Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :
Mathilde de TAFFIN de TILQUES

Le 29 Octobre 2013

Titre :

Contrôle transcriptionnel de l'identité musculaire chez la drosophile :
Modules Cis-Régulateurs et gènes cibles directs de Collier

ED BSB : Biologie du développement

Unité de recherche :

Centre de Biologie du Développement - UMR 5547

Directeur(s) de Thèse :

Dr. Alain VINCENT

Rapporteurs :

Dr. Florence MASCHAT
Pr. François KARCH - Dr. Yacine GRABA

Autre(s) membre(s) du jury :

Pr. David CRIBBS - président
Dr. Alain VINCENT - directeur de thèse

Manuscrit de thèse

Contrôle transcriptionnel de l'identité musculaire chez la drosophile : Modules Cis-Régulateurs et gènes cibles directs de Collier

Mathilde de TAFFIN de TILQUES

Sous la direction d'Alain Vincent

mathilde.de-taffin-de-tilques@univ-tlse3.fr

alain.vincent@univ-tlse3.fr

Remerciements

A Jonathan, ma chère Ginette, qui le premier m'a parlé du DA3, de Collier, de l'embryon de drosophile et de sa passion pour la recherche… J'ai emboîté ton pas et croqué la vie de labo à ta suite, te faisant crouler sous ma tonne de questions, entre rires et immuno-colorations, pondoirs par dizaine et chansons… Malgré tes « ¡ cállate !, ¡ cállate !, que me desesperas !!!! », tu n'as jamais flanché… merci !

A Alain, qui a osé me prendre sous son aile… sachant que je ne savais rien ni de la drosophile, ni de la biologie du développement !!! Tu as compris ma soif de savoir, as été là à chacune de mes joies comme de mes déceptions, m'as épaulé, encouragé, aidé, souri ou admonesté… bref, tu as été le chef, et j'étais heureuse d'être l'élève… Dans cette école de la vie, tu m'as énormément appris… merci !

A Michèle, qui avec Alain conserve si bien l'équilibre… attentive chaque jour, scrutant aussi bien les regards que le niveau de réactif B, pour t'assurer que tout allait bien… tu as été l'âme vigilante à mes côtés… merci !

A Justine, avec qui j'ai beaucoup échangé, au labo comme ailleurs… que de découvertes en ta compagnie ! Tu as su m'encourager… yes, we did it !!! Merci…

A Yannick, mon binôme à la paillasse… et ce jusqu'à Strasbourg ! Ta présence m'a tant rassurée, que malgré les difficultés nous y sommes arrivés… Certes il reste du travail, mais c'est certain, tu sauras t'en dépatouiller… bon courage et… merci !

A Laetitia qui a partagé un bout de paillasse avec moi… supportant avec le sourire mes ritournelles et autres jacasseries ! Sans toi, je ne sais pas comment je m'en serais sortie, de mes *in situ* et de l'imagerie… pour cela et bien plus encore… merci !

A Nathalie, Caroline (et Pierre !), Anurag, Ismaël et Isabelle, l'équipe de « glandeurs » (glande lymphatique s'entend !)… avec chacun de vous j'ai noué des liens particuliers, et votre amitié m'a permis d'avancer… merci !

A Laurence, Jean-Louis et Maxime, l'équipée plus musclée… pardon pour mes impatiences, et merci de n'avoir pas capitulé ! Aujourd'hui je passe le relais, je sais qu'il sera assuré… merci !

A Anaïs, « ma » stagiaire bio-informaticienne durant tout un été… que j'ai harcelé de « et ça tu saurais le faire ? Et ce motif, à quel position le trouves-tu ? »… Avec toi, j'ai mieux compris la bio-informatique… et maintenant c'est sûr, le motif Col est bien là… merci !

A Alice, ma tutrice qui, même tardivement prévenue, a pris sa tâche tant au sérieux… Tes conseils avisés ont beaucoup participé à mon avancée… merci !

A Luis mon mexicain préféré… no sé que decirte tanto tendría que decirte… si no hubieras estado aquí a mi lado, seguro que nunca hubiera podido terminar mi doctorado… en las risas y las penas, los cantos y los experimentos, arriba o abajo, siempre tu presencia me ayudó ¿ Y sabes qué ? muchísimas gracias…

A Muriel, Henri-Marc, David, Sandra, Clément et Aissette, l'équipe « d'en-face » ! Nos échanges ont été nombreux, et variés ! Pour chacune de ces mains tendues… merci !

A Delphine, Laurence, Marie-Anaïs et Marie-Amélie, ma promotion de thèse très féminine ! L'adum nous aura bien soudées, à défaut d'avoir d'autre utilité ! Pour votre présence et votre amitié… merci !

