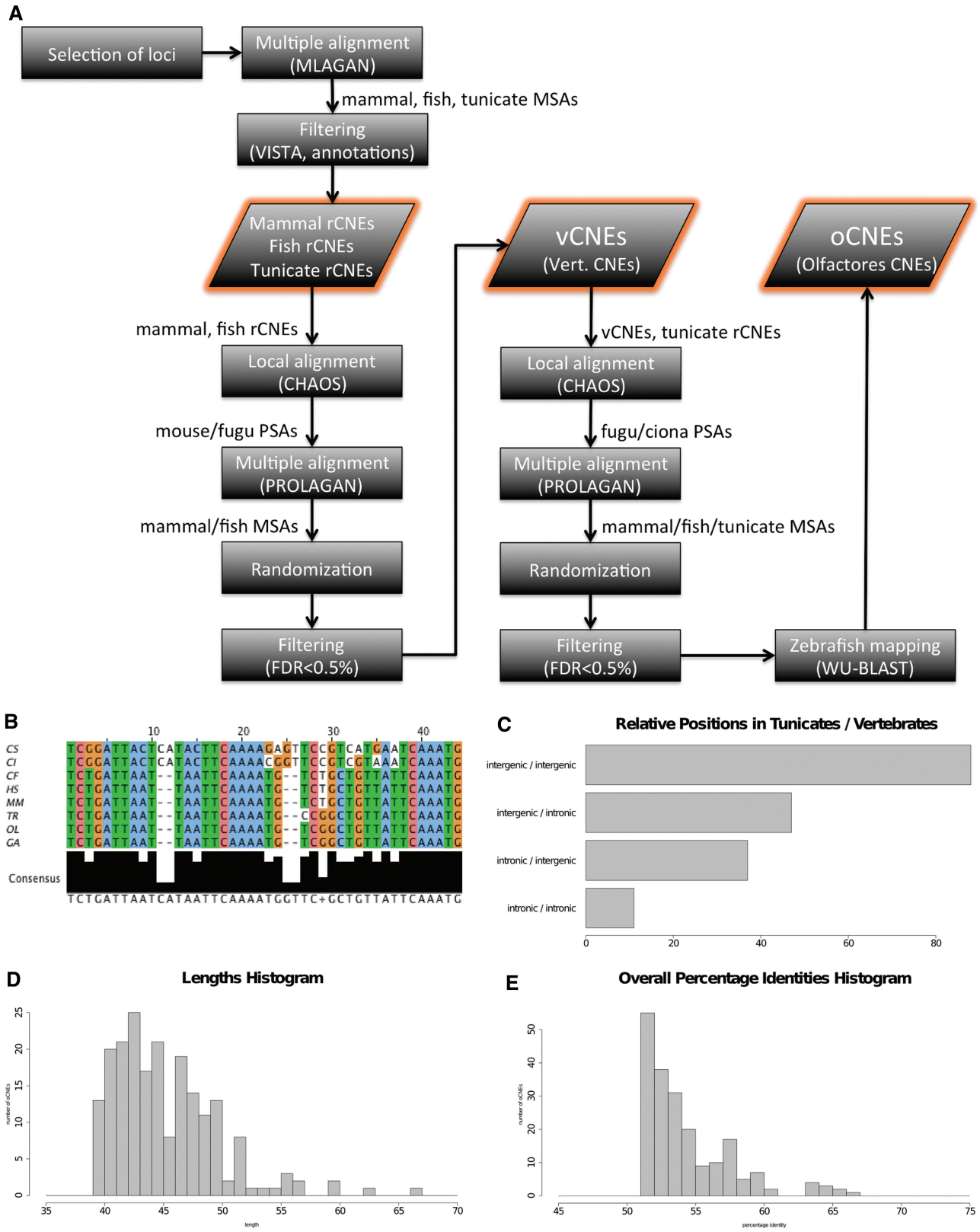


Figure 1. Description of oCNEs workflow and data: Panel **A** shows the schema representing the workflow of the pipeline herein presented. In the boxes are indicated the different steps of the procedure, out of the boxes the input and/or output of each step. MSA: multiple sequences alignment. PSA: pairwise sequences alignment. In **B** is shown an example of the conserved element (oCNE) discovered. Panel **C** indicates the number of oCNEs classified accordingly to their genomic locations relatively to the associated gene structure in tunicates and vertebrates. The majority of elements are conserved in intergenic regions in both organism groups. Finally, **D** and **E** plot the distributions of the length and of the overall percentage identity of the 183 oCNEs.

Table 1. Abbreviations commonly used for conserved elements

Abbreviation	Full name	Organismal group	Reference
CNG	Conserved non-genic elements	Mammals	Dermitzakis *et al.*, Nature 2002
UCE	Ultra conserved elements	Mammals, vertebrates	Bejerano *et al.*, Science 2004
CNE	Conserved non-coding elements	vertebrates	Woolfe *et al.*, Plos Biology 2005
SCE	Shuffled conserved elements	Vertebrates	Sanges *et al.*, Genome Biology 2006
PCNE	Phylogenetically conserved non-coding elements	Vertebrates, vertebrates + amphioxus	Hufton *et al.*, Genome Research 2009
rCNE	Regionally conserved non-coding elements	Mammals, fishes, tunicates	This work
vCNE	Vertebrate conserved non-coding elements	Vertebrates	This work
oCNE	Olfactores conserved non-coding elements	Vertebrates + tunicates	This work

The table indicates the abbreviations most commonly used to refer to conserved elements in this and other articles. For each acronym, we reported the full text, the group of organisms to which the elements are referring and the first article using it.

manually screened the genes flanking and overlapping (if any) oCNEs after having integrated the automatically verified Ensembl *Ciona* annotations with the ones downloaded from the Aniseed database (40). The curated annotations can be found in Supplementary Table S6. We noticed that, using the integrated annotations, a single oCNEs from the whole dataset can be considered as syntenic between vertebrates and ascidians. The element is found in an intron of the FoxP1 gene in both groups. The homology relationship between the two genes was not present in the version of the Ensembl database used in the analyses; therefore, it was classified as non-syntenic. We could not find any other missing relationship.

We also looked for specific duplication of oCNEs elements to check if duplicated elements could be found close to missing orthologous genes. This analysis allowed us to identify 14 duplicated elements of which three are present in all the vertebrate genomes analyzed, two are only found in mammals, two only in fishes, four exclusively in zebrafish (due to additional duplications of loci containing them) and three only in tunicates (see Supplementary Table S7). Elements specifically duplicated in vertebrates, mammals or fishes are associated to paralogous genes, which demonstrate that they were retained after local or whole genomic duplications. In *C. intestinalis*, on the other hand, the three duplicated elements are found in multiple copies within the same genomic region associated to the same genes. The result in *Ciona* could be due in some cases to assembly problems, and in fact in two cases the *C. savignyi* genome contains only a single copy of the same element. One element, however, is present in multiple copies in both genomes.

These results suggest either that oCNEs were eventually shuffled in tunicate genomes or that they were retained after genomic rearrangements and co-opted by different genes. Given the asymmetric design of the starting gene set, for which we used different numbers of genes from each group, we analyzed the number of genes with annotated orthologs, to verify whether the results obtained and the lack of synteny between oCNE containing loci could be due to a lack of inter-group ortholog annotation. Among the total set of 5786 analyzed *Ciona* gene loci, 3957 (68%) have an annotated ortholog in the set of fish analyzed loci (4107 in mammals), while of the 121 loci containing oCNEs, 61 (50%) have an annotated ortholog in fish (63 in mammals), but no conservation outside the coding exons could be detected between syntenic vertebrate and tunicate orthologous loci. The difference between the proportions is significant ($P = 2.1e-05$, Fisher exact test) suggesting that our results are not a random sampling of the starting dataset. These results indicate that oCNEs are found in regions showing significantly less annotated orthologous genes.

oCNEs can act as tissue-specific enhancers

CNEs have the ability to act as tissue-specific enhancers. Therefore, to test if oCNEs may carry enhancer activity, we have tested them in sea squirt and in zebrafish. Three genomic DNA fragment containing oCNEs (see Supplementary Table S5 for corresponding id) were chosen to be tested for specific enhancer activity within developing sea squirt and zebrafish embryos. We selected the genomic locus in *Ciona* containing the highest number of oCNEs. This region is a long intergenic region containing 10 oCNEs, which is found between the *ci0100140718* gene (a gene with no annotated homologs in vertebrates, which appears to be a reductase based on protein domain annotation) and a gene named *Ci-ATBF*. *Ci-ATBF* is a homeobox transcription factor representing the *Ciona* homolog of the mammalian *ATBF1* gene. It is involved in neuronal differentiation in vertebrates as well as invertebrates (64,65). In *Ciona*, *Ci-ATBF* is expressed in mesenchyme, tail epidermis, endoderm, visceral ganglion and nerve cord during development (66). *ATBF1* was previously shown to be associated with a cluster of group-specific conserved elements both in vertebrates and in worms (6). The top three most conserved oCNEs from this cluster of 10 were chosen to be validated (Figure 2A).

The first element (E1, id 1351907, 64% overall percentage identity), in mammals, is contained within a known UCE (enhancer browser id 189). This UCE was tested in transgenic mice by Pennacchio *et al.* (48) and the results in the enhancer browser indicate strong and restricted enhancer function in the neural tube at day 11.5. In mammals and fishes the element is localized in a gene desert upstream of the *Sox21* gene known to promote the progression of vertebrate neurogenesis (Figure 2B) (67). The second tested element (E2, id 1353058, 66% overall percentage identity) shows the highest

conservation score in vertebrates and is found downstream of the *Pax7* gene, which plays a role in neural crest development (Figure 2C) (68). The third element (E3, id 1352705, 60% overall percentage identity) maps upstream of *Prrxl1* (*Drgx*) a transcription factor involved in neuron migration, axonogenesis and nervous system development in vertebrates (Figure 2D). This element is associated with a UCE, which has been tested in mouse and resulted negative at 11.5 days (enhancer browser id 318) (48). Therefore, while these three elements are found inside the same gene desert in *Ciona*, they are present in three different regions among vertebrates. On the other hand, their genomic organization and the functions of the flanking genes are highly similar, as all elements are localized in intergenic regions flanking a transcription factor gene expressed in the developing neural system. These genes are *Sox21*, *Pax7* and *Prrxl1* in vertebrates and *Ci-ATBF* in tunicates.

Genomic fragments containing the *Ciona* sequences of the three selected oCNEs were cloned in Gateway constructs (43) upstream of the pFOG basal promoter and the *LacZ* reporter gene and electroporated in sea squirt embryos (see Figure 2E–G, 'Material and Methods' section and Supplementary Table S2). To verify if the elements could be categorized as positive enhancers, we calculated the total percentage of embryos expressing the reporter and used a minimum cutoff of 25% positive embryos as done in similar studies (69). From this analysis, the 600-bp fragment containing the E1 element showed strong enhancer activity exclusively in the mesenchyme (Figure 2E; 61% positive embryos). The 900-bp fragment containing the E2 element also resulted to be a functional enhancer, albeit weaker and in variable tissues (epidermis, muscle, mesenchyme and notochord), where the most representative and specific staining was in two cells at the tip of the tail (Figure 2F; 28% positive embryos). Finally, the 300-bp fragment containing the E3 gave weak mesenchyme staining in a lower number of embryos (20%) and was considered negative. Interestingly, the patterns of expression driven by the E1 and E2 constructs are in good agreement with the expression pattern of *Ci-ATBF* at the tailbud stage (66).

To evaluate the regulatory potential of these sequences in a vertebrate context, genomic fragments containing the zebrafish sequences of the same oCNEs were inserted in a construct containing a zebrafish *hsp70* minimal promoter and the *venus* reporter gene and microinjected into zebrafish embryos (see 'Material and Methods' section and Supplementary Table S2). Subsequently, fluorescence reporter activity was detected to assess their enhancer function (Figure 2H–L). Results from the 500-bp fragment containing E1 showed expression (44% of the embryos) of the reporter mainly in the telencephalic region but also extending posteriorly to the hindbrain in agreement with the expression pattern of *sox21* in zebrafish (Figure 2H). The 500-bp fragment containing the E2 element drives broad expression (67%) in the anterior neural tube, more specifically in whole forebrain and hindbrain regions. Additionally, ~20% of the injected embryos showed enhanced reporter expression in the skeletal muscle, which was not observed with the enhancer control or E1 construct injected embryos, suggesting that in addition to neuronal enhancer activity, the E2-containing fragment possesses a skeletal muscle–enhancer activity in this transient transgenic reporter assay (Figure 2I). The observed expression pattern was similar to that of the *pax7a* gene, upstream of which E2 is located. The 1100-bp genomic fragment containing E3, on the other hand, showed weak activity similar to the enhancer control construct with no emerging tissue-specific pattern (Supplementary Figure S1 for a complete outcome of the zebrafish experiments), and thus this element was not considered to act as an enhancer, similarly to the case for the corresponding *Ciona* element (Figure 2L and Supplementary Figure S1). These results show that two out of three genomic fragments containing the selected oCNEs have the ability to work as enhancers both in vertebrates and in tunicates, and their patterns of expression is in agreement with that of neighboring genes. The third element was consistently found not to act as an enhancer in both organisms tested and in mouse via the enhancer browser, suggesting it might require testing in different developmental stages, or that it might have other functions which were not tested. Interestingly, according to the Broad HMM classification based on chromatin states in several cell lines, the corresponding human genomic region is classified in the UCSC genome browser as a poised promoter in ES cells, i.e. presents bivalent histone marks, H3K27me3 and H3K4me1/2 (data not shown).

oCNEs can act as enhancers in cross-transgenesis experiments

We also performed cross-transgenesis experiment to evaluate if the fragments were capable to work in a different background. *Ciona* fragments were injected in zebrafish, and similarly, zebrafish elements were tested in *Ciona*. All *Ciona* elements enhanced the activity of a minimal promoter when tested in zebrafish embryos (Figure 2P–R and Supplementary Figure S2). Conversely, all zebrafish elements showed activity in *Ciona* embryos (Figure 2M–O and Supplementary Figure S3). The E1 fragments showed the strongest activity in these cross-transgenesis experiments. Interestingly, both *Ciona* and zebrafish E1 fragment displayed highest activity in fish anterior neural tissue. *Ciona* and fish E1 also showed overlapping activity in *Ciona* mesenchyme. Likewise, the *Ciona* fragment containing E2 showed neuronal and muscle activity in zebrafish experiments, similar to what is observed with zebrafish E2. Thus it seems that the two strongest oCNE enhancer we tested, the E1 and E2 elements, have conserved at least some of their cis-regulatory specificity between the two species. This cis-regulatory activity is, however, differently interpreted in the two organisms, brain in zebrafish versus mesenchyme or muscle in Ciona, possibly as a result of changes in the expression profiles of trans regulators between these two species (18).

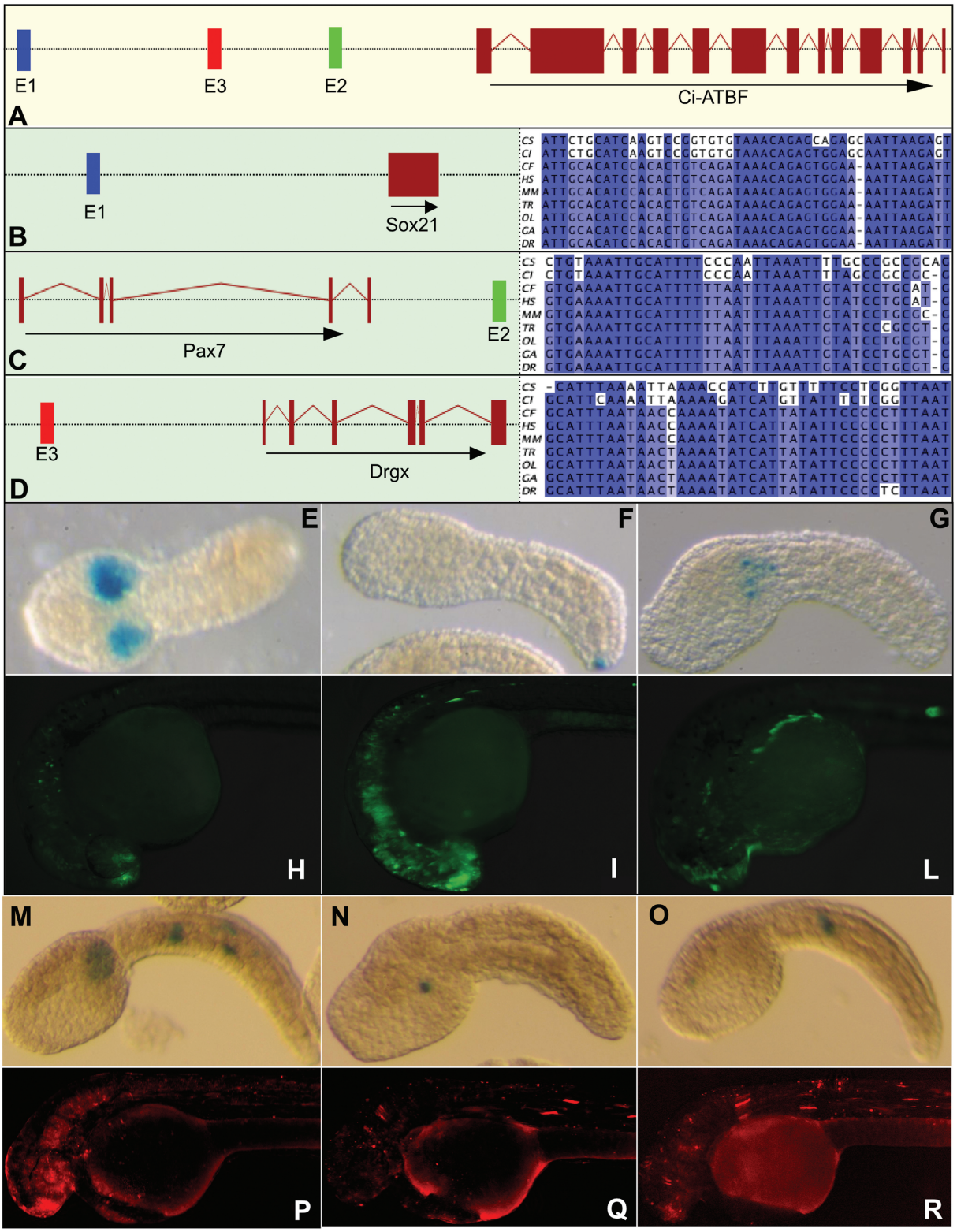


Figure 2. Functional validation of oCNEs enhancer function: three oCNEs were selected to be validated (E1, E2 and E3). Schemas represent the genomic intervals containing the selected oCNEs and the reciprocal positions of the elements and the associated genes. For clarity purposes the schemas are not respecting a specific scale. Panel **A** indicates that the three chosen elements are present in the same intergenic region in tunicates associated to the *Ci-ATBF* gene. In **B–D** are reported the three distinct vertebrate intervals containing the selected conserved sequences E1, E2, E3 and the respective oCNE alignments in all the analyzed species. Pictures **E–G** report the most representative expression pattern driven by the elements E1, E2 and E3, respectively, in *C. intestinalis*. Pictures **H, I, L** report the most representative expression pattern driven by the elements E1, E2 and E3, respectively, in *D. rerio*. **M–O** report the most representative expression pattern driven by the zebrafish elements E1, E2 and E3, respectively, injected in *C. intestinalis*. Finally **P–R** report the most representative expression pattern driven by the Ciona elements E1, E2 and E3, respectively, injected in zebrafish embryos.