A Tomasito, Carine, Benoît, Christophe, Fred, Aurélie, Manon, Julie, Dina, Gwenaëlle, Marion, Daniel, Fred, Julien, Thomas, Isabelle… vous avez contribué à rendre ma vie plus belle au CBD… merci !

A mes parents et frères et sœur Amaury, Bertrand, Maxence, Colomban et Clémence… vous avez suivi mes premiers pas dans la science… et ne vous êtes jamais lassés de m'entendre vous raconter le DA3…et mes déconvenues… merci !

A mes grands-parents, cousins et cousines, oncles et tantes, qui m'ont si bien entourée durant ces 4 années… en particulier Guillemette, la voix réconfortante durant mes nuits du confocal… du premier jour au dernier jour de ma thèse tu étais là… merci !

A mon arrière-garde de choc, la fmnd au grand complet, qui a tant entendu parler de mes petites mouches distinguées… pour votre soutien patient et appliqué … merci !

Et maintenant, place à la science… Oh, pas grand'chose en somme, juste quelques pattes de mouches… ou plutôt des muscles… et si les réponses tardent encore, la question, elle, n'est pas d'hier ! Voyez plutôt :

"Qui a disposé les membres d'un insecte et d'un moucheron, de manière à leur assigner une place, à leur donner une vie et un mouvement propres ? Prends et considère le plus chétif insecte, aussi petit que tu le voudras ; vois, si tu peux le comprendre, et l'ordre qui règne dans ses membres, et cette vie qui l'anime et le fait mouvoir" *S. Augustin, Discours sur le psaume CXLVIII.*

Résumé

Les facteurs de transcription (FT) métazoaires de la famille COE (Collier/Early B Cell Factor) participent au contrôle de divers processus biologiques : hématopoïèse, neurogenèse, établissement du patron des muscles. L'analyse de mutants chez divers organismes modèles montre des défauts de spécification de différents types cellulaires, dont des sous-types de neurones et, chez les mammifères, les lymphocytes B et les adipocytes bruns. Cependant les gènes cibles régulés par les facteurs COE restent majoritairement inconnus. La drosophile est un excellent modèle pour étudier la diversité fonctionnelle des FT COE. Collier (Col), le seul FT COE chez cet insecte, contrôle plusieurs processus au cours de l'embryogenèse : formation du segment céphalique intercalaire, spécification de l'identité de muscles squelettiques, de sous-types neuronaux du système nerveux central et périphérique, et de la « niche » dans l'organe hématopoïétique larvaire. Une approche gène-candidat a montré que Col régule des gènes spécifiques et différents dans chacun de ces tissus, mais les bases moléculaires de cette spécificité restaient inconnues. La musculature de l'embryon de drosophile est formée d'environ 30 muscles squelettiques différents dans chaque hemisegment, constitués chacun d'une seule fibre multi-nucléée. Il est maintenant bien établi que la combinatoire de facteurs de transcription (FT) exprimée dans les myoblastes fondateur des muscles contrôle l'identité musculaire, c'est-à-dire les caractéristiques morphologiques et fonctionnelles propres à chacun des muscles. Col est un FT identitaire, requis en particulier pour l'identité des muscles dorso-latéraux, dont le muscle DA3. Mon projet de thèse était de mettre en œuvre une recherche des gènes cibles à l'échelle génomique afin d'identifier les gènes régulés par Col au cours de la myogenèse embryonnaire et impliqués dans l'identité morphologique du muscle DA3.

Au cours de ma thèse, j'ai d'abord étudié à la régulation transcriptionnelle de *col* dans le lignage DA3, et caractérisé plus en détail les 2 Modules Cis-Régulateurs (CRM) contrôlant l'expression mésodermique de Col, respectivement aux étapes de spécification et de réalisation de l'identité musculaire. La dissection du CRM précoce (^{E}CRM) identifié par bio-informatique m'a permis d'identifier un fragment responsable uniquement de l'expression promusculaire de *col*. Le CRM tardif (^{L}CRM) avait été préalablement identifié. J'ai entrepris une étude de la fonction de sites de fixation prédits *in silico* et/ou *in vivo* pour les FT mésodermiques Twist et homéotiques Hox, par mutagenèse ponctuelle suivie de la construction de gènes rapporteurs, mais cette analyse n'a pas permis de conclusion définitive. J'ai donc commencé de développer une stratégie d'analyse des CRM de *col* dans leur contexte génomique, une stratégie qui permettra par ailleurs d'identifier les gènes cibles directs de Col selon les tissus où il est exprimé. En parallèle, j'ai contribué à l'étude du contrôle combinatoire de l'identité musculaire par Col et le FT mésodermique Nautilus (MyoD), qui montre que chaque FT contrôle des propriétés différentes du muscle DA3.