oCNEs overlap ultraconserved elements and are enriched in transcribed enhancers

UCEs are extremely conserved non-coding sequences, whose function is not yet completely understood. They were shown to regulate transcription in mammal embryos as well as to be transcribed (48,70). UCEs are mostly associated with developmentally regulated genes and are enriched in gene deserts (2). The observation that two out of three validated genomic fragments containing oCNEs overlap UCEs in vertebrates led us to verify the overall proportion of oCNEs overlapping UCEs. Therefore, we calculated the overlap of the oCNEs with the extended set of 5404 vertebrate UCEs reported by Stephen *et al*. (47) discovering that the majority of oCNEs overlap known vertebrate UCEs (145 out of 183, ~80%). The overlap is significantly higher than the overlap between vCNEs and UCEs (499 out of 900, 55%, $P = 3.8e-09$) pointing out that this is an oCNEs-specific enrichment. It is important to point out that such overlap could also be partially explained by our multi-steps procedure, which progressively selects for the highest conserved segments within each group. However, the usage of stringent parameters in the randomization steps rejects the hypothesis that these elements are conserved merely by chance. Extensive information about the functional validation of UCEs can be found in the enhancer browser (48). The database is composed of 1756 elements of which 50% (887) resulted to be positive as enhancers at the developmental stage tested. We calculated the overlap of these elements with the set of oCNEs. The results show that 85 out of 183 oCNEs overlap a conserved element tested in the enhancer browser. Of these, 66% (56) resulted to be functional enhancers in mouse. The difference between the two proportions is significant ($P = 3.9E-03$) indicating that oCNEs are enriched for elements found to be functional enhancers in mouse at developmental stage 11.5E. Several studies indicate that enhancers and UCEs can also be transcribed (49,70,71). To verify the possibility that oCNEs could act also at the transcript level, we tested the overlap between them and the published set of transcribed enhancers by Kim *et al*. (49). They showed that a subset of stimulus-dependent enhancers from mouse cortical neurons also show activity-regulated RNAPII binding, and therefore they are transcribed (49). The total set of intergenic functional enhancers contains 5117 genomic positions; of these, 2052 are also transcribed and are named eRNAs, while 3065 are not transcribed. The overlap between these two sets of enhancers and oCNEs shows that 15 oCNEs overlap eRNAs and only three overlap non-transcribed enhancers (Figure 3A). The differences between the overlaps in the two classes suggest that oCNEs are enriched for eRNAs ($P = 0.0078$). This is not a bias given by *a priori* enrichment in the vCNEs group, and indeed the same analysis in vCNEs group yield a similar proportion of conserved elements overlapping eRNAs and non-transcribed enhancers ($P = 0.57$). This result suggests that oCNEs could be enriched for a specific class of enhancers also able to be transcribed.

oCNEs can be transcribed

To validate if the identified elements are transcribed, we carried out RT-PCR from RNAs collected at four different stages of zebrafish development. Experiments were executed on the three elements for which we tested already the enhancer function. All the elements were found to be transcribed in a dynamic manner during development. While E1 and E2 appear to be expressed after shield stage, E3 starts to be transcribed at gastrula stage (Figure 3B). All three elements show a peak of transcription at 24 hpf, hinting at a potential role for transcription during the late gastrulation stages. Similar analyses were carried out in mouse as well as in sea squirt. In mouse, we tested expression at 8.5, 12.5 and adult stages. Expression was not detected at 12.5 and adult stages (data not shown), and conversely, expression of E1 and E2 but not E3 was detected at 8.5 (Figure 3C). Transcription of the three elements in *C. intestinalis* at the tail-bud stage showed expression of elements E2 and E3 (Figure 3D). These findings suggest that the identified oCNEs can be expressed during development in Olfactores and that the transcripts are produced at low levels as already shown for eRNA and more generally non-coding RNAs (49,72). Overlap analyses of zebrafish oCNEs against recently published dataset of RNA-seq data (73,74) did not give significant results, suggesting that their weak expression levels could need higher depth of sequencing to be detected.

To better understand if oCNEs transcripts could be associated to short or long RNAs, we analyzed the overlap of oCNEs with the UCSC human genome browser tracks collecting transcribed contigs from the ENCODE/CHSL RNA-seq data on Long (98 samples, 117 388 194 transcribed contigs) and Short (84 samples, 5 452 981 transcribed contigs) RNAs (53–55). The number of contigs overlapping each element indicates the number of samples in which an oCNEs overlap a transcribed region supporting the putative transcription of the element. Mapping of the oCNEs was compared with the mapping of a similar number of randomly selected elements. The results obtained indicate that 158 oCNEs overlap 4866 RNA contigs from the Long RNA-seq dataset, while 147 random elements overlap 3191 contigs. The difference between the two dataset is significant ($P = 2.5E-77$). The average number of samples in which each oCNEs result expressed in the Long RNA-seq dataset is ~30, while it is ~20 in the random set (Supplementary Figure S4). In the same analysis, using the Short RNA-seq data the oCNEs resulted to map on 18 contigs, while random regions overlap with seven; this difference is not significant ($P = 0.08$). We thus hypothesize, in the light of these results, that oCNEs are unlikely to be transcribed as short RNAs.

Functional enrichment analyses suggest oCNEs as hubs of homeobox gene regulatory networks

Genes and sequences associated with oCNEs were analyzed to define functional enrichments that could shed light on their specific cellular functions and origins.

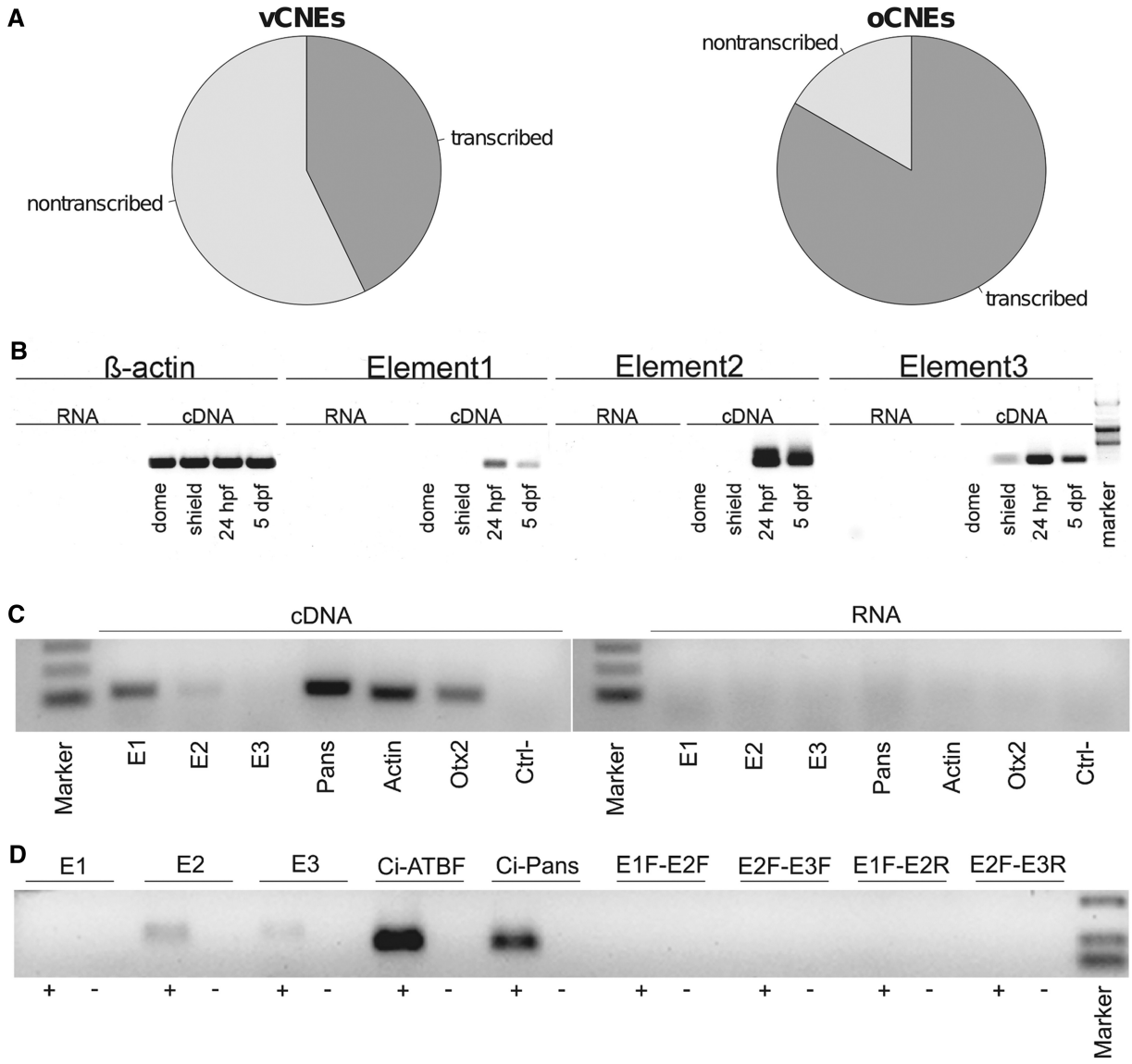


Figure 3. Potential transcription of oCNEs: The oCNEs result to be enriched for eRNAs. Pie-charts in (**A**) show how vCNEs and oCNEs overlapping enhancers from Kim *et al.* (49) segregate between the classes of eRNAs and non-transcribed enhancers. Twenty-eight vCNEs overlap enhancers from Kim *et al.* and ~40% of them overlap eRNAs. Conversely, 18 oCNEs overlap enhancers from Kim *et al.* and >80% of them overlap eRNAs. The three oCNEs used for the validation of the enhancer function were also validated for transcriptional activity in *D. rerio* (**B**), *M. musculus* (**C**) and *C. intestinalis* (**D**). Primers were designed to amplify a fragment of ~100 bp around each element. As positive control, we used the following coding transcripts: *bActin* (*D. rerio* and *M. musculus*), *Otx2* (*M. musculus*), *Ci-atbf* (*C. intestinalis*). Non-coding transcripts used were as follows: *Ci-Pans* (*C. intestinalis*) (51) and Pans (*Mm.221244*, the murine homolog of *Ci-Pans*) (52). All the used controls are known to be expressed at the time of the sampling. As negative control, we used DNAseI-digested RNA (indicated as RNA in B and C and as '-' in D). In *C. intestinalis*, we also used different combinations of the forward/reverse primers. The absence of signal in the cDNA template PCR is indicative of the absence of genomic DNA contamination in the cDNA preparation demonstrating that the amplicons are real RNA products.

All the functional associations were analyzed considering as reference 'universe' the groups of genes (or sequences) containing at least one rCNE in sea squirt or one vCNE in mouse. Such a strategy is necessary to avoid false enrichments resulting from the fact that these elements are primarily conserved inside each group. First of all, we decided to verify whether genomic regions containing a specific oCNE were enriched for genes containing the same specific protein domains both in vertebrates and tunicates. Therefore, domain enrichment analyses were performed by (i) identifying if and which length interval in mouse and sea squirt showed significantly enriched frequency of common domains; and (ii) checking for the specific significantly enriched domains. The protein domains identified from genes transcribed in genomic intervals containing each oCNE were compared with those identified in randomly paired vCNE/rCNE regions. We performed the analysis over three length intervals around oCNEs, and the significance of the associations decreases proportionally with respect to the extension of the window, disappearing at ~1 Mb (Supplementary Figure S5), in line with previous

observations for the range of action of long-distance enhancers (62). Focusing on a window of 500 kb (adjusted $P = 0.03$), the common domains resulting significantly enriched in oCNE regions as opposed to random vCNEs/rCNEs pairs are the homeobox (adjusted $P = 0.02$) and the helix-turn-helix (HTH) lambdarepressor (adjusted $P = 0.02$), as shown in Figure 4A. The homeobox gene superfamily encodes transcription factors that act as master regulators of development through their ability to activate or repress a diverse range of downstream target genes (75). The HTH domain is a common denominator in basal and specific transcription factors from the three superkingdoms of life and is frequently present in homeobox genes (76).

Then, to check if oCNEs may indicate a common conserved regulatory mechanism, we performed a similar analysis focused on transcription factor binding site enrichments taking into account as significant only binding sites significantly enriched in all the groups of organisms. To this aim we used the transcription factor binding matrices from the Jaspar Family database (56), which provides generic matrixes for major families of transcription factors. We found common significant enrichments for binding sites recognized by the homeobox (*Ciona* adjusted $P = 1.0E-13$), the high mobility group (HMG; *Ciona* adjusted $P = 4.3E-05$) and the forkhead (*Ciona* adjusted $P = 1.0E-03$) transcription factors classes within oCNEs sequences (see Figure 4B for results in *Ciona* and Supplementary Table S8 for results in all the tested species). Interestingly, the HMG proteins are a superfamily of nuclear proteins that bind to DNA and nucleosomes and induce structural changes in the chromatin. They are important in chromatin domains dynamics and in regulating the expression of specific genes during development (77). Forkhead box (Fox) proteins are a superfamily of evolutionarily conserved transcriptional regulators, which control a wide spectrum of biological processes and are heavily used in developmental processes (78). Finally, GO enrichment analysis was performed on the set of genes associated with oCNEs and compared with the genes associated with rCNEs found in *C. intestinalis*. GO classifications for *C. intestinalis* were extracted from the Aniseed annotation database (see 'Material and Methods' section). Figure 4C shows the GO classes resulting specifically enriched in *Ciona* oCNEs: multicellular organismal development (adjusted $P = 6.58E-06$), sequence-specific DNA binding (adjusted $P = 1.17E-05$), transcription (adjusted $P = 0.0007$), cell differentiation (adjusted $P = 0.0008$), transcription factor activity (adjusted $P = 0.008$) and calcium ion binding (adjusted $P = 0.017$). Taken together, these results clearly indicate that the genes surrounding oCNEs as well as the transcription factors potentially binding oCNEs are significantly associated with genes involved in development and, more specifically, to morphogenesis and differentiation and these enrichments are significantly more specific than the ones related to rCNEs. GO enrichment analyses performed in mouse using either DAVID (58) or FATIGO (59) gave similar results when we compared oCNEs or vCNEs with rCNEs, but no enrichment was found comparing mouse oCNEs with vCNEs, suggesting that, in vertebrates, oCNEs and vCNEs belong to similar functional classes (data not shown).

Conservation in the *Oikopleura* and amphioxus genomes

To understand if oCNEs are retained in other sequenced model chordates, we searched for their presence in the *Oikopleura dioica* and amphioxus (*Branchiostoma floridae*) genomes. The pipeline we presented needs at least two sequenced and well annotated genomes belonging to the same class to analyze that specific class of organisms, and therefore we could not analyze them using our pipeline. Moreover, as we did not detect Ciona oCNEs sequences in the *Oikopleura* and amphioxus genomes by using Blastn, we decided to use information from all the organisms in oCNE blocks to build HMM (41) matrices from each oCNE multiple alignment based on the sequences conserved in all the analyzed species. We then scanned the *Oikopleura* and amphioxus genomes with the HMMs thus generated. This search yielded nine conserved elements in the *Oikopleura* and 13 in the amphioxus. Genes flanking and overlapping the elements thus discovered were annotated using Blast2GO (42) and considered as putative target genes. Annotations were manually checked against *Ciona* and mouse overlapping and flanking genes for each respective element (see Supplementary Table S6). Again, these elements resulted not to be located in the vicinity of evolutionarily related genes, although they appear to be associated to genes functionally related. Indeed, according to the Blast2GO classification, the top scoring biological processes represented in the associated genes are related to development and regulation (Supplementary Table S9). The number of conserved elements is small, and therefore, we cannot test for significance; however, the biological functions annotated by Blast2GO are remarkably similar to those enriched in the 183 original oCNE dataset. It is particularly interesting the presence of oCNEs in *Oikopleura* genomic loci containing putative orthologous for the Bmp and Lim homeobox genes. Indeed, these genes are also associated to oCNEs in vertebrates and ascidians. In the amphioxus, interesting genes associated to oCNEs are the putative homologs of Jumonji, Argonaute and Znf729. We conclude that only a small number of oCNEs is represented in the *Oikopleura* and amphioxus genomes, and these elements are not syntenic with ascidians or vertebrates but, again, their genomic loci result to be associated to functionally similar regions.

Finally, we checked if oCNE neighborhoods lacking synteny between *Ciona* and vertebrates could show evidences for common origin when taking into account the genome of amphioxus (i.e. are close together on the amphioxus genome). The results indicated that only three oCNEs could be associated to pairs of putative target *Ciona*/mouse genes localized on the same scaffold in the amphioxus genome (with a distance between them of 263 617, 2 448 029 and 898 799 bp). Randomizations showed that this result is not significant (1000 randomization produced an average of 6.4 associations with a standard deviation of 2.8). Interestingly, Hufton *et al.*

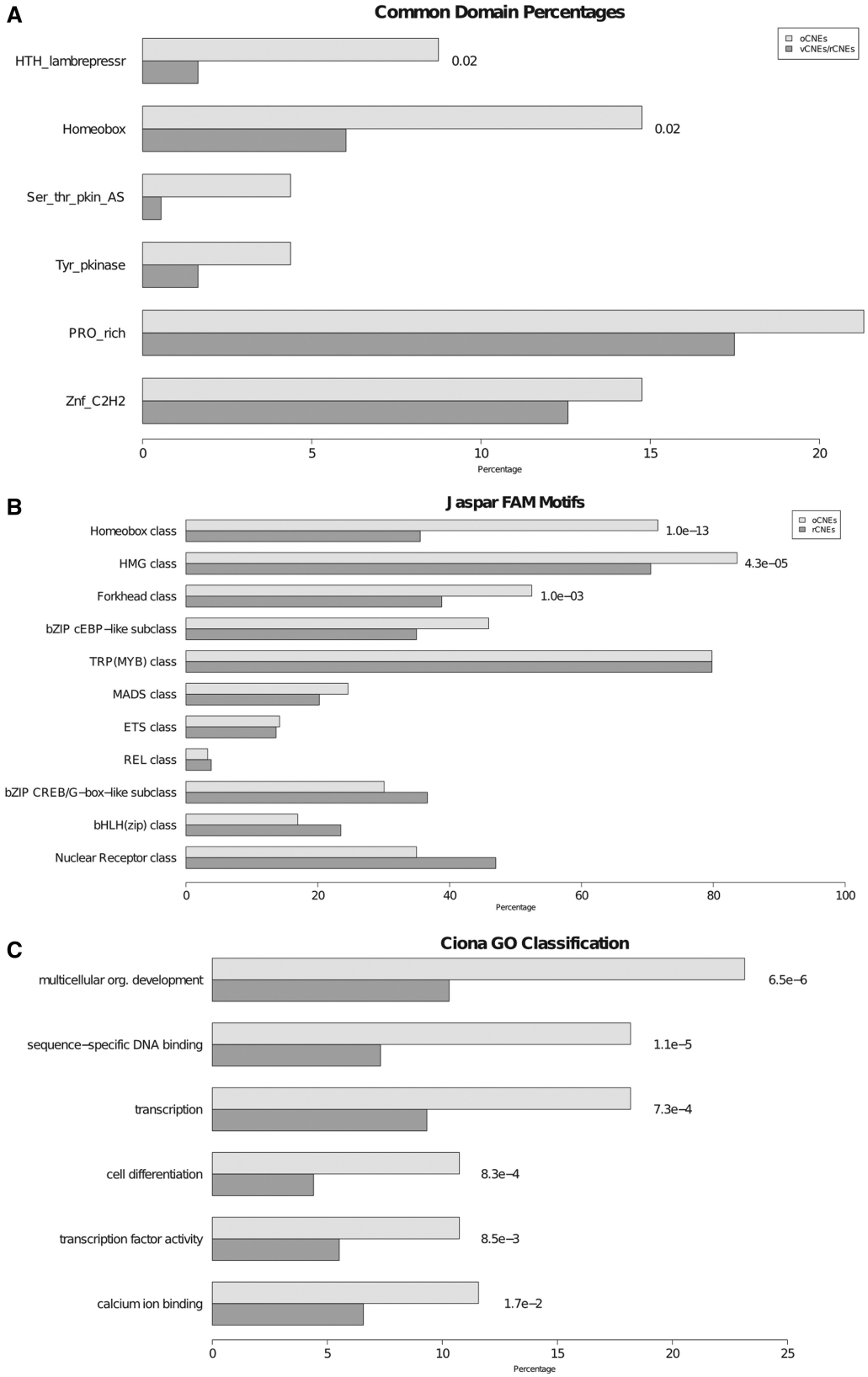


Figure 4. Functional enrichment analyses: **A** shows, for each domain, the percentage of oCNEs (light grey) and vCNEs/rCNEs random couples (dark grey) falling in intergenic regions associated to genes containing the same specific domain in all the species analyzed. Only domains for which the percentage is higher in oCNEs are reported. Adjusted *P*-values of the differences between the two groups are reported only if significant. Panel **B** shows, for each Jaspar fam motif, the percentage of *C. intestinalis* oCNEs (light grey) and *C. intestinalis* rCNEs (dark grey) containing at least one binding site for the specific motif. Adjusted *P*-values of the differences between the two groups are reported only if significant. **C** shows GO enrichments for each GO class associated to genes flanking tunicate oCNEs (light grey) and tunicate rCNEs (dark grey). Only oCNEs-associated significantly enriched classes are reported with the respective adjusted *P*-values.

reported the discovery of >1000 CNEs [defined as phylogenetically conserved non-coding elements (PCNEs)] among vertebrates or between vertebrates and amphioxus. Out of 183 oCNEs, 122 overlap the published set of vertebrate PCNEs, and 42 of them overlap the set conserved between vertebrates and amphioxus (data not shown). These PCNEs are conserved collinearly between vertebrates and amphioxus as a result of the methodology adopted. Our HMM approach could only map 4 out of these 42 oCNEs in Amphioxus, despite identifying some non-syntenic well-conserved oCNEs in this organism. This is probably because of the fact that the alignments by Hufton *et al.* were produced in a locus-specific way and with an estimated false-positive rate between 2 and 10% (based on two randomizations) as compared with our oCNE analysis, which was performed genome-wide at an FDR of 0.05%, and our HMM search, which was calibrated at high stringency, i.e. to yield only the original oCNE and close paralogs within the genome of origin, and thus only similar conserved elements in other genomes.

DISCUSSION

In this study, we developed a pipeline capable to identify, for the first time, CNEs spanning Olfactores genomes. Our analysis resulted in a set of 183 conserved non-coding blocks (oCNEs). We showed that oCNEs mainly overlap previously published UCEs and, although they are syntenic among vertebrates, they are found in non-syntenic loci in tunicates. Nevertheless, oCNEs are significantly associated with homeobox containing genes and genes involved in organismal development; also, they are significantly enriched for binding sites recognized by homeobox transcription factors. Such preponderance of homeobox genes associated to oCNEs, in the genomic context as well as in binding site predictions, could indicate a complex network of interactions which, during development, involve reciprocal regulatory relationship within this family of genes. The players of this network (usually defined as the 'input') appear to be the same genes in all the animal groups studied, but the regulatory interactions and the domains of expression encoded within these networks (often seen as the 'output'), appears to be different in distant groups [see Cameron and Davidson (26) for a first proposal of the input/output theory]. Genomic fragments containing oCNEs act as domain-specific enhancers in developing embryos of sea squirt, mouse and zebrafish without retaining the same domain specificity between the groups. The cross-transgenesis experiments indicate that despite the long evolutionary distance separating the species under investigation, conserved oCNEs can retain enhancer effect in cross-species analysis and support the functional significance of these conserved sequences. While the specificity of enhancer effects is not fully retained, at least in the case of *Ciona* E1, anterior telencephalic activity is enriched in zebrafish, which is reminiscent to the zebrafish orthologous element resulting mostly specific to the anterior telencephalon. It is noteworthy that all elements tested appear to enhance the activity of a minimal promoter in fish as well as in *Ciona*. We chose to amplify larger fragments because the conservation between vertebrates and ascidians is limited to short sequences of ~50 bps, which is unlikely to reflect the minimal functional unit. Consistent with this expectation oCNEs are anchored in longer regions conserved within each respective group. Thus oCNEs might represent a part of a specific regulatory element which, to work, would need support from sequence elements found in the flanking regions.

With constant refinements in the technologies capable to detect non-abundant transcripts, the observations that a large number of enhancers are also transcribed are tangibly increasing (49,54,70,71), suggesting that, at least in mammals, thousands of enhancers are transcribed. Interestingly, the oCNE dataset also shows significant overlap with the eRNA dataset. This enrichment is not a bias determined by the composition of vCNEs, indicating that oCNEs probably belong to a specific class of enhancers, which can also be transcribed. Furthermore, we indicate, by analyzing a large number of publicly available ENCODE datasets, that they are unlikely to transcribe short RNAs. It should be noted that for most eRNAs and UCEs analyzed, the full length and nature of the RNA molecules transcribed by these regions remains a largely unresolved question. Indeed, in this work, we demonstrated that oCNEs can effectively be transcribed even if we have not directly addressed the functional association between the transcription and the enhancer function. Further and more in-depth validations would need to be conducted to verify the extent, nature and specificity of oCNE expression.

It is important to specify that our results depend heavily on the methodology we used to identify oCNEs and that some homology relationships might be missing from current annotations. This raises the question whether oCNEs might be identified by mere chance. Our randomization-based filtering approach, which makes use of stringent FDR criteria indicating that <1 oCNE could be false, is pointing against this idea. On the contrary, given that other approaches were performed with more lenient statistical stringency, it is possible that we have missed some bona fide oCNEs, which might warrant future investigation. Similarly, our HMM search of oCNEs in other species such as amphioxus was performed stringently and might thus miss related and relevant CNEs, which could have diverged beyond the stringency of our approach. Manual curations of results and the significant overlaps with other relevant datasets such as eRNAs, UCEs, ENCODE data and the experimental evidence we produced are further proof of oCNEs' biological relevance. A different and altogether more complex issue is to what extent oCNE-like elements could arise by convergent evolution. We do not have sufficient data to tackle appropriately this issue but we speculate that it could be unlikely if we consider a parsimonious scenario for the evolution of such elements. Finally, assembly errors could have generated some of the extensive non-orthologous shuffling we have observed. This is an important concern to address because

many of these elements are found in gene deserts in which the lack of gene annotations can cause a higher proportion of assembly errors. However, in our pipeline this is unlikely because oCNEs originate from regionally conserved collinear regions in each group of organisms. Thus, to make an assembly error responsible for the generation of an oCNE, the same error should have occurred twice in the same collinear manner in at least two different organisms, which we believe to be highly improbable. It is possible, though, that assembly errors could cause some artificial duplication within the same genomic region of similar oCNEs, as seen in the duplication analysis within Ciona.

So, how can we explain the fact that such conserved regions are not conserved in a collinear fashion? The sequencing of new genomes could help us in shedding light on this point. Classically, CNEs are considered collinear regulatory regions conserved among lineages in terms of their position as well as in terms of their association to target genes whose sequences are conserved in their respective lineage but not among different lineages (6). oCNE elements do not appear to belong to this class, because they are well conserved among different lineages in terms of sequence while not being collinear. This is supported by the observation that, genes associated to oCNEs are significantly enriched for groups of genes in ascidians lacking clear vertebrate orthologs. Although they are not associated to the same potential target gene, they appear to maintain a clear preference for certain functional classes of genes. Despite a longer divergence time between amphioxus and vertebrates compared with *Ciona* and vertebrates, the conservation of synteny with vertebrates is greater for amphioxus than for *Ciona* (16). About 74% of amphioxus scaffolds show a significant presence of orthologs from the same human chromosome, while in *Ciona*, this proportion is ~9%. The *Oikopleura* is the only known chordate genome to show no significant conservation of gene neighborhood with other chordates (79). Our sensitive pipeline has been able to find a single collinear element conserved between vertebrates and ascidians, and analysis in the amphioxus and *Oikopleura* genomes show the presence of a minority of non-collinear oCNEs. Such observations lead to speculation that these elements could have been present in a chordate ancestor and have been differentially lost or co-opted by different genes during the dramatic changes that brought to the differentiation of the chordate lineages. Particularly intriguing are the findings that early vertebrate whole genome duplications were predated by a period of intense genome rearrangement (80) and that, in addition to whole genome duplications, segmental and single-gene duplications shaped the genomes of extant vertebrates (81). A mechanism that can be taken into account for the generation of non-syntenic conserved elements in such a scenario can be accounted by partial rediploidization following local- or whole-genome duplications, which, in vertebrates, have been demonstrated to be at the basis of the retention of regulatory regions deriving by exons of lost duplicated genes (82). We screened oCNEs for specific overlap to cDNAs and single whole genomes to understand if they could result from rediploidization events but no such results were found. A different scenario to justify the unexpected variability observed in oCNEs, in terms of their location as well as of their expression domains, could be addressed to several peculiarities of the tunicate genomes. First, tunicate genomes are highly re-arranged and experienced extensive gene losses as compared with the non-duplicated early chordate karyotype. Putnam *et al.* (16) have identified 8437 gene families with members in amphioxus and other chordates that represent the descendants of genes found in the last common chordate ancestor. They also estimate that subsequent family expansions have generated ~13 000 genes in amphioxus and vertebrates and ~7000 in *C. intestinalis*. The lower number of tunicate genes is believed to be due to an extensive gene loss, which caused ~2000 genes to be lost (83). The families of transcription factors that have lost the highest proportion of orthologs in tunicates are the homeobox, high-mobility group (HMG) and helix-loop-helix (HLH) [see (84) and its supplementary for a complete list of references and genes]. Intriguingly, these are the same gene families, which appear to be enriched in oCNEs. Hence, another mechanism that could justify the shuffling of oCNEs is that it could be associated with tunicate-specific gene losses and subsequent genomic rearrangements. If oCNEs were present in the chordate ancestor, they were probably co-opted by non-homologous but functionally similar genes, in tunicates, after the loss or the extreme derivation of the originally associated ones. A recent study shows that the roles of some Hox genes are not homologous to their vertebrate counterparts during *Ciona* larval development, further supporting the evidence that functional homology between tunicate and vertebrate genes is not always observed (85). In addition, gene expression dynamics of orthologous genes between developing *C. intestinalis* and *D. rerio* embryos were shown to be broadly divergent (18). Further support along this line is given by the fact that Hox and ParaHox genes in *C. intestinalis* are not organized in clusters, do not retain spatial and temporal developmental gene expression collinearity and contain transposable elements in their genomic loci (86,87). To us, this level of genomic and proteomic variability, unique to tunicates, could have occurred in concomitance with a peculiar rewiring of regulatory modules aimed at maintaining the chordate body plan. A final mechanism, which could be used to justify the shuffling of such elements, derives by the observation that they can be actively transcribed. Indeed, given that any type of RNA can serve as template for reverse transcription (88), the fact that oCNEs are transcribed suggests that they could have also been retrotransposed in new locations by the same mechanism involved, for example, in the creation of pseudogenes.

We thus propose that these conserved elements were shuffled either in an active (retroposition) or passive (rearrangements, rediploidization, derivation) fashion and co-opted by similar genes. The necessity for them to be shuffled is likely to have arisen during evolution of chordates to accommodate the coding variability, extensive gene gains and losses, genomic re-arrangements and

the establishment of different developmental times to maintain a similar body plan for all the chordates.

Unfortunately, the impossibility to find genomic relics of shuffling events related to oCNEs makes it extremely difficult to demonstrate which mechanism took the leading part in their evolution. We searched for any such relics, but did not find any enrichment for specific k-mers, repeats, pseudogenes, chromatin interaction features in the genomic intervals overlapping or surrounding oCNEs, nor did oCNEs result to be derived by lost coding or non-coding exons (data not shown). When more chordate genomes and transcriptomes will be sequenced, it will be possible to answer more in-depth questions related to the evolutionary history of chordate regulatory elements. Nevertheless, the analysis herein presented is the first report of a sensitive and stringent pipeline that could be adopted to look for conservation of non-coding elements in distant and derivate groups of genomes as soon as new genomes are published. Moreover, the data provided constitute the first collection of non-coding elements conserved among Olfactores and represent an extremely valuable resource for future comparative, evolutionary and developmental studies. Finally we provide initial evidence that oCNEs can act as enhancers (also in cross-transgenesis) and are transcribed in different organisms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–9 and Supplementary Figures 1–5.

ACKNOWLEDGEMENTS

The authors would like to acknowledge John Mattick for providing the extended set of UCEs data; Michael Greenberg, Tae-Kyung Kim and Jesse Gray for clarifications about the eRNAs dataset; Paolo Sordino and Salvatore D'Aniello for critical reading of the manuscript; Michael Brudno, Stefano Gustincich, Piero Carninci, Mariella Ferrante, Graziano Fiorito, Fiona McNish, Daniel Zerbino, Alison Meynert, Benedict Paten and Wolfgang Huber for helpful discussions and support. A special acknowledgement also goes to the late Parvesh Mahtani, who shared our enthusiasm for this project. We would like to thank two anonymous referees for their invaluable suggestions.

FUNDING

Seventh Framework Program of the European Commission (DOPAMINET and ZF Health projects) [223744 to E.S., F.M., P.L. and R.S., 242048 to F.M.]; the Italian 'Progetto Bandiera RITMARE' (to R.S.) R.S. has been supported by a Marie-Curie Research Early Stage Training Fellowship to develop part of the project at EBI. Funding for open access charge: RITMARE.

Conflict of interest statement. None declared.

REFERENCES

1. Dermitzakis,E.T., Reymond,A., Lyle,R., Scamuffa,N., Ucla,C., Deutsch,S., Stevenson,B.J., Flegel,V., Bucher,P., Jongeneel,C.V. *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, **420**, 578–582.
2. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
3. Woolfe,A., Goodson,M., Goode,D.K., Snell,P., McEwen,G.K., Vavouri,T., Smith,S.F., North,P., Callaway,H., Kelly,K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.
4. Plessy,C., Dickmeis,T., Chalmel,F. and Strähle,U. (2005) Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends Genet.*, **21**, 207–210.
5. Vavouri,T. and Lehner,B. (2009) Conserved noncoding elements and the evolution of animal body plans. *Bioessays*, **31**, 727–735.
6. Vavouri,T., Walter,K., Gilks,W., Lehner,B. and Elgar,G. (2007) Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol.*, **8**, R15.
7. Kermekchiev,M., Pettersson,M., Matthias,P. and Schaffner,W. (1991) Every enhancer works with every promoter for all the combinations tested: could new regulatory pathways evolve by enhancer shuffling? *Gene Expr.*, **1**, 71–81.
8. Kirchhamer,C.V., Yuh,C.H. and Davidson,E.H. (1996) Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. *Proc. Natl Acad. Sci. USA*, **93**, 9322–9328.
9. Visel,A., Akiyama,J.A., Shoukry,M., Afzal,V., Rubin,E.M. and Pennacchio,L.A. (2009) Functional autonomy of distant-acting human enhancers. *Genomics*, **93**, 509–513.
10. Hufton,A.L., Mathia,S., Braun,H., Georgi,U., Lehrach,H., Vingron,M., Poustka,A.J. and Panopoulou,G. (2009) Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. *Genome Res.*, **19**, 2036–2051.
11. Aparicio,S., Morrison,A., Gould,A., Gilthorpe,J., Chaudhuri,C., Rigby,P., Krumlauf,R. and Brenner,S. (1995) Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proc. Natl Acad. Sci. USA*, **92**, 1684–1688.
12. Royo,J.L., Maeso,I., Irimia,M., Gao,F., Peter,I.S., Lopes,C.S., D'Aniello,S., Casares,F., Davidson,E.H., Garcia-Fernández,J. *et al.* (2011) Transphyletic conservation of developmental regulatory state in animal evolution. *Proc. Natl Acad. Sci. USA*, **108**, 14186–14191.
13. Manzanares,M., Wada,H., Itasaki,N., Trainor,P.A., Krumlauf,R. and Holland,P.W. (2000) Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. *Nature*, **408**, 854–857.
14. Natale,A., Sims,C., Chiusano,M.L., Amoroso,A., D'Aniello,E., Fucci,L., Krumlauf,R., Branno,M. and Locascio,A. (2011) Evolution of anterior Hox regulatory elements among chordates. *BMC Evol. Biol.*, **11**, 330.
15. Holland,L.Z., Albalat,R., Azumi,K., Benito-Gutiérrez,E., Blow,M.J., Bronner-Fraser,M., Brunet,F., Butts,T., Candiani,S., Dishaw,L.J. *et al.* (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.*, **18**, 1100–1111.
16. Putnam,N.H., Butts,T., Ferrier,D.E.K., Furlong,R.F., Hellsten,U., Kawashima,T., Robinson-Rechavi,M., Shoguchi,E., Terry,A., Yu,J.-K. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.
17. Delsuc,F., Brinkmann,H., Chourrout,D. and Philippe,H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**, 965–968.
18. Sobral,D., Tassy,O. and Lemaire,P. (2009) Highly divergent gene expression programs can lead to similar chordate larval body plans. *Curr. Biol.*, **19**, 2014–2019.