Le cœur de mes travaux de thèse a consisté à identifier les gènes-cibles directs de Collier, en particulier au cours de la myogenèse, et à caractériser les modules cis-régulateurs associés afin de comprendre les bases contextuelles de la régulation transcriptionnelle tissu-spécifique par les protéines COE. Pour cela, j'ai réalisé des expériences d'immunoprécipitation de la chromatine à partir d'embryons entiers, suivi du séquençage systématique des fragments d'ADN précipités (ChIPseq) et de leur analyse par bio-informatique. Cette analyse m'a permis d'identifier le motif ADN sélectivement reconnu par Col *in vivo*, et de montrer que cette reconnaissance est contextuelle. Plusieurs gènes cibles ont été validés par des expériences d'hybridation *in situ* et d'analyse fonctionnelle de leurs CRM, parmi lesquels une majorité d'autres FTs. L'ensemble des résultats révèle une complexité inattendue des réseaux de régulation transcriptionnelle contrôlant l'identité musculaire chez la drosophile et confirme que Col est un acteur majeur de différents réseaux dans différents tissus embryonnaires. Au vu de la conservation des FTs COE au cours de l'évolution, les conclusions de cette étude modèle chez la drosophile apportent un éclairage nouveau sur les études en cours sur des modèles mammifères.

Summary

The COE (**Co**llier/**E**arly B cell Factor) family is a metazoan-specific family of transcription factors (TF) that are involved in the control of numerous biological processes, including hematopoiesis, neurogenesis and muscle identity. Mutant analysis of COE TFs across several organisms showed defects in the specification of different cell types, like neuron subtypes or, in mammalians, B lymphocytes and brown adipocytes. However, the COE target genes are mostly unknown. *Drosophila* (fruit fly) is an excellent model to study the functional diversity of COE TFs. Collier (Col), the only COE member in this insect, controls several processes during embryogenesis: intercalary segment formation in the head, specification of somatic muscle identity, of subtypes of neurons both in the central and peripheral nervous system, and specification of the larval hematopoietic "niche". A gene candidate approach identified a few Col target genes, which appear specific and different in each of these tissues, but the molecular basis of this specificity remains unknown. The *Drosophila* embryonic musculature is composed of 30 different muscles per trunk hemisegment, each muscle constituted by a single multinucleate fiber. It is now well established that the combination of TFs expressed in the founder myoblast of a muscle controls the identity, i.e. morphological and functional properties which are characteristic of this muscle. Col acts as an identity TF in dorso-lateral muscles, in particular the DA3 muscle. My PhD project was to set up a genome wide approach to identify direct Col target genes in the mesoderm which control DA3 muscle morphology.

I first studied *col* transcriptional regulation in the DA3 muscle lineage by characterizing in more details the 2 Cis-Regulatory Modules (CRM) controlling *col* expression in the mesoderm, respectively during the specification and realization phases of muscle identity. The dissection of the Early CRM (^{E}CRM), initially identified by bio-informatics, allowed me to define a shorter fragment solely responsible for Col expression in the promuscular cluster. The late CRM (^{L}CRM) was previously characterized in the lab. I tested the function of *in silico* predicted and/or *in vivo* bound motifs for the mesodermal TF Twist, and homeotic Hox proteins, by point mutations of these motifs in a reporter gene assay. Unfortunately, this assay did not give us better insight into the direct regulatory control of *col* transcription via the ^{L}CRM. I therefore started developing a novel strategy to analyze CRMs in their genomic context, based on BAC recombineering, a strategy that will also serve to identify Col direct target genes in a tissue-specific way. In parallel, I contributed analyzing the combinatorial control of DA3 muscle identity by Col and Nautilus (MyoD), showing that each TF regulates different properties of this muscle.

The core of my PhD work was the identification of Collier direct target genes in the DA3 muscle lineage, and the characterization of the corresponding CRM to better understand how COE proteins activate specific target genes in a tissue-dependent manner. I performed chromatin immuno-precipitation on whole embryos followed by systematic sequencing of the immuno-precipitated fragments (ChIPseq). By bio-informatics, I identified Col *in vivo* binding motif and showed that Col binding *in vivo* is context-dependent. Several candidate genes were validated by *in situ* hybridizations and functional analysis of the Col binding CRM. TF are over-represented among these targets. All together, the results reveal an unexpected complexity of gene regulatory networks that control muscle identity in *Drosophila* and confirm the critical role for Col in several transcription regulatory networks in the embryo. Considering the evolutionary conservation of COE proteins and their *in vivo* DNA binding properties, these results bring new insight into the complexity of COE function in other organisms, including mammals.