19. Lemaire,P., Smith,W.C. and Nishida,H. (2008) Ascidians and the plasticity of the chordate developmental program. *Curr. Biol.*, **18**, R620–R631.
20. Britten,R.J. and Davidson,E.H. (1969) Gene regulation for higher cells: a theory. *Science*, **165**, 349–357.
21. Zuckerkandl,E. (1994) Molecular pathways to parallel evolution: I. Gene nexuses and their morphological correlates. *J. Mol. Evol.*, **39**, 661–678.
22. García-Bellido,A. (1996) Symmetries throughout organic evolution. *Proc. Natl Acad. Sci. USA*, **93**, 14229–14232.
23. Tsong,A.E., Miller,M.G., Raisner,R.M. and Johnson,A.D. (2003) Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell*, **115**, 389–399.
24. Ihmels,J., Bergmann,S., Gerami-Nejad,M., Yanai,I., McClellan,M., Berman,J. and Barkai,N. (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, **309**, 938–940.
25. Prud'homme,B., Gompel,N., Rokas,A., Kassner,V.A., Williams,T.M., Yeh,S.-D., True,J.R. and Carroll,S.B. (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, **440**, 1050–1053.
26. Cameron,R.A. and Davidson,E.H. (2009) Flexibility of transcription factor target site position in conserved cis-regulatory modules. *Dev. Biol.*, **336**, 122–135.
27. Oda-Ishii,I., Bertrand,V., Matsuo,I., Lemaire,P. and Saiga,H. (2005) Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians Halocynthia roretzi and Ciona intestinalis. *Development*, **132**, 1663–1674.
28. Lowe,C.B., Bejerano,G. and Haussler,D. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl Acad. Sci. USA*, **104**, 8005–8010.
29. Sanges,R., Kalmar,E., Claudiani,P., D'Amato,M., Muller,F. and Stupka,E. (2006) Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol.*, **7**, R56.
30. Chuzhanova,N.A., Krawczak,M., Nemytikova,L.A., Gusev,V.D. and Cooper,D.N. (2000) Promoter shuffling has occurred during the evolution of the vertebrate growth hormone gene. *Gene*, **254**, 9–18.
31. Ueda,M., Arimura,S., Yamamoto,M.P., Takaiwa,F., Tsutsumi,N. and Kadowaki,K. (2006) Promoter shuffling at a nuclear gene for mitochondrial RPL27. Involvement of interchromosome and subsequent intrachromosome recombinations. *Plant Physiol.*, **141**, 702–710.
32. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
33. Margulies,E.H., Cooper,G.M., Asimenos,G., Thomas,D.J., Dewey,C.N., Siepel,A., Birney,E., Keefe,D., Schwartz,A.S., Hou,M. *et al.* (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.*, **17**, 760–774.
34. Hubbard,T.J.P., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
35. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
36. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
37. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.
38. Mayor,C., Brudno,M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer,K.A., Pachter,L.S. and Dubchak,I. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046–1047.
39. Brudno,M., Chapman,M., Göttgens,B., Batzoglou,S. and Morgenstern,B. (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.
40. Tassy,O., Dauga,D., Daian,F., Sobral,D., Robin,F., Khoueiry,P., Salgado,D., Fox,V., Caillol,D., Schiappa,R. *et al.* (2010) The ANISEED database: digital representation, formalization, and elucidation of a chordate developmental program. *Genome Res.*, **20**, 1459–1468.
41. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
42. Götz,S., García-Gómez,J.M., Terol,J., Williams,T.D., Nagaraj,S.H., Nueda,M.J., Robles,M., Talón,M., Dopazo,J. and Conesa,A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.*, **36**, 3420–3435.
43. Roure,A., Rothbächer,U., Robin,F., Kalmar,E., Ferone,G., Lamy,C., Missero,C., Mueller,F. and Lemaire,P. (2007) A multicassette Gateway vector set for high throughput and comparative analyses in ciona and vertebrate embryos. *PLoS One*, **2**, e916.
44. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
45. Gehrig,J., Reischl,M., Kalmár,E., Ferg,M., Hadzhiev,Y., Zaucker,A., Song,C., Schindler,S., Liebel,U. and Müller,F. (2009) Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nat. Methods*, **6**, 911–916.
46. Kawakami,K. (2004) Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element. *Methods Cell Biol.*, **77**, 201–222.
47. Stephen,S., Pheasant,M., Makunin,I.V. and Mattick,J.S. (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol. Biol. Evol.*, **25**, 402–408.
48. Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A., Lewis,K.D. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
49. Kim,T.-K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K., Kuersten,S. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.
50. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
51. Alfano,C., Teresa Russo,M. and Spagnuolo,A. (2007) Developmental expression and transcriptional regulation of Ci-Pans, a novel neural marker gene of the ascidian, Ciona intestinalis. *Gene*, **406**, 36–41.
52. Karali,M., Peluso,I., Marigo,V. and Banfi,S. (2007) Identification and characterization of microRNAs expressed in the mouse eye. *Invest. Ophthalmol. Vis. Sci.*, **48**, 509–515.
53. Meyer,L.R., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Kuhn,R.M., Wong,M., Sloan,C.A., Rosenbloom,K.R., Roe,G., Rhead,B. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
54. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
55. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
56. Bryne,J.C., Valen,E., Tang,M.-H.E., Marstrand,T., Winther,O., Da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
57. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

58. Dennis,G., Sherman,B., Hosack,D., Yang,J., Gao,W., Lane,H. and Lempicki,R. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, R60.
59. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
60. Brudno,M., Malde,S., Poliakov,A., Do,C.B., Couronne,O., Dubchak,I. and Batzoglou,S. (2003) Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, **19**, i54–i62.
61. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*, **57**, 289–300.
62. Vavouri,T., McEwen,G.K., Woolfe,A., Gilks,W.R. and Elgar,G. (2006) Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. *Trends Genet.*, **22**, 5–10.
63. Kikuta,H., Laplante,M., Navratilova,P., Komisarczuk,A.Z., Engström,P.G., Fredman,D., Akalin,A., Caccamo,M., Sealy,I., Howe,K. *et al.* (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.*, **17**, 545–555.
64. Miura,Y., Tam,T., Ido,A., Morinaga,T., Miki,T., Hashimoto,T. and Tamaoki,T. (1995) Cloning and characterization of an ATBF1 isoform that expresses in a neuronal differentiation-dependent manner. *J. Biol. Chem.*, **270**, 26840–26848.
65. Jung,C.-G., Kim,H.-J., Kawaguchi,M., Khanna,K.K., Hida,H., Asai,K., Nishino,H. and Miura,Y. (2005) Homeotic factor ATBF1 induces the cell cycle arrest associated with neuronal differentiation. *Development*, **132**, 5137–5145.
66. Miwata,K., Chiba,T., Horii,R., Yamada,L., Kubo,A., Miyamura,D., Satoh,N. and Satou,Y. (2006) Systematic analysis of embryonic expression profiles of zinc finger genes in Ciona intestinalis. *Dev. Biol.*, **292**, 546–554.
67. Sandberg,M., Källström,M. and Muhr,J. (2005) Sox21 promotes the progression of vertebrate neurogenesis. *Nat. Neurosci.*, **8**, 995–1001.
68. Basch,M.L., Bronner-Fraser,M. and García-Castro,M.I. (2006) Specification of the neural crest occurs during gastrulation and requires Pax7. *Nature*, **441**, 218–222.
69. Corbo,J.C., Levine,M. and Zeller,R.W. (1997) Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. *Development*, **124**, 589–602.
70. Licastro,D., Gennarino,V.A., Petrera,F., Sanges,R., Banfi,S. and Stupka,E. (2010) Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. *BMC Genomics*, **11**, 151.
71. De Santa,F., Barozzi,I., Mietton,F., Ghisletti,S., Polletti,S., Tusi,B.K., Muller,H., Ragoussis,J., Wei,C.-L. and Natoli,G. (2010) A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol.*, **8**, e1000384.
72. Mattick,J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genet.*, **5**, e1000459.
73. Ulitsky,I., Shkumatava,A., Jan,C.H., Sive,H. and Bartel,D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, **147**, 1537–1550.
74. Pauli,A., Valen,E., Lin,M.F., Garber,M., Vastenhouw,N.L., Levin,J.Z., Fan,L., Sandelin,A., Rinn,J.L., Regev,A. *et al.* (2011) Systematic identification of long non-coding RNAs expressed during zebrafish embryogenesis. *Genome Res.*, **22**, 577–591.
75. Christensen,K.L., Patrick,A.N., McCoy,E.L. and Ford,H.L. (2008) The six family of homeobox genes in development and cancer. *Adv. Cancer Res.*, **101**, 93–126.
76. Aravind,L., Anantharaman,V., Balaji,S., Babu,M.M. and Iyer,L.M. (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.*, **29**, 231–262.
77. Bianchi,M.E. and Agresti,A. (2005) HMG proteins: dynamic players in gene regulation and differentiation. *Curr. Opin. Genet. Dev.*, **15**, 496–506.
78. Hannenhalli,S. and Kaestner,K.H. (2009) The evolution of Fox genes and their role in development and disease. *Nat. Rev. Genet.*, **10**, 233–240.
79. Denoeud,F., Henriet,S., Mungpakdee,S., Aury,J.-M., Da Silva,C., Brinkmann,H., Mikhaleva,J., Olsen,L.C., Jubin,C., Canestro,C. *et al.* (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, **330**, 1381–1385.
80. Hufton,A.L., Groth,D., Vingron,M., Lehrach,H., Poustka,A.J. and Panopoulou,G. (2008) Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res.*, **18**, 1582–1591.
81. Olinski,R.P., Lundin,L.-G. and Hallböök,F. (2006) Conserved synteny between the Ciona genome and human paralogons identifies large duplication events in the molecular evolution of the insulin-relaxin gene family. *Mol. Biol. Evol.*, **23**, 10–22.
82. Dong,X., Navratilova,P., Fredman,D., Drivenes,Ø., Becker,T.S. and Lenhard,B. (2010) Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. *Nucleic Acids Res.*, **38**, 1071–1085.
83. Hughes,A.L. and Friedman,R. (2005) Loss of ancestral genes in the genomic evolution of Ciona intestinalis. *Evol. Dev.*, **7**, 196–200.
84. Imai,K.S., Hino,K., Yagi,K., Satoh,N. and Satou,Y. (2004) Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. *Development*, **131**, 4047–4058.
85. Ikuta,T., Satoh,N. and Saiga,H. (2010) Limited functions of Hox genes in the larval development of the ascidian Ciona intestinalis. *Development*, **137**, 1505–1513.
86. Ferrier,D.E.K. and Holland,P.W.H. (2002) Ciona intestinalis ParaHox genes: evolution of Hox/ParaHox cluster integrity, developmental mode, and temporal colinearity. *Mol. Phylogenet. Evol.*, **24**, 412–417.
87. Ikuta,T., Yoshida,N., Satoh,N. and Saiga,H. (2004) Ciona intestinalis Hox gene cluster: its dispersed structure and residual colinear expression in development. *Proc. Natl Acad. Sci. USA*, **101**, 15118–15123.
88. Brosius,J. (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, **238**, 115–134.

DISCUSSION

Mon travail de thèse a permis de mettre en lumière un nouvel acteur de l'activité des enhancers : la séquence des spacers, situés entre les sites de fixation des facteurs de transcription. Nos résultats indiquent que ces séquences peuvent avoir un effet global sur la fixation des TFs, indépendant de l'affinité de la séquence de leur site de fixation telle que décrite par les données de SELEX-seq. Mais la fixation in vitro des TFs sur les différents enhancers ne corrèle pas toujours avec leur niveau d'activité, et comprendre le mode d'action des spacers est loin d'être trivial.

Dans cette dernière partie, je commenterai les questions suivantes, qui restent largement sans réponses et suggèrent un nouvel angle d'approche pour la compréhension des enhancers.

Comment les spacers modifient-ils le niveau d'activité des enhancers ?

Les spacers sont-ils des acteurs de la robustesse des enhancers ?

« Finalement, ça a l'air facile de faire des enhancers …»

Comment les spacers modifient-ils le niveau d'activité des enhancers ?

En regardant les séquences des spacers des enhancers randomisés testés pendant ma thèse, aucune base à aucune position ne paraît corréler avec leur niveau d'activité. Mais le petit nombre d'éléments testés ne permet ni de conclure, ni de faire des statistiques. L'analyse des spacers des plus de 150 000 éléments testés par Emma Farley n'a pas non plus permis d'identifier des séquences communes aux éléments actifs, allant dans le sens de mes observations (Farley et al., 2015). Ils ont conclu que toute l'information réside dans les sites de fixation des facteurs de transcription et n'ont donc pas considéré que la séquence des spacers puisse influencer le niveau d'activité d'un enhancer.

Ce n'est pas une surprise, la simple comparaison des séquences nucléotidiques des spacers des éléments randomisés est insuffisante pour comprendre leur activité. Mais l'ADN peut encoder de façon complexe et enchevêtrée une grande diversité des messages qui se lisent sur la base de triplets dans le cas des gènes codants par exemple, et des répétitions de dinucléotides (Yáñez-Cuna et al., 2014) ou des fréquences d'apparition de dinucléotides (Khoueiry et al., 2010) corrèlent avec la présence d'enhancers. La fixation des facteurs de transcription est influencée par la forme de l'hélice qui dépend elle-même de sa séquence, et de façon non triviale, similitudes de formes et de séquences (et vice versa) ne corrèlent pas nécessairement (Stephen C J Parker & Tullius, 2011).

Nous avons donc cherché à comprendre quelle est l'action des spacers et quel type d'information ils contiennent.

Les expériences de retard sur gel ont montré dans certains cas que la fixation d'Ets1/2 sur un fragment de l'enhancer contenant son site de fixation dépend de la taille du fragment considéré, et que plus la fenêtre est grande, plus sa fixation reflète sa fixation sur l'enhancer entier, suggérant un effet global des spacers sur la fixation. Dans d'autres cas, les 10pb centrées sur son site de fixation semblent suffisantes pour expliquer sa fixation.

Nous avons montré que la séquence de spacers peut influencer la fixation d'Ets1/2 sur l'enhancer, alors que l'affinité prédite pour son site de fixation semble inchangée. De même, dans les données d'Emma Farley, des enhancers synthétiques ayant des sites de fixation prédits pour être au moins aussi bons que ceux de l'élément a, présentent des niveaux d'activités très

variés (plus d'un 1/7 sont inactifs et la même proportion seulement est plus active que le WT (<0,01RPM et >4RPM respectivement)).

Cette apparente indépendance entre l'action des séquences des TFBS et des spacers sur le niveau d'activité des enhancers peut être testée : il suffit de prendre une petite dizaine d'éléments randomisés ayant des niveaux d'activités variés, et de créer deux versions de chacun d'entre eux, où un des sites ETS porte une mutation sensée accroitre et diminuer son affinité respectivement. Une totale indépendance serait mise en évidence par une augmentation (resp. diminution) linéaire du niveau d'activité des variants « améliorés » (resp. affaiblis). Néanmoins, cette hypothèse est peu probable et chaque séquence randomisée pourrait agir différemment.

Les spacers peuvent agir sur la fixation des facteurs de transcription à différents niveaux (Figure III.1). Nos résultats de retards sur gel suggèrent une possible implication de la forme de l'hélice (Figure III.1 C).

Nous avons débuté une collaboration avec une équipe de l'IRB à Barcelone pour modéliser les propriétés physiques de l'hélice d'ADN pour chacune des séquences testées. Les premiers résultats, obtenus sur les séquences de 6 éléments-a randomisés et du sauvage, n'ont pas permis d'associer certaines propriétés de l'hélice d'ADN au niveau d'activité de leur séquence et nous attendons les résultats sur l'ensemble des séquences des éléments testés avec impatience.

La création de nouveaux sites de fixation pour des facteurs de transcription me semble être un phénomène anecdotique et ne permet pas d'expliquer le niveau d'activité des éléments randomisés. Toutefois, il est à prendre en compte dans l'interprétation des résultats, dont ils complexifient l'interprétation. Par exemple, la forte activité de l'élément aR43 est due à la fixation d'Ets1/2 sur un nouveau site crée dans les spacers, alors que les sites E1 et E2 ne sont pas fixés. Les séquences randomisées empêcheraient donc la fixation d'Ets1/2 sur les deux sites originaux, mais pas sur ce nouveau site ? La création de ce nouveau motif fonctionnel permet donc d'expliquer l'activité de l'enhancer, mais n'apporte aucun indice pour comprendre comment les spacers jouent sur la fixation des facteurs de transcription.

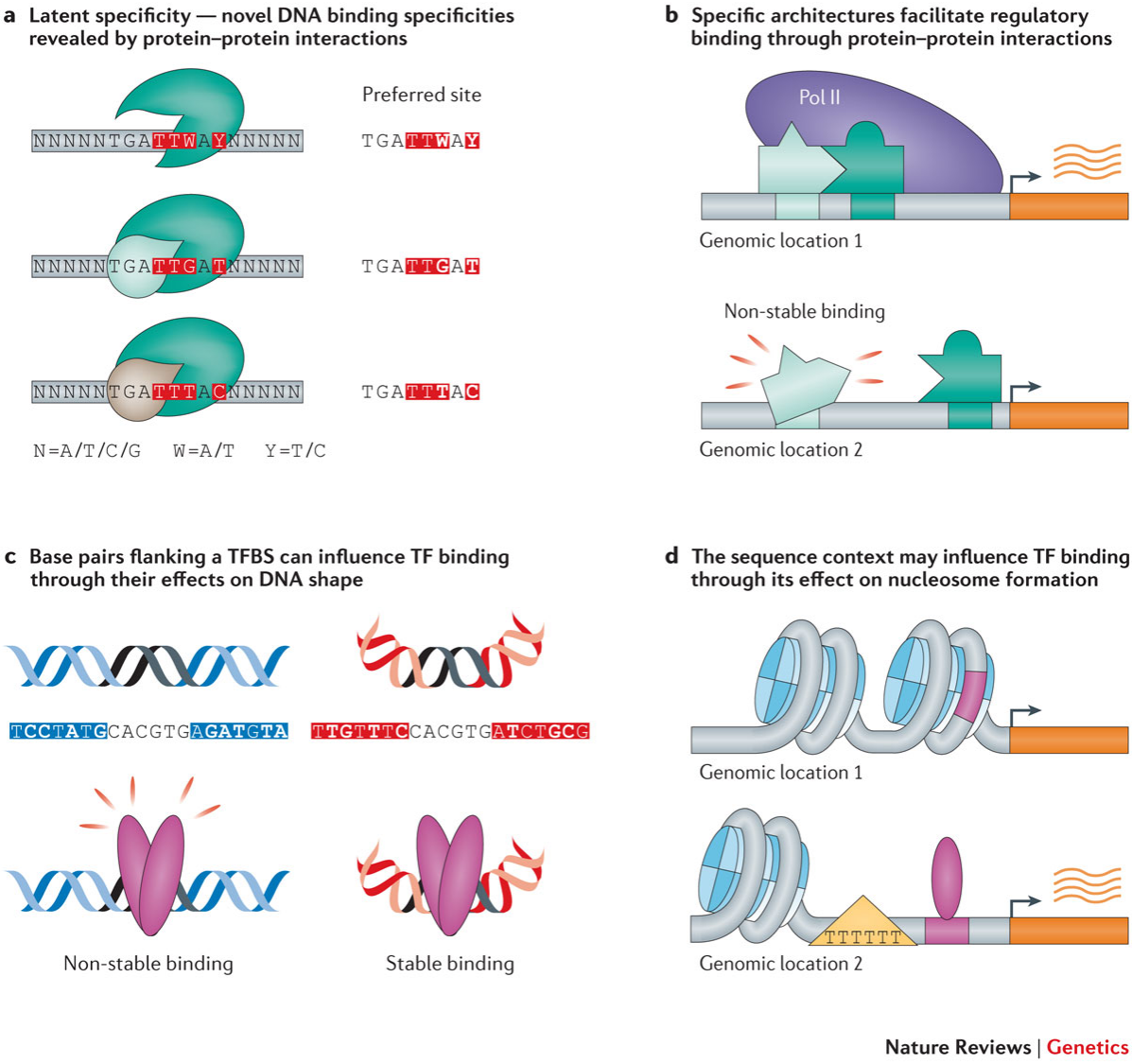


Figure III.1 Les séquences intercalantes sont impliquées dans différents mécanismes affectant la fixation des facteurs de transcription. (Levo & Ségal 2014)

Les spacers sont-ils des acteurs de la robustesse des enhancers ?