Index des sigles et abréviations

AbdA/B : Abdominal A/B
Ama : Amalgam
Aret : Arrest/Bruno
AMP : Adult Muscle Progenitor, cellule "souche" musculaire adulte
Antp : Antennapedia
BLRP : Biotin Ligase Recognition Peptide
ChAPseq : Chromatin Affinity Purification & sequencing
ChIPseq : Chromatin Immuno-Precipitation & sequencing
Cnc : Cap'n'collar
CNE : Conserved Non-coding Element
Col : Collier
COE (famille) : Collier/Olfactory-1/Early B-cell Factor
CRM : Cis Regulatory Module, Module Cis-Régulateur
^{E}CRM : Early CRM, CRM précoce de Col
^{L}CRM : Late CRM, CRM tardif de Col
DA3 (muscle) : muscle Dorsal Aigu 3
DBD : DNA Binding Domain, domaine de fixation à l'ADN
Dpp : Decapentaplegic
Dys : Dystrophin
EBF : Early B-cell Factor
Eve : Even-skipped
Eya : Eyes absent
FC : Founder Cell, cellule fondatrice
FCM : Fusion Competent Myoblast, myoblaste naïf
FT : Facteur de Transcription
FTi : Facteur de Transcription identitaire
FTSS : Facteur de Transcription Site-Spécifique
GFP : Green Fluorescent Protein
Hh : Hedgehog
HLH : Helix-Loop-Helix
IPT : Immunoglobulin - Plexin - Transcription factor-like
Kr : Krüppel
Kuz : Kuzbanian
Mbl : Muscleblind
Mef2 : Myocyte enhancer factor 2
MRF : Myogenic Regulatory Factor, facteur de regulation myogénique
Mrtf ; Myocardin related transcription factor
Nau : Nautilus
Nerfin-1 : Nervous finger 1
PC : Progenitor Cell, Progéniteur
Phyl : Phyllopod
Ppk : Pick-pocket
Pum : Pumillo

PWM : Position-Weight Matrix, Matrice Poids-Position
Px : Plexus
qPCR : quantitative Polymerase Chain Reaction
Salr : Spalt-related
Sens-2 : Sensless 2
SIT : Site d'Initiation de la Transcription
Sli : Slit
Smr : Smrter
SNC : Système Nerveux Central
SNP : Système Nerveux Périphérique
So : Sine oculis
Ten-m : Tenascin major
Tin : Tinman
Tkv : Thickvein
Tl : Toll
Tup : Tailup
Twi : Twist
Ubx : Ultrabithorax
Wg : Wingless

Sommaire

I – Introduction

Une Drosophile…environ 16 000 gènes... des centaines de processus de développement parfaitement contrôlés pour, à partir d'une cellule unique, aboutir à un organisme complexe – utilisé aujourd'hui comme organisme modèle sur les études de physiologie, comportement, immunité, …, et seulement 755 facteurs de transcription prédits reconnaissant des motifs spécifiques dans l'ADN (**F**acteurs de **T**ranscription **S**ite-**S**pécifiques, FTSS) (Hens et al., 2011)…

Comment, avec ce nombre restreint de FTSS, la diversité des réseaux de régulation transcriptionnelle se met-elle en place ? Comment chaque FTSS module-t-il l'activité transcriptionnelle d'ensembles spécifiques de gènes au cours du développement ? Comment l'expression de chaque facteur est-elle elle-même régulée spatialement et temporellement ? Ces questions sont depuis bien longtemps abordées par de nombreux laboratoires (ex : (Biggin and Tjian, 1989) ; (Seyres et al., 2012)) et les résultats obtenus ont certes apporté quelques réponses… mais davantage de questions encore sur la mise en place de tels réseaux de régulation !

Depuis plusieurs années, l'équipe d'Alain Vincent et Michèle Crozatier étudie les rôles du facteur de transcription Collier chez la Drosophile. Ce facteur de transcription, identifié en 1996, est le seul membre chez la drosophile d'une famille de FTSS caractérisée par son domaine de liaison à l'ADN. Collier (Col) est exprimé dans une grande diversité de tissus au cours du développement embryonnaire et larvaire de la Drosophile. C'est donc un modèle de choix pour l'étude des propriétés régulatrices des facteurs de transcription en fonction du contexte cellulaire. Cependant, très peu de ses cibles directes ont été identifiées, par des approches classiques basées sur des gènes candidats. Les travaux réalisés depuis plusieurs années au sein de l'équipe pour décrypter le rôle de Col dans le processus de spécification de l'identité musculaire, joints aux données acquises à l'échelle génomique, en particulier par le laboratoire d'E. Furlong, sur les réseaux de régulation actifs au sein du mésoderme, créaient un contexte idéal pour caractériser les réseaux de gènes en amont et en aval de Col au cours de la myogenèse embryonnaire. C'est dans ce contexte que se placent mes travaux de thèse.

Pour commencer, je rappellerai très brièvement des notions générales sur les réseaux de régulation transcriptionnelle, puis ferai un état des lieux des connaissances acquises sur le facteur de transcription Collier et son rôle dans la mise en place de l'identité musculaire au cours de la myogenèse dans l'embryon de drosophile.

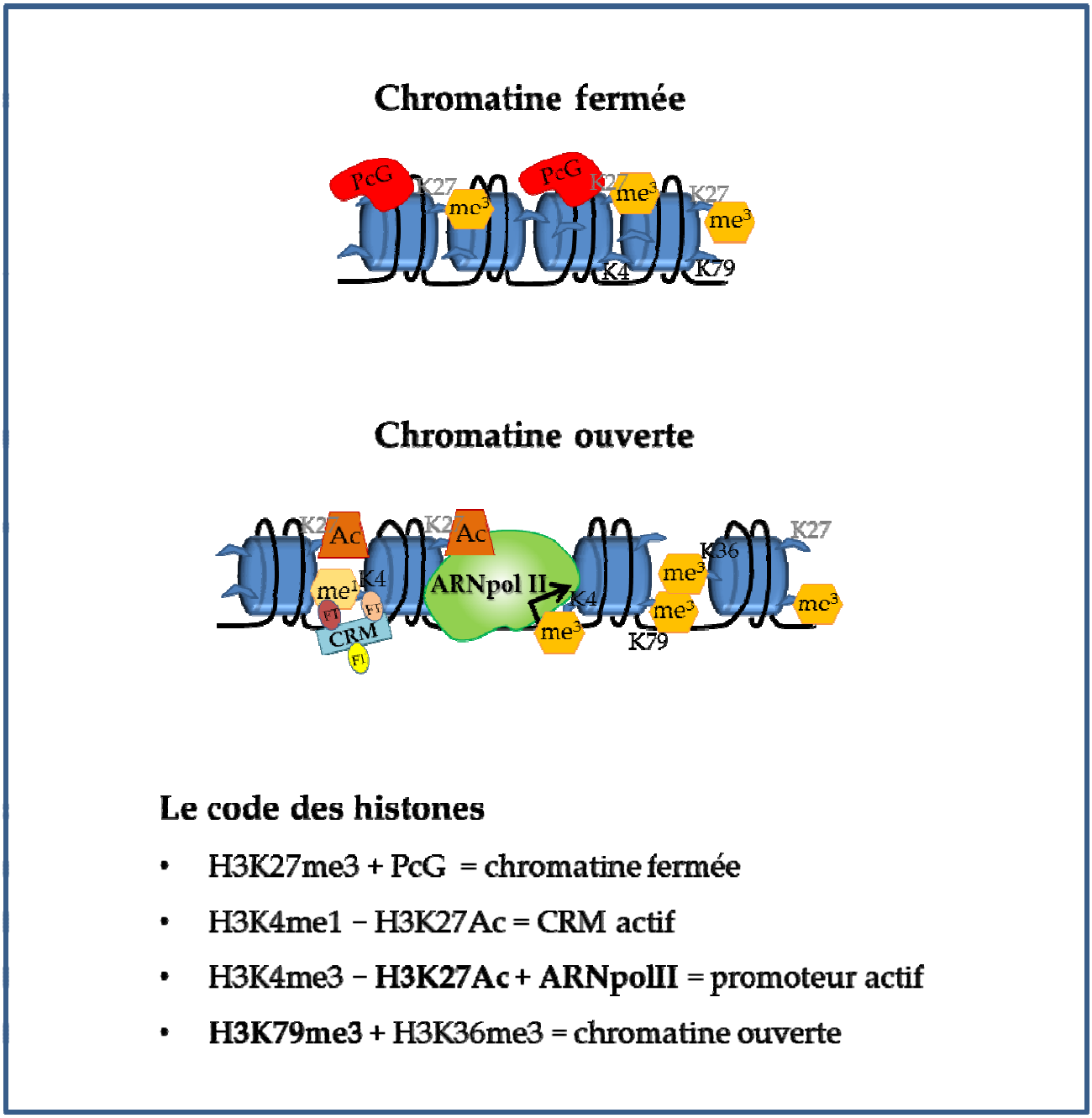


Fig. i1 – Le « code des histones »

Représentations schématiques de « l'état » ouvert ou fermé de la chromatine, associé à des modifications de résidus de l'histone H3 et à la fixation de protéines chromatinienne (PcG, ARNpolII). Les modifications post-traductionnelles des histones sont des marques de l'activité de la chromatine. En gras, les marques les plus prédictives d'après (Bonn et al., 2012).