Le comptage des embryons électroporés donnait l'impression que, d'une expérience à l'autre, l'activité des éléments randomisés est plus variable que celle de l'élément naturel, en termes de niveau et de profil spatial d'activité.

Si le pourcentage d'embryons où LacZ est exprimé reflète de façon quantitative le niveau d'activité des enhancers (C. D. Brown et al., 2007), je n'ai pas réussi à quantifier cette variabilité. La normalisation des résultats est efficace, mais limitée pour les éléments très actifs ou très faibles, à cause des effets de seuil. L'utilisation d'un contrôle interne, linéaire, pourrait permettre d'obtenir des mesures du niveau d'activité plus faciles à comparer.

Plusieurs études ont montré que la sous-optimisation d'une entité fonctionnelle est une stratégie utilisée par l'évolution pour gagner en spécificité et/ou en robustesse (par exemple Crocker et al., 2015; Farley et al., 2015). Une sous-optimisation des spacers dans les enhancers naturels, au sens où ils ne confèrent pas un niveau d'activité maximal à l'enhancer, pourrait donc être impliquée dans la spécificité de son profil spatial d'activité. Cette hypothèse permettrait d'expliquer pourquoi l'élément-a n'est pas capable de répondre à de faibles niveaux de signalisation FGF, contrairement à certains éléments randomisés actifs dans un plus grand nombre de cellules.

« Finalement, ça a l'air facile de faire des enhancers …»

Cette phrase résume notre surprise au vu des résultats des expériences de randomisation des spacers.

La difficulté à « créer » un enhancer synthétique, provient peut-être plus de limitations techniques que théoriques : ce type d'expérience n'a jamais été mené chez les principaux organismes modèles. *Ciona intestinalis* est une espèce où la transgénèse est facile et rapide, et l'élément a est un petit enhancer. Ce n'est sans doute pas anodin si cette approche a été réalisée au même moment et indépendamment par deux équipes.

Mais la difficulté à trouver dans le génome de *Ciona* des regroupements de sites de fixation pour Ets et GATA ayant une activité enhancer laissait également penser qu'il fallait un contexte très particulier dans les spacers pour que l'enhancer fonctionne. Au contraire, au vu de mes résultats, il semblerait que ce soit l'identité « non-enhancer » qui est codée par certaines séquences des spacers. Et il ne semble même pas nécessaire d'essayer d'homogénéiser la composition en base d'un enhancer synthétique avec celle du génome cible : les éléments

randomisés ont un GC moyen de 50%, alors que le génome de *Ciona* est riche en AT. Ce qui pose de nombreuses questions et force à considérer la « question enhancer » sous un autre angle, moins intuitif.

Se fourvoie-t-on en cherchant uniquement des signatures des enhancers, faut-il également s'intéresser à ce qui empêche un regroupement de TFBS d'agir en tant qu'enhancer ? Doit-on procéder par élimination des séquences inactives plutôt que par sélection de celles qui cumulent le maximum de signatures cis-régulatrices lorsque qu'on cherche à identifier des enhancers à grande échelle ? Comment combiner ces deux logiques ?

L'inactivité de certains de ces regroupements de TFBS génomiques inactifs est-elle sous pression de sélection, afin d'empêcher l'apparition d'enhancers inopportuns ? Cette stratégie, s'il en est, semble peu logique et très coûteuse : pourquoi maintenir un très grand nombre de regroupements de sites Ets et GATA inactifs, alors qu'il suffirait de garder l'activité d'une minorité sous pressions de sélection. Mais de nombreux exemples ont déjà montré que l'évolution, ni élégante ni logique, relève du bricolage (Jacob 1977 Orgogozo review). Ainsi il ne serait pas totalement absurde que les séquences des spacers des enhancers fonctionnels soient peu contraintes, ce qui permettrait une plus grande dynamique d'apparition de nouveaux enhancers (réservoirs de séquences régulatrices), une plus grande robustesse des enhancers aux mutations (moins « fragiles », donc moindre « coût » de pressions de sélection sur eux), contrebalancée par des pressions de sélections maintenant les autres regroupements de TFBS inactifs.

D'autre part, nous avons montré que deux regroupements génomiques de sites de fixation pour Ets et pour GATA, N83 et N26, ont une activité enhancer, mais nous n'avons pas pu lier cette activité à celle d'un gène voisin, exprimé dans les lignages neuraux précoces. Les enhancers identifiés par Khoueiry et ses collègues étaient en effet tous les 4 à proximité de marqueurs neuraux précoces, Nodal, Elk, Prickle et Erf-a (Khoueiry et al., 2010). La principale différence entre ces enhancers neuraux « avérés » et N26/N83 est que les premiers sont situés dans des régions conservées avec *Ciona savignyi.*

N83 est situé dans une région très conservée entre les génomes de ciones type A et B. Le gène le plus proche est à plus de 13kb, et les suivants à 30kb minimum. Aucune donnée d'expression n'est disponible pour ces gènes, et seul le gène le plus proche,

(KH2012:KH.L128.7) a une annotation de domaine protéique (interpro) : “C-type lectin fold” et “C-type lectin like”. N26 est situé dans une région partiellement conservée entre type A et type B (0 à 95% de conservation). Le 5UTR’ du gène le plus proche est situé à 1kb : KH2012:KH.C10.118 ne semble pas exprimé au stade 32 cellules. De l’autre côté, à 2,2kb, se situe la région 3’ du gène KH2012:KH.C10.161, pour lequel nous ne disposons d’aucune donnée d’expression.

Une explication séduisante serait de penser qu’il s’agit de shadow enhancers, les premiers enhancers agissant à longue distance identifiés chez les ascidies ! Des expériences de capture de la conformation de la chromatine permettraient de tester cette hypothèse. Mais des données d’ARN-seq et d’ATAC-seq obtenues par l’équipe (Jacques Piette, Marta Magri et Alicia Madgwick, non publié) semblent montrer que ces deux éléments ne sont pas des enhancers neuraux précoces dans leur contexte génomique : en effet, ils ne semblent pas transcrits (contrairement à la majorité des enhancers de *Ciona* identifiés), et sont localisés dans de la chromatine inactive.

Ces résultats pourraient suggérer qu’il y a deux poids deux mesures dans la gestion des regroupements de TFBS dans un génome au cours de l’évolution :
a) Les regroupements localisés dans de la chromatine inactive seraient systématiquement inactivés par cette dernière et donc moins soumis à des pressions de sélection sur l’activité enhancer. Ce qui expliquerait qu’ils puissent, ou non, sortis de leur contexte chromatinien activer la transcription d’un gène rapporteur. Toutefois, GATA étant un facteur pionnier, capable d’ouvrir la chromatine, il est nécessaire de bien définir le niveau de compaction minimum de l’ADN garantissant l’inactivation de l’enhancer.
b) Les regroupements localisés dans de la chromatine ouverte seraient sous fortes pression de sélections pour le maintien de leur activité ou inactivité, comme nous l’avons proposé plus haut.

Il est probable que ces deux logiques de limitation du nombre d’enhancers actifs dans le génome coexistent, associées à des pressions de sélections dans les enhancers actifs agissant sur l’identité, l’organisation, l’affinité de leurs TF un nouveau site BSs, la forme de l’hélice d’ADN, déjà identifiés comme cibles de l’évolution pour le maintient de l’identité enhancer (Crocker et al., 2015; Stephen C J Parker & Tullius, 2011).

Il me semble donc fondamental de considérer le rôle des spacers d'un point de vue évolutif. Sans ça, nous ne saurons pas si la compréhension mécanistique, même poussée, de leur mode d'action reflète leur logique in vivo ou en est une interprétation élégante (comme dirait Dobianszki).

Pendant ma thèse, j'ai voulu éclairer mon approche en aveugle d'une lumière évolutive, en tirant parti de l'important polymorphisme des génomes de ciones. J'ai donc collecté des échantillons de sperme de *Ciona intestinalis* de types A et B venant de différents horizons (Japon, Nouvelle Zélande, Naples, Angleterre, Roscoff, Californie, Quebec) afin de créer une banque d'ADN génomique de diverses populations de ciones. La stratégie était simple : séquencer pour chaque individu les enhancers les mieux caractérisés chez la cione : ceux de Bra, Elk, Otx, FoxF, SFRP1/5 et les analyser par des approches de microévolution pour détecter quelles pressions de sélections agissent sur quels éléments des enhancers (test McDonald Kreitman, ref droso). Des résultats préliminaires très prometteurs sur un petit nombre de séquences ont été obtenus par Andrès Garcia de la Filia Molina et montrent que ces enhancers sont sous pression de sélection (négatives dans la plupart des cas), et que les séquences des sites de fixation des facteurs de transcription ne sont pas plus contraintes que celles des spacers. Nous n'avons pas pu poursuivre ce projet, mais je suis convaincue que cette approche serait complémentaire et informative.

Elle permettrait notamment de répondre à une des questions existentielles de ma thèse : Est-ce pertinent de travailler avec un enhancer minimal ? Quel est le sens biologique de mes résultats ?
En effet, pour pouvoir tester l'activité et analyser des enhancers, il est nécessaire d'en définir arbitrairement les limites. Si un enhancer minimal est bien défini au niveau fonctionnel, il ne l'est pas au sens biologique. Par exemple, l'enhancer neural d'O*tx* fait plus de 55 paires de bases et contient donc plus de sites ETS et GATA que l'élément a. Il est évident que mes expériences de randomisation étaient limitées techniquement par la taille maximale des oligonucléotides pouvant être synthétisés, mais on ne sait pas comment les variants se seraient comportés si j'avais inclus des régions voisines. L'analyse de microévolution pourrait permettre d'identifier les fenêtres les plus opportunes à considérer pour étudier un enhancer, voire des phénomènes d'interdépendance entre différentes positions, de compensation de certaines mutations a priori délétères qui ajoutent à la complexité de l'information codée par les enhancers

Perspectives

Je ne pense pas vraiment qu'un « code cis-régulateur » universel puisse être un jour découvert, mais l'information cis-régulatrice étant portée intrinsèquement par la séquence d'ADN, comprendre comment elle encode ses différents niveaux d'information me semble un enjeu fondamental pour l'identification d'enhancers à l'échelle génomique.

L'amélioration, l'automatisation des techniques, le prix décroissant du séquençage haut débit permettent de multiplier les approches de ChIP-seq, ATAC-seq, etc, mais leur pouvoir d'identification est limité à des corrélations et aux types cellulaires, stades de développement et conditions environnementales étudiés, et peuvent être difficiles à mettre en œuvre dans certaines espèces.

Ciona intestinalis est un organisme modèle qui a un « petit » génome, où il est facile de tester l'activité d'un grand nombre d'enhancers, et dont le niveau de polymorphisme permet des approches de micro-évolution même dans les séquences non-codantes. C'est le modèle « parfait » pour essayer de comprendre d'un point de vue évolutif les règles définissant l'identité enhancer, et de les tester in vivo.

BIBLIOGRAPHIE

Abdul-Wajid, S., Veeman, M. T., Chiba, S., Turner, T. L., & Smith, W. C. (2014). Exploiting the extraordinary genetic polymorphism of ciona for developmental genetics with whole genome sequencing. *Genetics*, *197*(1), 49–59. doi:10.1534/genetics.114.161778

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., … Mann, R. S. (2015). Deconvolving the recognition of DNA shape from sequence. *Cell*, *161*(2), 307–18. doi:10.1016/j.cell.2015.02.008

Abitua, P. B., Wagner, E., Navarrete, I. a, & Levine, M. (2012). Identification of a rudimentary neural crest in a non-vertebrate chordate. *Nature*, *492*(7427), 104–7. doi:10.1038/nature11589

Akam, M. (1989). Drosophila development: making stripes inelegantly. *Nature*, *341*(6240), 282–283. doi:10.1038/341282a0

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Developmental Cell*, *16*(1), 47–57. doi:10.1016/j.devcel.2008.11.011

Andersson, R., Sandelin, A., & Danko, C. G. (2015). A unified architecture of transcriptional regulatory elements. *Trends in Genetics*, *31*(8), 426–433. doi:10.1016/j.tig.2015.05.007

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., … Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science (New York, N.Y.)*, *340*(6137), 1234167. doi:10.1126/science.1234167

Aristote. Hist. An. Lib IV, capVI

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.)*, *339*(6123), 1074–7. doi:10.1126/science.1232542

Arnone, M. I., & Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development*, *124*(10), 1851–1864.

Arnosti, D. N., & Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, *94*(5), 890–898. doi:10.1002/jcb.20352

Atanesyan, L., Günther, V., Dichtl, B., Georgiev, O., & Schaffner, W. (2012). Polyglutamine tracts as modulators of transcriptional activation from yeast to mammals. *Biological*

Chemistry, *393*(1-2), 63–70. doi:10.1515/BC-2011-252

Auger, H., Lamy, C., Haeussler, M., Khoueiry, P., Lemaire, P., & Joly, J.-S. (2009). Similar regulatory logic in Ciona intestinalis for two Wnt pathway modulators, ROR and SFRP-1/5. *Developmental Biology*, *329*(2), 364–73. doi:10.1016/j.ydbio.2009.02.018

Aza-Blanc, P., Ramírez-Weber, F. A., Laget, M. P., Schwartz, C., & Kornberg, T. B. (1997). Proteolysis that is inhibited by hedgehog targets Cubitus interruptus protein to the nucleus and converts it to a repressor. *Cell*, *89*(7), 1043–53.

Banerjee, A. R., Kim, Y. J., & Kim, T. H. (2014). A novel virus-inducible enhancer of the interferon-β gene with tightly linked promoter and enhancer activities. *Nucleic Acids Research*, *42*(20), 12537–54. doi:10.1093/nar/gku1018

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2 Pt 1), 299–308. doi:10.1016/0092-8674(81)90413-X

Barolo, S. (2012). Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *34*(2), 135–41. doi:10.1002/bies.201100121

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., & Natoli, G. (2014). Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Molecular Cell*, *54*(5), 844–57. doi:10.1016/j.molcel.2014.04.006

Barthel, K. K. B., & Liu, X. (2008). A transcriptional enhancer from the coding region of ADAMTS5. *PloS One*, *3*(5), e2184. doi:10.1371/journal.pone.0002184

Bassett, A., Cooper, S., Wu, C., & Travers, A. (2009). The folding and unfolding of eukaryotic chromatin. *Current Opinion in Genetics & Development*, *19*(2), 159–65. doi:10.1016/j.gde.2009.02.010

Beh, J., Shi, W., Levine, M., Davidson, B., & Christiaen, L. (2007). FoxF is essential for FGF-induced migration of heart progenitor cells in the ascidian Ciona intestinalis. *Development (Cambridge, England)*, *134*(18), 3297–305. doi:10.1242/dev.010140

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science (New York, N.Y.)*, *304*(5675), 1321–5. doi:10.1126/science.1098119

Benoist, C., & Chambon, P. (1981). In vivo sequence requirements of the SV40 early promotor region. *Nature*, *290*(5804), 304–10. doi:10.1038/290304a0

Berger, M. F., & Bulyk, M. L. (2006). Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods in Molecular Biology (Clifton, N.J.)*, *338*, 245–60. doi:10.1385/1-59745-097-

9:245

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

Bertrand, V., Hudson, C., Caillol, D., Popovici, C., & Lemaire, P. (2003). Neural tissue in ascidian embryos is induced by FGF9/16/20, acting via a combination of maternal GATA and Ets transcription factors. *Cell*, *115*(5), 615–27.

Bishop, E. P., Rohs, R., Parker, S. C. J., West, S. M., Liu, P., Mann, R. S., … Tullius, T. D. (2011). A Map of Minor Groove Shape and Electrostatic Potential from Hydroxyl Radical Cleavage Patterns of DNA. *ACS Chem Biol*, *6*(12), 1314–1320. doi:10.1021/cb200155t.A

Blackwood, E. M. (1998). Going the Distance: A Current View of Enhancer Action. *Science*, *281*(5373), 60–63. doi:10.1126/science.281.5373.60

Blackwood, E. M., & Kadonaga, J. T. (1998). Going the distance: a current view of enhancer action. *Science (New York, N.Y.)*, *281*(5373), 60–3.

Borok, M. J., Tran, D. A., Ho, M. C. W., & Drewell, R. A. (2010). Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. *Development (Cambridge, England)*, *137*(1), 5–13. doi:10.1242/dev.036160

Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R., & Nasmyth, K. (1985). Characterization of a “silencer” in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell*, *41*(1), 41–48.

Britten, R. J., & Davidson, E. H. (1969). Gene Regulation for Higher Cells: A Theory. *Science*, *165*(3891), 349–357. doi:10.1126/science.165.3891.349

Broos, S., Soete, A., Hooghe, B., Moran, R., van Roy, F., & De Bleser, P. (2013). PhysBinder: Improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties. *Nucleic Acids Research*, *41*(Web Server issue), W531–4. doi:10.1093/nar/gkt288

Brown, C. D., Johnson, D. S., & Sidow, A. (2007). Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science (New York, N.Y.)*, *317*(5844), 1557–60. doi:10.1126/science.1145893

Brown, J. B., & Celniker, S. E. (2015). Lessons from modENCODE. *Annual Review of Genomics and Human Genetics*, *16*, 31–53. doi:10.1146/annurev-genom-090413-025448

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–8. doi:10.1038/nmeth.2688

Bulger, M., & Groudine, M. (2011). Functional and mechanistic diversity of distal transcription

enhancers. *Cell*, *144*(3), 327–39. doi:10.1016/j.cell.2011.01.024

Butler, J. E., & Kadonaga, J. T. (2001). Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes & Development*, *15*(19), 2515–9. doi:10.1101/gad.924301

Calhoun, V. C., Stathopoulos, A., & Levine, M. (2002). Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9243–7. doi:10.1073/pnas.142291299

Campos, E. I., & Reinberg, D. (2009). Histones: annotating chromatin. *Annual Review of Genetics*, *43*, 559–99. doi:10.1146/annurev.genet.032608.103928

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., … Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, *38*(6), 626–635. doi:10.1038/ng1789

Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, *134*(1), 25–36. doi:10.1016/j.cell.2008.06.030

Chabry, L. (1887). *Embryologie normale et tératologique des Ascidies*. Felix Alcan Editeur, Paris.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., … Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (New York, N.Y.)*, *327*(5963), 302–5. doi:10.1126/science.1182213

Chen, J., Zhang, Z., Li, L., Chen, B.-C., Revyakin, A., Hajj, B., … Liu, Z. (2014). Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell*, *156*(6), 1274–85. doi:10.1016/j.cell.2014.01.062

Chinnadurai, G. (2007). Transcriptional regulation by C-terminal binding proteins. *The International Journal of Biochemistry & Cell Biology*, *39*(9), 1593–607. doi:10.1016/j.biocel.2007.01.025

Christiaen, L., Bourrat, F., & Joly, J.-S. (2005). A modular cis-regulatory system controls isoform-specific pitx expression in ascidian stomodaeum. *Developmental Biology*, *277*(2), 557–66. doi:10.1016/j.ydbio.2004.10.008

Christiaen, L., Davidson, B., Kawashima, T., Powell, W., Nolla, H., Vranizan, K., & Levine, M. (2008). The transcription/migration interface in heart precursors of Ciona intestinalis. *Science (New York, N.Y.)*, *320*(5881), 1349–52. doi:10.1126/science.1158170

Christiaen, L., Stolfi, A., Davidson, B., & Levine, M. (2009). Spatio-temporal intersection of Lhx3 and Tbx6 defines the cardiac field through synergistic activation of Mesp. *Developmental Biology*, *328*(2), 552–60. doi:10.1016/j.ydbio.2009.01.033

Christiaen, L., Wagner, E., Shi, W., & Levine, M. (2009a). Electroporation of transgenic DNAs in the sea squirt Ciona. *Cold Spring Harbor Protocols*, *2009*(12), pdb.prot5345. doi:10.1101/pdb.prot5345

Christiaen, L., Wagner, E., Shi, W., & Levine, M. (2009b). Microinjection of morpholino oligos and RNAs in sea squirt (Ciona) embryos. *Cold Spring Harbor Protocols*, *2009*(12), pdb.prot5347. doi:10.1101/pdb.prot5347

Conaway, R. C., & Conaway, J. W. (2011). Origins and activity of the Mediator complex. *Seminars in Cell & Developmental Biology*, *22*(7), 729–734. doi:10.1016/j.semcdb.2011.07.021

Conklin, E. G. (1905). The organization and cell lineage of the ascidian egg. *J. Acad. Nat. Sci. Phila.*, *13*, 1–119.