I.1 –Réseaux de régulation transcriptionnelle : Chromatine, Facteurs de transcription et Modules Cis-Régulateurs …

Plusieurs niveaux interconnectés de mécanismes de régulations transcriptionnelles au sein d'une cellule doivent être considérés. D'abord la structure de la chromatine, ouverte ou fermée qui favorise ou réprime la transcription ; ensuite la structure des **M**odules **C**is-**R**égulateurs (CRM), séquences d'ADN regroupant des sites de fixation pour des facteurs de transcription séquence –spécifiques ; enfin les facteurs de transcription qui, en se liant à leurs sites au sein de CRM activent ou répriment la transcription des gènes associés à ces CRM.

I.1.1 – La structure de la chromatine comme révélateur de l'activité transcriptionnelle

La chromatine peut être définie comme l'ensemble de l'ADN nucléaire et des protéines et ARN qui lui sont associés. Le protéome de la chromatine est très complexe ; il comporte les histones autour desquelles s'entoure l'ADN et qui forment le squelette de la chromatine. Les histones sont soumises à de nombreuses modifications post-traductionnelles, correspondant à une chromatine dite ouverte ou fermée et qui résultent de et participent à l'interaction entre les histones et d'autres protéines chromatiniennes qui interagissent aussi entre-elles (van Bemmel et al., 2013). La composition protéique de la chromatine varie au long de la chaîne d'ADN et selon le type cellulaire. « L'état » de la chromatine, c'est-à-dire la composition en protéines liées à l'ADN et les modifications post-traductionnelles apportées à certains résidus des histones, tient une grande part dans la régulation du patron d'expression des gènes ou la réplication de l'ADN (Filion et al., 2010; Karlić et al., 2010; Levy and Noll, 1981) (Fig. I1). Par exemple la triméthylation de la lysine 27 des histones de type 3 (H3K27me3), associée à des protéines de type Polycomb définit des régions « fermée » où l'expression des gènes est réprimée (Müller et al., 2002). A l'opposé, la monométhylation de la lysine 4 des histones H3 (H3K4me1) combinée à l'acétylation de la lysine 27 des histones H3 (H3K27ac) sont des marques des éléments cis-régulateurs actifs au sein de la chromatine (Creyghton et al., 2010). L'acétylation des histones H3 sur la lysine 27 (H3K27ac), accompagnée de la triméthylation de la lysine 4 des histones H3 (H3K4me3) et de la fixation de l'ARN polymérase Pol II, identifie des régions promotrices actives (Gaertner et al., 2012; Heintzman et al., 2007). Les gènes activement transcrits sont très

souvent associés à des triméthylations des lysines 79 et 36 des histones H3 (H3K79me3 et H3K36me3), bien que la triméthylation de la lysine 36 ne soit pas requise pour une activité transcriptionnelle (Filion et al., 2010). La combinaison des marques des histones, appelée signature de la chromatine varie de manière très dynamique au cours du temps, reflétant ainsi la dynamique d'expression des gènes. L'analyse de la signature de la chromatine au cours du temps et selon les types cellulaires est un très bon révélateur de l'activité transcriptionnelle des cellules. En outre, des études récentes sur la localisation de l'ARN Polymérase II au sein de la chromatine ont montré un « pré-chargement » de la polymérase en amont de gènes clef du développement, précédent l'activation de la transcription. Ce processus de « poising » est essentiel à la synchronisation de l'expression de ces gènes lorsque les FTSS requis deviennent disponibles (Boettiger and Levine, 2009; Gaertner et al., 2012; Zeitlinger et al., 2007). Dans ce contexte, l'équipe d'E. Furlong a étudié les variations de l'état de la chromatine à différents temps durant la myogenèse embryonnaire de la Drosophile (Bonn et al., 2012). De cette analyse à l'échelle génomique, il ressort que la présence de l'ARN Polymérase II associée à l'acétylation de H3K27 sont les deux éléments les plus prédictifs des promoteurs actifs, tandis que la triméthylation de H3K79 permet une très bonne localisation des gènes en cours de transcription. Je reviendrai sur ces données tissulaires ultérieurement.