Corbo, J. C., Levine, M., & Zeller, R. W. (1997). Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. *Development (Cambridge, England)*, *124*(3), 589–602.

Courey, A. J., & Jia, S. (2001). Transcriptional repression: the long and the short of it. *Genes & Development*, *15*(21), 2786–96. doi:10.1101/gad.939601

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563. doi:10.1038/227561a0

Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., … Stern, D. L. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, *160*(1-2), 191–203. doi:10.1016/j.cell.2014.11.041

Cunliffe, V. T. (2008). Eloquent silence: developmental functions of Class I histone deacetylases. *Current Opinion in Genetics & Development*, *18*(5), 404–10. doi:10.1016/j.gde.2008.10.001

Cuvier, G. (1817). *Mémoires pour servir à l'histoire et à l'anatomie des mollusques*. Paris: Deterville.

Davidson, E. H., & Erwin, D. H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science (New York, N.Y.)*, *311*, 796–800. doi:10.1126/science.1126454

Deaton, A., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, *25*(10), 1010–1022. doi:10.1101/gad.2037511.1010

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., … Rokhsar, D. S. (2002). The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science (New York, N.Y.)*, *298*(5601), 2157–67. doi:10.1126/science.1080049

Delsuc, F., Brinkmann, H., Chourrout, D., & Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, *439*(7079), 965–8. doi:10.1038/nature04336

Delsuc, F., Tsagkogeorga, G., Lartillot, N., & Philippe, H. (2008). Additional molecular support for the new chordate phylogeny. *Genesis (New York, N.Y. : 2000)*, *46*(11), 592–604. doi:10.1002/dvg.20450

DeMare, L. E., Leng, J., Cotney, J., Reilly, S. K., Yin, J., Sarro, R., & Noonan, J. P. (2013). The genomic landscape of cohesin-associated chromatin interactions. *Genome Research*, *23*(8), 1224–34. doi:10.1101/gr.156570.113

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., … Blobel, G. A. (2012). Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. *Cell*, *149*(6), 1233–1244. doi:10.1016/j.cell.2012.03.051

Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.-M., Da Silva, C., Brinkmann, H., … Chourrout, D. (2010). Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science (New York, N.Y.)*, *330*(6009), 1381–5. doi:10.1126/science.1194167

Dhar, R., Weissman, S., Zain, B., Pan, J., & Lewis, A. J. (1974). The nucleotide sequence preceding an RNA polymerase initiation site on SV40 DNA. Part 2. The sequence of the early strand transcript. *Nucleic Acids Research*, (4), 595–611.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–80. doi:10.1038/nature11082

Dror, I., Golan, T., Levy, C., Rohs, R., & Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment on transcription factor binding across diverse protein families. *Genome Research*, *25*(9), 1268–80. doi:10.1101/gr.184671.114

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a., Doyle, F., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

Durán, E., Djebali, S., González, S., Flores, O., Mercader, J. M., Guigó, R., … Orozco, M. (2013). Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. *Nucleic Acids Research*, *41*(15), 7220–30. doi:10.1093/nar/gkt511

Ea, V., Sexton, T., Gostan, T., Herviou, L., Baudement, M.-O., Zhang, Y., … Forné, T. (2015). Distinct polymer physics principles govern chromatin dynamics in mouse and Drosophila topological domains. *BMC Genomics*, *16*, 607. doi:10.1186/s12864-015-1786-8

Ellington, A. D., & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature*, *346*(6287), 818–22. doi:10.1038/346818a0

Emera, D., Casola, C., Lynch, V. J., Wildman, D. E., Agnew, D., & Wagner, G. P. (2012). Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. *Molecular Biology and Evolution*, *29*(1), 239–47. doi:10.1093/molbev/msr189

Emera, D., & Wagner, G. P. (2012). Transposable element recruitments in the mammalian placenta: impacts and mechanisms. *Briefings in Functional Genomics*, *11*(4), 267–76. doi:10.1093/bfgp/els013

Erceg, J., Saunders, T. E., Girardot, C., Devos, D. P., Hufnagel, L., & Furlong, E. E. M. (2014). Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity. *PLoS Genetics*, *10*(1), e1004060. doi:10.1371/journal.pgen.1004060

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., & Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science (New York, N.Y.)*, *350*(6258), 325–328. doi:10.1126/science.aac6948

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., … Segal, E. (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Computational Biology*, *4*(11), e1000216. doi:10.1371/journal.pcbi.1000216

Fisher, A. L., Ohsako, S., & Caudy, M. (1996). The WRPW motif of the hairy-related basic helix-loop-helix repressor proteins acts as a 4-amino-acid transcription repression and protein-protein interaction domain. *Molecular and Cellular Biology*, *16*(6), 2670–7.

Fisher, W. W., Li, J. J., Hammonds, A. S., Brown, J. B., Pfeiffer, B. D., Weiszmann, R., … Celniker, S. E. (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(52), 21330–5. doi:10.1073/pnas.1209589110

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Lassmann, T., … Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–70. doi:10.1038/nature13182

Francetic, T., Le May, M., Hamed, M., Mach, H., Meyers, D., Cole, P. A., … Li, Q. (2012). Regulation of Myf5 Early Enhancer by Histone Acetyltransferase p300 during Stem Cell Differentiation. *Molecular Biology*, *1*. doi:10.4172/2168-9547.1000103

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., & Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, *466*(7305), 490–3. doi:10.1038/nature09158

Fried, M., & Crothers, D. M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Research*, *9*(23), 6505–25.

Fujioka, M., Emi-Sarker, Y., Yusibova, G. L., Goto, T., & Jaynes, J. B. (1999). Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development (Cambridge, England)*, *126*(11), 2527–38.

Gallant, J. R., Imhoff, V. E., Martin, A., Savage, W. K., Chamberlain, N. L., Pote, B. L., … Mullen, S. P. (2014). Ancient homology underlies adaptive mimetic diversity across butterflies. *Nature Communications*, *5*, 4817. doi:10.1038/ncomms5817

Gannon, F., O'Hare, K., Perrin, F., LePennec, J.P. Benoist, C., Cochet, M., Breathnach, R., … Chambon, P. (1979). Gene, Organisation and sequences at the 5' end of a cloned complete ovalbumin. *Saudi Med J*, (278), 428–434. doi:10.1073/pnas.0703993104

Garner, M. M., & Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Research*, *9*(13), 3047–60.

Garton, M., Najafabadi, H. S., Schmitges, F. W., Radovani, E., Hughes, T. R., & Kim, P. M. (2015). A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. *Nucleic Acids Research*, *43*(19), 9147–57. doi:10.1093/nar/gkv919

Geggier, S., & Vologodskii, A. (2010). Sequence dependence of DNA bending rigidity. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(35), 15421–6. doi:10.1073/pnas.1004809107

Gehrig, J., Reischl, M., Kalmár, E., Ferg, M., Hadzhiev, Y., Zaucker, A., … Müller, F. (2009). Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nature Methods*, *6*(12), 911–916. doi:10.1038/nmeth.1396

Gibcus, J. H., & Dekker, J. (2013). The Hierarchy of the 3D Genome. *Molecular Cell*, *49*(5), 773–782. doi:10.1016/j.molcel.2013.02.011

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, *17*(6), 877–885. doi:10.1101/gr.5533506

Gisselbrecht, S. S., Barrera, L. A., Porsch, M., Aboukhalil, A., Estep, P. W., Vedenko, A., … Bulyk, M. L. (2013). Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. *Nature Methods*, *10*(8), 774–80. doi:10.1038/nmeth.2558

Gohl, D., Müller, M., Pirrotta, V., Affolter, M., & Schedl, P. (2008). Enhancer blocking and transvection at the Drosophila apterous locus. *Genetics*, *178*(1), 127–43. doi:10.1534/genetics.107.077768

Gómez-Marín, C., Tena, J. J., Acemel, R. D., López-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., … Gómez-Skarmeta, J. L. (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(24), 7542–7. doi:10.1073/pnas.1505463112

Gray, S., & Levine, M. (1996). Transcriptional repression in development. *Current Opinion in*

Cell Biology, *8*(3), 358–364.

Haberle, V., Li, N., Hadzhiev, Y., Plessy, C., Previti, C., Nepal, C., … Lenhard, B. (2014). Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, *507*(7492), 381–385. doi:10.1038/nature12974

Haeckel, E. (1904). Kunstformen der Natur

Haeussler, M., Jaszczyszyn, Y., Christiaen, L., & Joly, J.-S. (2010). A cis-regulatory signature for chordate anterior neuroectodermal genes. *PLoS Genetics*, *6*(4), e1000912. doi:10.1371/journal.pgen.1000912

Harafuji, N., Keys, D. N., & Levine, M. (2002). Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(10), 6802–5. doi:10.1073/pnas.052024999

Hardison, R. C., & Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews. Genetics*, *13*(7), 469–83. doi:10.1038/nrg3242

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R., & Eisen, M. B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genetics*, *4*(6), e1000106. doi:10.1371/journal.pgen.1000106

Harmston, N., Baresic, A., Lenhard, B., & B, P. T. R. S. (2013). The mystery of extreme non-coding conservation The mystery of extreme non-coding conservation, (November).

He, Q., Bardet, A. F., Patton, B., Purvis, J., Johnston, J., Paulson, A., … Zeitlinger, J. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nature Genetics*, *43*(5), 414–20. doi:10.1038/ng.808

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., … Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–8. doi:10.1038/ng1966

Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols*, *2*(8), 1849–61. doi:10.1038/nprot.2007.249

Hergeth, S. P., Schneider, R., Hergeth, S. P., & Schneider, R. (2015). The H 1 linker histones : multifunctional proteins beyond the nucleosomal core particle, *16*(11), 1–15. doi:10.15252/embr.201540749

Hikosaka Akira, Takehiro Kusakabe, N. S. (1994). Hikosaka1994.pdf.

Hill, M. M., Broman, K. W., Stupka, E., Smith, W. C., Jiang, D., & Sidow, A. (2008). The C. savignyi genetic map and its integration with the reference sequence facilitates insights

into chordate genome evolution. *Genome Research*, *18*(8), 1369–79. doi:10.1101/gr.078576.108

Hirth, F. (2003). An urbilaterian origin of the tripartite brain: developmental genetic insights from Drosophila. *Development*, *130*(11), 2365–2373. doi:10.1242/dev.00438

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., … Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, *155*(4), 934–47. doi:10.1016/j.cell.2013.09.053

Hoekstra, H. E., & Coyne, J. A. (2007). The locus of evolution: evo devo and the genetics of adaptation. *Evolution; International Journal of Organic Evolution*, *61*(5), 995–1016. doi:10.1111/j.1558-5646.2007.00105.x

Holmqvist, P.-H., & Mannervik, M. (2013). Genomic occupancy of the transcriptional co-activators p300 and CBP. *Transcription*, *4*(1), 18–23. doi:10.4161/trns.22601

Hoshino, Z. 'ichiro, & Tokioka, T. (1967). An unusually robust Ciona from the northeastern coast of Honsyu island, Japan. *PUBLICATIONS OF THE SETO MARINE BIOLOGICAL LABORATORY*, *15*(4), 275–290.

Hotta, K., Mitsuhara, K., Takahashi, H., Inaba, K., Oka, K., Gojobori, T., & Ikeo, K. (2007). A web-based interactive developmental table for the ascidian Ciona intestinalis, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. *Developmental Dynamics : An Official Publication of the American Association of Anatomists*, *236*(7), 1790–805. doi:10.1002/dvdy.21188

Hozumi, A., Yoshida, R., Horie, T., Sakuma, T., Yamamoto, T., & Sasakura, Y. (2013). Enhancer activity sensitive to the orientation of the gene it regulates in the chordategenome. *Developmental Biology*, *375*(1), 79–91. doi:10.1016/j.ydbio.2012.12.012

Hsin, J., & Manley, J. L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes and Development*, 2119–2137. doi:10.1101/gad.200303.112.Transcription

Hudson, C., & Lemaire, P. (2001). Induction of anterior neural fates in the ascidian Ciona intestinalis. *Mechanisms of Development*, *100*(2), 189–203.

Ikuta, T., & Saiga, H. (2007). Dynamic change in the expression of developmental genes in the ascidian central nervous system: Revisit to the tripartite model and the origin of the midbrain-hindbrain boundary region. *Developmental Biology*, *312*(2), 631–643. doi:10.1016/j.ydbio.2007.10.005

Imai, K. S., Hino, K., Yagi, K., Satoh, N., & Satou, Y. (2004). Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. *Development (Cambridge, England)*, *131*(16), 4047–4058. doi:10.1242/dev.01270

Imai, K. S., Levine, M., Satoh, N., & Satou, Y. (2006). Regulatory blueprint for a chordate embryo. *Science (New York, N.Y.)*, *312*(5777), 1183–1187. doi:10.1126/science.1123404

Irvine, S. Q. (2013). Study of Cis-regulatory Elements in the Ascidian Ciona intestinalis. *Current Genomics*, *14*(1), 56–67. doi:10.2174/138920213804999192

Jack, J., Dorsett, D., Delotto, Y., & Liu, S. (1991). Expression of the cut locus in the Drosophila wing margin is required for cell type specification and is regulated by a distant enhancer. *Development (Cambridge, England)*, *113*(3), 735–47.

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, *3*(3), 318–356. doi:10.1016/S0022-2836(61)80072-7

Jennings, B. H., & Ish-Horowicz, D. (2008). The Groucho/TLE/Grg family of transcriptional co-repressors. *Genome Biology*, *9*(1), 205. doi:10.1186/gb-2008-9-1-205

Jenuwein, T., & Allis, C. D. (2001). Translating the histone code. *Science (New York, N.Y.)*, *293*(5532), 1074–80. doi:10.1126/science.1063127

Johnson, D. S., Davidson, B., Brown, C. D., Smith, W. C., & Sidow, A. (2004). Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Research*, *14*(12), 2448–56. doi:10.1101/gr.2964504

Johnson, D. S., Zhou, Q., Yagi, K., Satoh, N., Wong, W., & Sidow, A. (2005). De novo discovery of a tissue-specific gene regulatory module in a chordate. *Genome Research*, *15*(10), 1315–24. doi:10.1101/gr.4062605

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., … Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, *20*(6), 861–73. doi:10.1101/gr.100552.109

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., … Kivioja, T. (2013). Resource DNA-Binding Specificities of Human Transcription Factors. *Cell*, *152*(1-2), 327–339. doi:10.1016/j.cell.2012.12.009

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., … Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*. doi:10.1038/nature15518

Jonkers, I., & Lis, J. T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews. Molecular Cell Biology*, *16*(3), 167–177. doi:10.1038/nrm3953

Joshi, R., Passner, J. M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M. A., … Mann, R. S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, *131*(3), 530–43. doi:10.1016/j.cell.2007.09.024

Kadonaga, J. T. (2012). Perspectives on the RNA Polymerase II Core Promoter. *Wiley*

Interdisciplinary Reviews. Developmental Biology, *1*(1), 40–51. doi:10.1002/wdev.21

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., … Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, *467*(7314), 430–5. doi:10.1038/nature09380

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., … Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, *458*(7236), 362–6. doi:10.1038/nature07667

Karaiskou, A., Swalla, B. J., Sasakura, Y., & Chambon, J.-P. (2015). Metamorphosis in solitary ascidians. *Genesis*, *53*(1), 34–47. doi:10.1002/dvg.22824

Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K., & Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. *Nature Methods*, *11*(2), 203–9. doi:10.1038/nmeth.2766

Kawai, N., Takahashi, H., Nishida, H., & Yokosawa, H. (2005). Regulation of NF-kappaB/Rel by IkappaB is essential for ascidian notochord formation. *Developmental Biology*, *277*(1), 80–91. doi:10.1016/j.ydbio.2004.09.007

Kedes, L. H. (1979). Histone genes and histone messengers. *Annual Review of Biochemistry*, *48*, 837–70. doi:10.1146/annurev.bi.48.070179.004201

Kellum, R., & Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, *64*(5), 941–950. doi:10.1016/0092-8674(91)90318-S

Keys, D. N., Lee, B., Di Gregorio, A., Harafuji, N., Detter, J. C., Wang, M., … Richardson, P. M. (2005). A saturation screen for cis-acting regulatory DNA in the Hox genes of Ciona intestinalis. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(3), 679–83. doi:10.1073/pnas.0408952102

Kharchenko, P. V, Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., … Park, P. J. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, *471*(7339), 480–5. doi:10.1038/nature09725

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., … Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, *23*(5), 800–11. doi:10.1101/gr.144899.112

Khoueiry, P., Rothbächer, U., Ohtsuka, Y., Daian, F., Frangulian, E., Roure, A., … Lemaire, P. (2010). A cis-regulatory signature in ascidians and flies, independent of transcription factor binding sites. *Current Biology : CB*, *20*(9), 792–802. doi:10.1016/j.cub.2010.03.063

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., … Becker, T. S. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research*, *17*(5), 545–55. doi:10.1101/gr.6086307

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, *188*(4184), 107–16.

Kleinjan, D. A., & van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *American Journal of Human Genetics*, *76*(1), 8–32. doi:10.1086/426833

Kowalevsky, A. (1866). Entwickelungsgeschichte der Einfachen Ascidien. *Mémoires de l'Académie Impériale Des Sciences de St.-Pétersbourg, VIIè Série.*, *10*(15), 1–16.