I.1.2 – Les facteurs de transcription site-spécifique

Outre les protéines impliquées dans la modification des histones et d'autres protéines chromatiniennes, il convient de distinguer deux grandes classes de facteurs de régulation de la transcription des gènes par l'ARN polymérase II: 1) les facteurs généraux associés à cette polymérase et requis pour l'initiation ou l'élongation dans toutes les cellules. On peut ranger dans cette classe les protéines du complexe Médiateur, même si la fonction de ce complexe apparait modulaire, en fonction du tissu ou de la cellule (Ries and Meisterernst, 2011). 2) les facteurs se liant à des motifs spécifiques d'ADN au sein des CRM, facteurs dits de transcription site-spécifiques (FTSS), responsables de la spécificité cellulaire/tissulaire et/ou des variations de transcription en fonction du cycle ou de la physiologie cellulaire. Cette distinction en 2 grandes classes est forcément schématique puisque certains FTSS peuvent avoir des fonctions ubiquistes. Ces facteurs n'agissent en général pas seuls mais en combinaison avec d'autres facteurs de transcription (Ravasi et al., 2010) (exemples : (Dubois et al., 2007; Halfon et al., 2000; Junion et al., 2012)). Ils favorisent le recrutement d'enzymes de modification de la chromatine qui facilitent

ou répriment l'accès de la machinerie transcriptionnelle au niveau des promoteurs des gènes qu'ils régulent (Bonn et al., 2012; Fedorova and Zink, 2008) . Les FTSS sont généralement classés en familles selon leur domaine de fixation à l'ADN.

On trouve par exemple les FTSS à homéodomaine, une des premières familles de FTSS caractérisée. Cette famille de facteurs regroupe les gènes homéotiques tels que *deformed* (*dfd*), *antennapedia* (*antp*), *ultrabithorax* (*ubx*) ou *abdominalA* ou *B* (*abdA*/*abdB*) chez la drosophile. Ce sont des acteurs majeurs du contrôle spatial du développement embryonnaire selon l'axe antéro-postérieur (Gehring, 1985). Ils se fixent sur un motif nucléotidique appelé homéobox grâce à leur homéodomaine constitué d'une structure Hélice-Tour-Hélice (HTH : 3 hélices alpha séparées par de courtes boucles ; la plus longue hélice établit la liaison spécifique avec l'ADN tandis que les deux autres forment la structure de l'homéodomaine). Bien que très similaires, les homéobox présentent tout de même de légères variations dans leur séquence qui définissent des sous-classes de FTSS à homéodomaine (Sorge et al., 2012). L'homéodomaine peut être associé à un autre domaine de liaison à l'ADN, comme dans les protéines à Paired-homéodomaine, ou un domaine d'interaction protéine-protéine spécifique tel que le domaine LIM.

Une autre famille de FTSS comprenant de nombreux membres est la famille des FT à doigts de zinc de type C2H2 (Schuh et al., 1986). Le nombre de doigts de zinc constituant leur domaine de fixation est variable et permet ainsi la reconnaissance de motifs plus ou moins étendus sur l'ADN, chaque doigt établissant un contact avec un motif de 3 à 5 nucléotides de l'ADN. Il existe d'autres types de protéines à doigts de zinc avec un domaine de liaison à l'ADN de taille fixe, tels les récepteurs nucléaires.

La famille des FTSS à domaine b-HLH (basic Helix-Loop-Helix) est caractérisée par un domaine de fixation à l'ADN composé de 2 hélices alpha reliées par une boucle. Ces facteurs se lient sous forme de dimères à une séquence consensus de type CANNTG appelée E-box. Les résidus basiques facilitent la fixation à l'ADN tandis que les hélices participent à la dimérisation (homo- ou hétéro-dimérisation) de ces facteurs (Massari and Murre, 2000).

Un motif apparenté au motif HLH a été identifié au sein d'une autre famille de FTSS, les protéines COE (Dubois and Vincent, 2001; Liberg et al., 2002; Simionato et al., 2007). J'y reviendrai.

Il existe bien d'autres familles de FTSS aujourd'hui caractérisées, que je ne détaillerai pas ici, comme la famille des FTSS à domaine MADS. Pour un bon nombre de FTSS, leurs motifs de reconnaissance/fixation sur l'ADN ont été caractérisés et sont répertoriés dans des banques de données telles que JASPAR (Portales-Casamar et al., 2010), facilitant ainsi la recherche de ces motifs sur le génome ou l'identification *de novo* de motifs sur des séquences nucléotidiques.