Kozhevnikova, E. N., van der Knaap, J. A., Pindyurin, A. V, Ozgur, Z., van Ijcken, W. F. J., Moshkin, Y. M., & Verrijzer, C. P. (2012). Metabolic enzyme IMPDH is also a transcription factor regulated by cellular state. *Molecular Cell*, *47*(1), 133–9. doi:10.1016/j.molcel.2012.04.030

Kuratani, S., Wada, H., Kusakabe, R., & Agata, K. (2006). Evolutionary embryology resurrected in Japan with a new molecular basis: Nori Satoh and the history of ascidian studies originating in Kyoto during the 20th century. *The International Journal of Developmental Biology*, *50*(5), 451–4. doi:10.1387/ijdb.062154sk

Kusakabe, T., Yoshida, R., Ikeda, Y., & Tsuda, M. (2004). Computational discovery of DNA motifs associated with cell type-specific gene expression in Ciona. *Developmental Biology*, *276*(2), 563–80. doi:10.1016/j.ydbio.2004.09.037

Kvon, E. Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J. O., Pagani, M., Schernhuber, K., … Stark, A. (2014). Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature*. doi:10.1038/nature13395

Kvon, E. Z., Stampfel, G., Yáñez-Cuna, J. O., Dickson, B. J., & Stark, A. (2012). HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes & Development*, *26*(9), 908–13. doi:10.1101/gad.188052.112

Lagha, M., Bothma, J. P., & Levine, M. (2012). Mechanisms of transcriptional precision in animal development. *Trends in Genetics : TIG*, *28*(8), 409–16. doi:10.1016/j.tig.2012.03.006

Lai, F., & Shiekhattar, R. (2014). Enhancer RNAs: The new molecules of transcription. *Current Opinion in Genetics and Development*, *25*(1), 38–42. doi:10.1016/j.gde.2013.11.017

Lamarck, J.-B. (1816). *Histoire naturelle des animaux sans vertèbres*. Paris: Verdière.

Lamy, C., Rothbächer, U., Caillol, D., & Lemaire, P. (2006). Ci-FoxA-a is the earliest zygotic determinant of the ascidian anterior ectoderm and directly activates Ci-sFRP1/5. *Development (Cambridge, England)*, *133*(15), 2835–44. doi:10.1242/dev.02448

Lang, G., Gombert, W. M., & Gould, H. J. (2005). A transcriptional regulatory element in the coding sequence of the human Bcl-2 gene. *Immunology*, *114*(1), 25–36. doi:10.1111/j.1365-2567.2004.02073.x

Lee, A. P., & Venkatesh, B. (2013). Ultraconserved Elements (UCEs) in the Human Genome. *DOI: 10.1002/9780470015902.a0020842.pub2*. doi:10.1002/9780470015902.a0020842.pub2

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., … Przeworski, M. (2012). Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology*, *10*(9), e1001388. doi:10.1371/journal.pbio.1001388

Lemaire, P. (2009). Unfolding a chordate developmental program, one cell at a time: invariant cell lineages, short-range inductions and evolutionary plasticity in ascidians. *Developmental Biology*, *332*(1), 48–60. doi:10.1016/j.ydbio.2009.05.540

Lemaire, P. (2011). Evolutionary crossroads in developmental biology: the tunicates. *Development (Cambridge, England)*, *138*(11), 2143–52. doi:10.1242/dev.048975

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., … de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*(14), 1725–35.

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., … Ren, B. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, *518*(7539), 350–354. doi:10.1038/nature14217

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Current Biology : CB*, *20*(17), R754–63. doi:10.1016/j.cub.2010.06.070

Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, *424*(6945), 147–51. doi:10.1038/nature01763

Levo, M., & Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nature Reviews. Genetics*, *15*(7), 453–68. doi:10.1038/nrg3684

Li, J. J., & Herskowitz, I. (1993). Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science (New York, N.Y.)*, *262*(5141), 1870–4.

Li, L. M., & Arnosti, D. N. (2011). Long- and short-range transcriptional repressors induce distinct chromatin states on repressed genes. *Current Biology : CB*, *21*(5), 406–12. doi:10.1016/j.cub.2011.01.054

Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., … Biggin, M. D. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biology*, *6*(2), e27. doi:10.1371/journal.pbio.0060027

Li, X., & Noll, M. (1994). Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo. *The EMBO Journal*, *13*(2), 400–6.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A.,

… Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289–93. doi:10.1126/science.1181369

Lin, Y. C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., … Murre, C. (2012). Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nature Immunology*, *13*(12), 1196–204. doi:10.1038/ni.2432

Linné, C. von, & Salvius, L. (1758). *Caroli Linnaei...Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.* (Vol. v.1). Holmiae : Impensis Direct. Laurentii Salvii,.

Liu, H., Zhang, R., Xiong, W., Guan, J., Zhuang, Z., & Zhou, S. (2013). A comparative evaluation on prediction methods of nucleosome positioning. *Briefings in Bioinformatics*. doi:10.1093/bib/bbt062

Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., … Young, R. A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, *153*(2), 320–34. doi:10.1016/j.cell.2013.03.036

Ludwig, M. Z., Manu, Kittler, R., White, K. P., & Kreitman, M. (2011). Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. *PLoS Genetics*, *7*(11), e1002364. doi:10.1371/journal.pgen.1002364

Luster, T. A., & Rizzino, A. (2003). Regulation of the FGF-4 gene by a complex distal enhancer that functions in part as an enhanceosome. *Gene*, *323*, 163–72.

Lynch, V. J., Leclerc, R. D., May, G., & Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, *43*(11), 1154–9. doi:10.1038/ng.917

Lynch, V. J., Tanzer, A., Wang, Y., Leung, F. C., Gellersen, B., Emera, D., & Wagner, G. P. (2008). Adaptive changes in the transcription factor HoxA-11 are essential for the evolution of pregnancy in mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(39), 14928–33. doi:10.1073/pnas.0802355105

Lynch, V. J., & Wagner, G. P. (2008). Resurrecting the role of transcription factor change in developmental evolution. *Evolution; International Journal of Organic Evolution*, *62*(9), 2131–54. doi:10.1111/j.1558-5646.2008.00440.x

Magnani, L., Eeckhoute, J., & Lupien, M. (2011). Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in Genetics : TIG*, *27*(11), 465–74. doi:10.1016/j.tig.2011.07.002

Makeev, V. J., Lifanov, A. P., Nazina, A. G., & Papatsenko, D. A. (2003). Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Research*, *31*(20), 6016–26.

Man, T. K., & Stormo, G. D. (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Research*, *29*(12), 2471–2478. doi:10.1093/nar/29.12.2471

Marinić, M., Aktas, T., Ruf, S., & Spitz, F. (2013). An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape. *Developmental Cell*, *24*(5), 530–42. doi:10.1016/j.devcel.2013.01.025

Markstein, M., Markstein, P., Markstein, V., & Levine, M. S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(2), 763–8. doi:10.1073/pnas.012591199

Martinez, C. A., & Arnosti, D. N. (2008). Spreading of a corepressor linked to action of long-range repressor hairy. *Molecular and Cellular Biology*, *28*(8), 2792–802. doi:10.1128/MCB.01203-07

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-Y., Denay, G., Lee, J., … Wasserman, W. W. (2015). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*. doi:10.1093/nar/gkv1176

Mathelier, A., & Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS Computational Biology*, *9*(9), e1003214. doi:10.1371/journal.pcbi.1003214

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)*, *337*(6099), 1190–5. doi:10.1126/science.1222794

McLean, C. Y., Reno, P. L., Pollen, A. A., Bassan, A. I., Capellini, T. D., Guenther, C., … Kingsley, D. M. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, *471*(7337), 216–9. doi:10.1038/nature09774

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., … Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, *30*(3), 271–277. doi:10.1038/nbt.2137

Minokawa, T., Yagi, K., Makabe, K. W., & Nishida, H. (2001). Binary specification of nerve cord and notochord cell fates in ascidian embryos. *Development (Cambridge, England)*, *128*(11), 2007–17.

Miwata, K., Chiba, T., Horii, R., Yamada, L., Kubo, A., Miyamura, D., … Satou, Y. (2006). Systematic analysis of embryonic expression profiles of zinc finger genes in Ciona intestinalis. *Developmental Biology*, *292*(2), 546–54. doi:10.1016/j.ydbio.2006.01.024

Miya, T., & Nishida, H. (2003). An Ets transcription factor, HrEts, is target of FGF signaling

and involved in induction of notochord, mesenchyme, and brain in ascidian embryos. *Developmental Biology*, *261*(1), 25–38.

Monod, J., & Jacob, F. (1961). Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology*, *26*, 389–401.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., … Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. *Cell*, *147*(5), 1132–45. doi:10.1016/j.cell.2011.10.023

Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M., & Chambon, P. (1981). The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Research*, *9*(22), 6047–6068.

Morgan, T. H. (1944). Some further data on self fertilization in Ciona. *Journal of Experimental Zoology*, *97*(3), 231–248. doi:10.1002/jez.1400970303

Mullen, A. C., Orlando, D. A., Newman, J. J., Lovén, J., Kumar, R. M., Bilodeau, S., … Young, R. A. (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell*, *147*(3), 565–76. doi:10.1016/j.cell.2011.08.050

Najafabadi, H. S., Mnaimneh, S., Schmitges, F. W., Garton, M., Lam, K. N., Yang, A., … Hughes, T. R. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology*, *33*(5), 555–62. doi:10.1038/nbt.3128

Nakashima, K., Yamada, L., Satou, Y., Azuma, J. I., & Satoh, N. (2004). The evolutionary origin of animal cellulose synthase. *Development Genes and Evolution*, *214*(2), 81–88. doi:10.1007/s00427-003-0379-8

Nam, J., & Davidson, E. H. (2012). Barcoded DNA-tag reporters for multiplex cis-regulatory analysis. *PloS One*, *7*(4), e35934. doi:10.1371/journal.pone.0035934

Narasimhan, K., Lambert, S. A., Yang, A. W., Riddell, J., Mnaimneh, S., Zheng, H., … Hughes, T. R. (2015). Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. *eLife*, *4*. doi:10.7554/eLife.06967

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., … Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, *489*(7414), 83–90. doi:10.1038/nature11212

Nishida, H. (1987). Cell Lineage Analysis in Ascidian Embryos by Intracellular Injection of a Tracer Enzyme, *541*, 526–541.

Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., … Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, *4*. doi:10.7554/eLife.04837

Noordermeer, D., Leleu, M., Schorderet, P., Joye, E., Chabaud, F., & Duboule, D. (2014). Temporal dynamics and developmental memory of 3D chromatin architecture at Hox

gene loci. *eLife*, *3*, e02557.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., … Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–5. doi:10.1038/nature11049

Oda-Ishii, I., Bertrand, V., Matsuo, I., Lemaire, P., & Saiga, H. (2005). Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians Halocynthia roretzi and Ciona intestinalis. *Development (Cambridge, England)*, *132*(7), 1663–74. doi:10.1242/dev.01707

Ohler, U., & Wassarman, D. a. (2010). Promoting developmental transcription. *Development (Cambridge, England)*, *137*(1), 15–26. doi:10.1242/dev.035493

Ohta, N., & Satou, Y. (2013). Multiple signaling pathways coordinate to induce a threshold response in a chordate embryo. *PLoS Genetics*, *9*(10), e1003818. doi:10.1371/journal.pgen.1003818

Ohtsuki, S., Levine, M., & Cai, H. N. (1998). Different core promoters possess distinct regulatory activities in the Drosophila embryo. *Genes & Development*, *12*(4), 547–56.

Ong, C.-T., & Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, *12*(4), 283–293. doi:10.1038/nrg2957

Orenstein, Y., & Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research*, *42*(8), e63. doi:10.1093/nar/gku117

Orphanides, G., Lagrange, T., & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes & Development*, *10*(21), 2657–2683. doi:10.1101/gad.10.21.2657

Panne, D. (2008). The enhanceosome. *Current Opinion in Structural Biology*, *18*(2), 236–42. doi:10.1016/j.sbi.2007.12.002

Papatsenko, D., & Levine, M. (2007). A rationale for the enhanceosome and other evolutionarily constrained enhancers. *Current Biology*, *17*(22), 955–957. doi:10.1016/j.cub.2007.09.035

Parker, S. C. J., Hansen, L., Abaan, H. O., Tullius, T. D., & Margulies, E. H. (2009). Local DNA topography correlates with functional noncoding regions of the human genome. *Science (New York, N.Y.)*, *324*(5925), 389–392. doi:10.1126/science.1169050

Parker, S. C. J., & Tullius, T. D. (2011). DNA shape, genetic codes, and evolution. *Current Opinion in Structural Biology*, *21*(3), 342–347. doi:10.1016/j.sbi.2011.03.002

Passamaneck, Y. J., Katikala, L., Perrone, L., Dunn, M. P., Oda-Ishii, I., & Di Gregorio, A. (2009). Direct activation of a notochord cis-regulatory module by Brachyury and FoxA in the ascidian Ciona intestinalis. *Development (Cambridge, England)*, *136*(21), 3679–89.

doi:10.1242/dev.038141

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., … Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, *30*(3), 265–270. doi:10.1038/nbt.2136

Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., Struhl, K., & Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. *Genome Research*, *17*(8), 1170–7. doi:10.1101/gr.6101007

Pennati, R., Ficetola, G. F., Brunetti, R., Caicci, F., Gasparini, F., Griggio, F., … Manni, L. (2015). Morphological Differences between Larvae of the Ciona intestinalis Species Complex: Hints for a Valid Taxonomic Definition of Distinct Species. *Plos One*, *10*(5), e0122879. doi:10.1371/journal.pone.0122879

Pérez-Rueda, E., & Collado-Vides, J. (2000). The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Research*, *28*(8), 1838–47.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W. M., Solovei, I., Brugman, W., … van Steensel, B. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular Cell*, *38*(4), 603–13. doi:10.1016/j.molcel.2010.03.016

Perry, M. W., Boettiger, A. N., Bothma, J. P., & Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. *Current Biology : CB*, *20*(17), 1562–7. doi:10.1016/j.cub.2010.07.043

Perry, M. W., Boettiger, A. N., & Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13570–5. doi:10.1073/pnas.1109873108

Persikov, A. V, Wetzel, J. L., Rowland, E. F., Oakes, B. L., Xu, D. J., Singh, M., & Noyes, M. B. (2015). A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. *Nucleic Acids Research*, *43*(3), 1965–84. doi:10.1093/nar/gku1395

Peter, I., & Davidson, E. H. (2015). *Genomic Control Process: Development and Evolution*. Academic Press.

Phillips-Cremins, J. E. (2014). Unraveling architecture of the pluripotent genome. *Current Opinion in Cell Biology*, *28*, 96–104. doi:10.1016/j.ceb.2014.04.006

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., … Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, *153*(6), 1281–95. doi:10.1016/j.cell.2013.04.053

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin

accessibility data. *Genome Research*, *21*(3), 447–55. doi:10.1101/gr.112623.110

Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., … Cramer, P. (2015). Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature*, *518*(7539), 376–80. doi:10.1038/nature14229

Porrua, O., & Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol*, *16*(3), 190–202. doi:10.1038/nrm3943

Pott, S., & Lieb, J. D. (2015). What are super-enhancers ? *Nature Publishing Group*, *47*(1), 8–12. doi:10.1038/ng.3167

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(3), 784–8. doi:10.1073/pnas.72.3.784

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., … Rokhsar, D. S. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, *453*(7198), 1064–71. doi:10.1038/nature06967

Rebeiz, M., Pool, J. E., Kassner, V. A., Aquadro, C. F., & Carroll, S. B. (2009). Stepwise modification of a modular enhancer underlies adaptation in a Drosophila population. *Science (New York, N.Y.)*, *326*(5960), 1663–7. doi:10.1126/science.1178357

Reddy, K. L., Zullo, J. M., Bertolino, E., & Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, *452*(7184), 243–7. doi:10.1038/nature06727

Reverberi, G., & Minganti, A. (1946). Cell Lineage and Embryo Patterning. *Staz. Zool. Napoli*, *20*(13), 199–252.

Richmond, T. J., & Davey, C. A. (2003). The structure of DNA in the nucleosome core. *Nature*, *423*(6936), 145–50. doi:10.1038/nature01595

Ritter, D. I., Dong, Z., Guo, S., & Chuang, J. H. (2012). Transcriptional enhancers in protein-coding exons of vertebrate developmental genes. *PloS One*, *7*(5), e35202. doi:10.1371/journal.pone.0035202

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry*, *79*, 233–69. doi:10.1146/annurev-biochem-060408-091030

Roose, J., & Clevers, H. (1999). TCF transcription factors: molecular switches in carcinogenesis. *Biochimica et Biophysica Acta*, *1424*(2-3), M23–37.

Rose, S. M. (1939). Embryonic Induction in the Ascidia. *Biological Bulletin*, *77*(2), 216–232.

Rosenberg, M., & Court, D. (1979). Regulatory sequences involved in the promotion and termination of RNA transcription. *Annual Review of Genetics*, *13*, 319–353.

doi:10.1146/annurev.ge.13.120179.001535

Rotem, A., Ram, O., Shoresh, N., Sperling, R. a, Goren, A., Weitz, D. a, & Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, (October). doi:10.1038/nbt.3383

Rothbächer, U., Bertrand, V., Lamy, C., & Lemaire, P. (2007). A combinatorial code of maternal GATA, Ets and beta-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development (Cambridge, England)*, *134*(22), 4023–32. doi:10.1242/dev.010850

Roure, A., Lemaire, P., & Darras, S. (2014). An otx/nodal regulatory signature for posterior neural development in ascidians. *PLoS Genetics*, *10*(8), e1004548. doi:10.1371/journal.pgen.1004548

Ruan, Q., Kameswaran, V., Tone, Y., Li, L., Liou, H.-C., Greene, M. I., … Chen, Y. H. (2009). Development of Foxp3(+) regulatory t cells is driven by the c-Rel enhanceosome. *Immunity*, *31*(6), 932–40. doi:10.1016/j.immuni.2009.10.006

Sackerson, C., Fujioka, M., & Goto, T. (1999). The even-skipped locus is contained in a 16-kb chromatin domain. *Developmental Biology*, *211*(1), 39–52. doi:10.1006/dbio.1999.9301

Saksouk, N., Simboeck, E., & Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. *Epigenetics & Chromatin*, *8*, 3. doi:10.1186/1756-8935-8-3

Sardet, C., Paix, A., Prodon, F., Dru, P., & Chenevert, J. (2007). From oocyte to 16-cell stage: cytoplasmic and cortical reorganizations that pattern the ascidian embryo. *Developmental Dynamics : An Official Publication of the American Association of Anatomists*, *236*(7), 1716–31. doi:10.1002/dvdy.21136

Sathyapriya, R., Vijayabaskar, M. S., & Vishveshwara, S. (2008). Insights into protein-DNA interactions through structure network analysis. *PLoS Computational Biology*, *4*(9), e1000170. doi:10.1371/journal.pcbi.1000170

Satoh, N., Satou, Y., Davidson, B., & Levine, M. (2003). Ciona intestinalis: an emerging model for whole-genome analyses. *Trends in Genetics : TIG*, *19*(7), 376–81. doi:10.1016/S0168-9525(03)00144-6

Satou, Y., Imai, K. S., & Satoh, N. (2001). Early embryonic expression of a LIM-homeobox gene Cs-lhx3 is downstream of beta-catenin and responsible for the endoderm differentiation in Ciona savignyi embryos. *Development (Cambridge, England)*, *128*(18), 3559–3570.

Satou, Y., Takatori, N., Yamada, L., Mochizuki, Y., Hamaguchi, M., Ishikawa, H., … Satoh, N. (2001). Gene expression profiles in Ciona intestinalis tailbud embryos. *Development (Cambridge, England)*, *128*(15), 2893–2904.

Sayou, C., Monniaux, M., Nanao, M. H., Moyroud, E., Brockington, S. F., Thévenon, E., … Dumas, R. (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA

binding specificity. *Science (New York, N.Y.)*, *343*(6171), 645–8. doi:10.1126/science.1248229

Schaller, H., Gray, C., & Herrmann, K. (1975). Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(2), 737–41.

Schmidt, D., Schwalie, P. C., Ross-Innes, C. S., Hurtado, A., Brown, G. D., Carroll, J. S., … Odom, D. T. (2010). A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research*, *20*(5), 578–88. doi:10.1101/gr.100479.109

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., … Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)*, *328*(5981), 1036–40. doi:10.1126/science.1186176

Schoborg, T. A., & Labrador, M. (2010). The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. *Journal of Molecular Evolution*, *70*(1), 74–84. doi:10.1007/s00239-009-9310-x

Schwarzer, W., & Spitz, F. (2014). The architecture of gene expression: integrating dispersed cis-regulatory modules into coherent regulatory domains. *Current Opinion in Genetics & Development*, *27*, 74–82. doi:10.1016/j.gde.2014.03.014

Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *73*(3), 804–8.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., … Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, *442*(7104), 772–8. doi:10.1038/nature04979

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., & Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, *451*(7178), 535–40. doi:10.1038/nature06496

Segal, E., & Widom, J. (2009). From DNA sequence to transcriptional behaviour: a quantitative approach. *Nature Reviews. Genetics*, *10*(7), 443–56. doi:10.1038/nrg2591

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., … Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, *30*(6), 521–530. doi:10.1038/nbt.2205

Shefer, K., Sperling, J., & Sperling, R. (2014). The Supraspliceosome — A Multi-Task Machine for Regulated Pre-mRNA Processing in the Cell Nucleus. *Computational and Structural Biotechnology Journal*, *11*(19), 113–122. doi:10.1016/j.csbj.2014.09.008

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., … Ren, B. (2012). A map of

the cis-regulatory sequences in the mouse genome. *Nature*, *488*(7409), 116–20. doi:10.1038/nature11243

Shenkar, N., & Swalla, B. J. (2011). Global diversity of Ascidiacea. *PloS One*, *6*(6), e20657. doi:10.1371/journal.pone.0020657

Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., … Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, *32*(2), 171–8. doi:10.1038/nbt.2798

Shirangi, T. R., Dufour, H. D., Williams, T. M., & Carroll, S. B. (2009). Rapid evolution of sex pheromone-producing enzyme expression in Drosophila. *PLoS Biology*, *7*(8), e1000168. doi:10.1371/journal.pbio.1000168

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews. Genetics*, *15*(4), 272–86. doi:10.1038/nrg3682

Siepel, A., & Arbiza, L. (2014). Cis-regulatory elements and human evolution. *Current Opinion in Genetics & Development*, *29*, 81–9. doi:10.1016/j.gde.2014.08.011

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., & Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, *39*(9), 381–399. doi:10.1016/j.tibs.2014.07.002

Sleator, R. D. (2010). An overview of the current status of eukaryote gene prediction strategies. *Gene*, *461*(1-2), 1–4. doi:10.1016/j.gene.2010.04.008

Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., … Cherry, J. M. (2015). ENCODE data at the ENCODE portal. *Nucleic Acids Research*. doi:10.1093/nar/gkv1160

Small, K. S., Brudno, M., Hill, M. M., & Sidow, A. (2007). A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. *Genome Biology*, *8*(3), R41. doi:10.1186/gb-2007-8-3-r41

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews. Genetics*, *13*(9), 613–26. doi:10.1038/nrg3207

Spitz, F., Gonzalez, F., & Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell*, *113*(3), 405–417. doi:10.1016/S0092-8674(03)00310-6

Squarzoni, P., Parveen, F., Zanetti, L., Ristoratore, F., & Spagnuolo, A. (2011). FGF/MAPK/Ets signaling renders pigment cell precursors competent to respond to Wnt signal by directly controlling Ci-Tcf transcription. *Development (Cambridge, England)*, *138*(7), 1421–32. doi:10.1242/dev.057323

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., & Stark, A. (2015).

Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature*. doi:10.1038/nature15545

Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., … Stamatoyannopoulos, J. A. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, *515*(7527), 365–370. doi:10.1038/nature13972

Stern, C. D. (2005). Neural induction: old problem, new findings, yet more questions. *Development*, *132*(9), 2007–2021. doi:10.1242/dev.01794

Stern, D. L., & Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? *Evolution; International Journal of Organic Evolution*, *62*(9), 2155–77. doi:10.1111/j.1558-5646.2008.00450.x

Stockley, P. G., & Persson, B. (2009). Surface plasmon resonance assays of DNA-protein interactions. *Methods in Molecular Biology (Clifton, N.J.)*, *543*, 653–69. doi:10.1007/978-1-60327-015-1_38

Stolfi, A., Gandhi, S., Salek, F., & Christiaen, L. (2014). Tissue-specific genome editing in Ciona embryos by CRISPR/Cas9. *Development (Cambridge, England)*, *141*(21), 4115–20. doi:10.1242/dev.114488

Struhl, K., & Segal, E. (2013). Determinants of nucleosome positioning. *Nature Structural & Molecular Biology*, *20*(3), 267–73. doi:10.1038/nsmb.2506

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L., & Jones, R. T. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology*, *203*(2), 439–55.

Tassy, O., Daian, F., Hudson, C., Bertrand, V., & Lemaire, P. (2006). A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis. *Current Biology : CB*, *16*(4), 345–58. doi:10.1016/j.cub.2005.12.044

Tempel, S. (2012). Using and understanding RepeatMasker. *Methods in Molecular Biology (Clifton, N.J.)*, *859*, 29–51. doi:10.1007/978-1-61779-603-6_2

Teng, L., Firpi, H. A., & Tan, K. (2011). Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. *Nucleic Acids Research*, *39*(17), 7371–9. doi:10.1093/nar/gkr476

Thanos, D., & Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell*, *83*(7), 1091–100.

Thomas, M. C., & Chiang, C.-M. (2006). The general transcription machinery and general cofactors. *Critical Reviews in Biochemistry and Molecular Biology*, *41*(3), 105–78. doi:10.1080/10409230600648736

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., …

Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82. doi:10.1038/nature11232

Tokuoka, M., Kumano, G., & Nishida, H. (2007). FGF9/16/20 and Wnt-5alpha signals are involved in specification of secondary muscle fate in embryos of the ascidian, Halocynthia roretzi. *Development Genes and Evolution*, *217*(7), 515–27. doi:10.1007/s00427-007-0160-5

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., & de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, *10*(6), 1453–65.

Treen, N., Yoshida, K., Sakuma, T., Sasaki, H., Kawai, N., Yamamoto, T., & Sasakura, Y. (2014). Tissue-specific and ubiquitous gene knockouts by TALEN electroporation provide new approaches to investigating gene function in Ciona. *Development (Cambridge, England)*, *141*(2), 481–7. doi:10.1242/dev.099572

Trompouki, E., Bowman, T. V, Lawton, L. N., Fan, Z. P., Wu, D.-C., DiBiase, A., … Zon, L. I. (2011). Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell*, *147*(3), 577–89. doi:10.1016/j.cell.2011.09.044

Tsagkogeorga, G., Turon, X., Hopcroft, R. R., Tilak, M.-K., Feldstein, T., Shenkar, N., … Delsuc, F. (2009). An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models. *BMC Evolutionary Biology*, *9*(1), 187. doi:10.1186/1471-2148-9-187

Tümpel, S., Cambronero, F., Sims, C., Krumlauf, R., & Wiedemann, L. M. (2008). A regulatory module embedded in the coding region of Hoxa2 controls expression in rhombomere 2. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(51), 20077–82. doi:10.1073/pnas.0806360105

Ulianov, S. V, Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., … Razin, S. V. (2015). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Research*. doi:10.1101/gr.196006.115

van Arensbergen, J., van Steensel, B., & Bussemaker, H. J. (2014). In search of the determinants of enhancer–promoter interaction specificity. *Trends in Cell Biology*, *24*(11), 695–702. doi:10.1016/j.tcb.2014.07.004

van Steensel, B., & Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. *Nature Biotechnology*, *28*(10), 1089–95. doi:10.1038/nbt.1680

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature Reviews. Genetics*, *10*(4), 252–63. doi:10.1038/nrg2538

Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., … Akey,

J. M. (2012). Personal and population genomics of human regulatory variation. *Genome Research*, *22*(9), 1689–1697. doi:10.1101/gr.134890.111

Vieux-Rochas, M., Fabre, P. J., Leleu, M., Duboule, D., & Noordermeer, D. (2015). Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(15), 4672–7. doi:10.1073/pnas.1504783112

Vinson, J. P., Jaffe, D. B., O'Neill, K., Karlsson, E. K., Stange-Thomann, N., Anderson, S., … Lander, E. S. (2005). Assembly of polymorphic genomes: Algorithms and application to Ciona savignyi. *Genome Research*, *15*(8), 1127–1135. doi:10.1101/gr.3722605

Wai, H., Johzuka, K., Vu, L., Eliason, K., Kobayashi, T., Horiuchi, T., & Nomura, M. (2001). Yeast RNA polymerase I enhancer is dispensable for transcription of the chromosomal rRNA gene and cell growth, and its apparent transcription enhancement from ectopic promoters requires Fob1 protein. *Molecular and Cellular Biology*, *21*(16), 5541–53. doi:10.1128/MCB.21.16.5541-5553.2001

Wang, D., & Bodovitz, S. (2010). Single cell analysis: the new frontier in "omics". *Trends in Biotechnology*, *28*(6), 281–90. doi:10.1016/j.tibtech.2010.03.002

Wang, W., & Christiaen, L. (2012). Transcriptional enhancers in ascidian development. *Current Topics in Developmental Biology*, *98*, 147–72. doi:10.1016/B978-0-12-386499-4.00006-9

Wei, G.-H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., … Taipale, J. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO Journal*, *29*(13), 2147–60. doi:10.1038/emboj.2010.106

Weingarten-Gabbay, S., & Segal, E. (2014). The grammar of transcriptional regulation. *Human Genetics*, *133*(6), 701–11. doi:10.1007/s00439-013-1413-1

Weirauch, M. T., & Hughes, T. R. (2010). Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends in Genetics : TIG*, *26*(2), 66–74. doi:10.1016/j.tig.2009.12.002

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., … Hughes, T. R. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, *158*(6), 1431–1443. doi:10.1016/j.cell.2014.08.009

Wen, P., Crawford, N., & Locker, J. (1993). A promoter-linked coupling region required for stimulation of alpha-fetoprotein transcription by distant enhancers. *Nucleic Acids Research*, *21*(8), 1911–8.

White, M. a, Myers, C. a, Corbo, J. C., & Cohen, B. a. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11952–7. doi:10.1073/pnas.1307449110

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., … Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–19. doi:10.1016/j.cell.2013.03.035

Wilczyński, B., & Furlong, E. E. M. (2010). Dynamic CRM occupancy reflects a temporal map of developmental progression. *Molecular Systems Biology*, *6*, 383. doi:10.1038/msb.2010.35

Will, C. L., & Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, *3*(7), 1–2. doi:10.1101/cshperspect.a003707

Wilson, S. I., & Edlund, T. (2001). Neural induction: toward a unifying mechanism. *Nature Neuroscience*, *4 Suppl*, 1161–8. doi:10.1038/nn747

Wingender, E., Schoeps, T., & Dönitz, J. (2013). TFClass: An expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, *41*(D1), 165–170. doi:10.1093/nar/gks1123

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature*, *430*(6995), 85–8. doi:10.1038/nature02698

Worsley Hunt, R., & Wasserman, W. W. (2014). Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biology*, *15*(7), 412. doi:10.1186/s13059-014-0412-4

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews. Genetics*, *8*(3), 206–16. doi:10.1038/nrg2063

Wunderlich, Z., & Leonid, A. M. (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics : TIG*, *25*(10), 429–34. doi:10.1016/j.tig.2009.09.002

Yamada, L., Shoguchi, E., Wada, S., Kobayashi, K., Mochizuki, Y., Satou, Y., & Satoh, N. (2003). Morpholino-based gene knockdown screen of novel genes with developmental function in Ciona intestinalis. *Development (Cambridge, England)*, *130*(26), 6485–95. doi:10.1242/dev.00847

Yáñez-Cuna, J. O., Arnold, C. D., Stampfel, G., Boryń, L. M., Gerlach, D., Rath, M., & Stark, A. (2014). Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Research*, *24*(7), 1147–56. doi:10.1101/gr.169243.113

Yasuo, H., & Hudson, C. (2007). FGF8/17/18 functions together with FGF9/16/20 during formation of the notochord in Ciona embryos. *Developmental Biology*, *302*(1), 92–103. doi:10.1016/j.ydbio.2006.08.075

Yuh, C. H., & Davidson, E. H. (1996). Modular cis-regulatory organization of Endo16, a gut-specific gene of the sea urchin embryo. *Development (Cambridge, England)*, *122*(4), 1069–1082.

Zabidi, M. a., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., & Stark, A. (2014). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, *518*(7540), 556–559. doi:10.1038/nature13994

Zaret, K. S., & Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, *25*(21), 2227–41. doi:10.1101/gad.176826.111

Zentner, G. E., & Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. *Nature Structural & Molecular Biology*, *20*(3), 259–66. doi:10.1038/nsmb.2470

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., … Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences*, 201422023. doi:10.1073/pnas.1422023112

Zhu, J., Adli, M., Zou, J. Y., Verstappen, G., Coyne, M., Zhang, X., … Bernstein, B. E. (2013). Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, *152*(3), 642–54. doi:10.1016/j.cell.2012.12.033

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., & Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, *462*(7269), 65–70. doi:10.1038/nature08531

Title Sequence determinants of enhancer activity during *Ciona intestinalis* development

Abstract

Enhancers are crucial elements for the control of gene expression during embryonic development. The ascidian *Ciona intestinalis* offers unique experimental features to study these *cis*-regulatory sequences: enhancers are generally small and compact and their activity can be tracked at the single cell level thanks to the invariant cell lineage of ascidian embryos.

Previous work identified two independent signatures associated with enhancer activity: the presence of specific transcription factors binding sites (TFBS) and a global dinucleotide signature along enhancers (Khoueiry, 2010). Although they correlate with enhancer activity, these signatures are insufficient to identify enhancer sequences from their sole sequence.

During my thesis, I used a well-characterized early neural *Ciona* enhancer, the a-element of the Otx gene, as a model enhancer. This small (55pb) enhancer, is bound by GATA-a and ETS1/2 and is activated by the FGF pathway. To better understand the determinants of early neural enhancer activity, I tested the impact of point mutations affecting the affinity of the a-element TFBS for their binding TF and of the randomization of the spacer sequences that separate the TFBS in four ETS and GATA binding site clusters.

Our results suggest at least two levels of *cis*-regulatory control: spatiotemporal specificity of enhancer activity is encoded in the identity of TF-binding sites, while the level of enhancer activity is set both by the affinity of TFs for their binding sites and by the composition of the spacer sequences. A surprisingly high number of variants of the a-element with randomized spacers are active, always in the same cell lineages as the WT. These variants, however, display a wide range of activity levels. This effect is also observed when the spacers in another active ETS/GATA cluster are randomized. Randomization of the spacers can even confer enhancer activity to a large fraction of inactive cluster variants. Consistent with their early neural activity and with the presence of ETS- and GATA-binding sites, these variants are, like the a-element, responsive to the FGF neural inducer.

We could not link the action of the spacers on enhancer activity to any simple nucleotide or dinucleotide sequence features and it currently remains unclear why it is so easy to create a synthetic enhancer while most putative genomic ETS/GATA clusters are inactive. Using in vitro transcription factor binding assays, we showed that randomization of spacer sequences can affect TF binding to the a-element without changing the primary sequence of the binding site, and that extended minimal TFBS do not always recapitulate binding to the whole element. These results suggest that the physical structure of the DNA helix around the binding sites may play an important role in the control of enhancer activity.

Keywords Enhancers ; transcription ; transcription factors

Titre Signatures nucléotidiques de l'activité des enhancers développementaux chez l'ascidie *Ciona intestinalis*

Résumé

Les enhancers sont des régulateurs cruciaux de l'expression des gènes pendant le développement embryonnaire. L'ascidie *Ciona intestinalis* est un organisme-modèle qui se prête à l'étude de ces séquences *cis*-régulatrices car ses enhancers sont généralement petits et compacts, et le lignage invariant des cellules chez l'embryon permet de visualiser leur activité avec une résolution cellulaire.

Deux signatures indépendantes associées à l'activité d'un enhancer avaient été identifiées : la présence de sites de fixation pour des facteurs de transcription spécifiques, et une signature dinucléotidique globale à l'échelle des enhancers (Khoueiry 2010). Cependant, si ces signatures corrèlent avec l'activité des enhancers, elles ne permettent pas d'identifier de nouveaux enhancers grâce à leur séquence.

Pendant ma thèse, j'ai utilisé un enhancer neural précoce de *Ciona*, le très bien caractérisé élément-a du gène Otx, comme enhancer-modèle. Ce petit enhancer (55pb), est lié par les facteurs de transcription GATA-a et ETS1/2 et activé par la voie de signalisation FGF. Afin de mieux comprendre les déterminants de l'activité neurale précoce d'un enhancer, j'ai testé l'impact de mutations ponctuelles affectant l'affinité de sites de fixation de l'élément-a pour les facteurs de transcription. J'ai également randomisé les spacers, séquences situées entre les sites de fixation pour ETS et GATA dans quatre clusters de ces sites.

Nos résultats suggèrent au moins deux niveaux de contrôle de la régulation en *cis* : i) la spécificité spatiotemporelle de l'activité d'un enhancer est définie par l'identité des sites de fixation des facteurs de transcription, et ii) son niveau d'activité dépend à la fois de l'affinité des facteurs de transcription pour leurs sites de fixation et la composition des spacers. La majorité des variants randomisés de l'élément-a sont actifs dans les mêmes lignées cellulaires que le sauvage et leurs niveaux d'activité sont très divers. Le même résultat est obtenu en randomisant les spacers d'un autre cluster ETS/GATA actif. La randomisation de ces séquences a même conféré de l'activité enhancer à de nombreux variants de clusters inactifs. En accord avec leur activité neurale précoce et la présence de sites de fixations pour ETS et GATA, ces variants, comme l'élément-a, répondent à l'induction neurale de FGF.

Nous n'avons pas réussi à expliquer l'action des spacers sur l'activité des enhancers par des caractéristiques simples de leurs séquences (nucléotidique ou dinucléotidique), et l'on ne comprend pas pourquoi il est si simple de créer un enhancer synthétique quand la majorité des clusters génomiques de sites de fixations putatifs pour ETS et GATA sont inactifs. En utilisant une approche de fixation *in vitro* des facteurs de transcription, nous avons montré que la randomisation des spacers peut affecter la fixation d'un facteur de transcription sur l'élément a, sans changer la séquence primaire du site de fixation, mais que la fixation sur l'élément entier ne peut pas toujours être expliquée par la fixation sur les sites isolées. Ces résultats suggèrent que la structure physique de l'hélice d'ADN des enhancers peut jouer un rôle important dans le contrôle de l'activité d'un gène.

Mots-clefs Enhancers ; transcription ; facteurs de transcription