L'extraordinaire développement de ces banques de motifs, combiné à des algorithmes de recherche de motifs de plus en plus sophistiqués, favorise la recherche *in silico* de Modules Cis-Régulateurs (Cis-Regulatory Modules ou CRM), sur la base de combinaisons de motifs de fixation de FTSS, définissant à priori un « code de régulation transcriptionnelle » (voir ci-dessous).

I.1.3 – Les CRM : plateforme d'intégration des FTSS

Les Modules Cis-Régulateurs ou CRM, sont des régions génomiques non codantes associées à un gène et participant à la régulation de sa transcription. Un CRM peut réguler la transcription de plusieurs gènes, de même qu'un gène peut être régulé par plusieurs CRM. Les CRM regroupent des sites de fixation pour différents facteurs de transcription agissant positivement ou négativement et peuvent ainsi être considérés comme des plateformes d'intégration des FTSS (Nelson and Wardle, 2013).

Les CRM ont d'abord été identifiés à partir des régions non codantes amont ou aval des gènes par la technique des gènes rapporteurs dans lesquels la séquence codante (lacZ ou GFP par exemple) précédée d'un promoteur minimal est placée en aval du fragment génomique à tester. Un patron spatio-temporel d'expression du gène rapporteur reproduisant celui du gène endogène est un bon révélateur d'une région de contrôle transcriptionnel.

Le choix des fragments testés pour l'étude d'un gène particulier est soit guidée (hypersensibilité à la DNAse (John et al., 2013) ou historiquement carte de restriction), soit réalisée de manière systématique. La majorité des CRM étudiés ont historiquement été localisés par dissection des régions en amont, quelquefois en aval des gènes, plus rarement dans les introns. Cette technique classique d'analyse a par exemple permis d'identifier de nombreux éléments régulateurs du gène *collier*, dont le [1]CRM mésodermique sur lequel je reviendrai (Crozatier and Vincent, 1999; Dubois et al., 2007). Une des premières recherches à grande échelle (sans à priori préalable) d'éléments « enhancer » (un terme souvent utilisé pour des CRM régulant positivement l'expression des gènes) a été réalisée en 1989 par l'insertion aléatoire d'un gène rapporteur lacZ dans le génome de la drosophile, via la mobilisation d'un élément P (Bellen et al., 1989). Plus de 500 lignées ont ainsi été générées et la caractérisation des patrons d'expression du rapporteur lacZ a identifié des éléments cis-régulateurs majoritairement en amont et dans le proche voisinage des gènes. Plus récemment, une analyse systématique des régions non codantes, y compris introniques, des gènes connus ou potentiellement impliqués dans la formation du système nerveux de la drosophile a été réalisée, qui a nécessité de générer une collection d'environ 5000 lignées (Pfeiffer et al., 2008). Chaque région génomique testée a été découpée en

fragments d'environ 3 kb, chevauchant sur 1kb, selon le principe du tuilage, et chaque fragment cloné en amont du gène de l'activateur de transcription gal4 (Fischer et al., 1988). L'analyse d'un certain nombre de ces lignées a confirmé la diversité et la complexité des CRM contrôlant l'expression d'un gène. Ainsi par exemple, l'analyse systématique de ces fragment dans la région génomique de *tail-up* (*tup*) a fait apparaître 3 séquences cis-régulatrices pour ce gène (résultats du laboratoire, non publiés).

Avec l'accroissement du nombre de génomes entièrement séquencés, des nouvelles méthodes d'identification de CRM ont été mises au point à l'échelle du génome. J'en citerai deux : 1) la méthode d'immunoprécipitation de la chromatine avec des anticorps dirigés soit contre un facteur de transcription donné soit contre des marques chromatiniennes spécifiques. Cette immunoprécipitation est suivie de l'identification des fragments immunoprécipités par puce (ChIP-chip) ou séquençage (ChIPseq) (Visel et al., 2009) (ex. (Sandmann et al., 2006b; Treiber et al., 2010b)), 2) L'utilisation d'algorithmes de recherche *in silico* s'appuyant sur la connaissance préalable d'éléments de régulation du gène tels que les motifs associés à la fixation d'un facteur ou d'une combinaison de facteurs (Hartmann et al., 2013) ou des voies de signalisation régulant l'expression du gène permettant la recherche des motifs de fixation des effecteurs de ces voies aux alentours du gène (Berman et al., 2002; Berman et al., 2004; Philippakis et al., 2006).

Une des limites de l'identification à grande échelle de CRM potentiels reste la validation fonctionnelle *in vivo* de prédictions bio-informatiques ou de données de ChIPseq, qui ne peut être réalisée à la même échelle ! Afin de sélectionner les CRM potentiels les plus prometteurs, un critère additionnel souvent utilisé est la conservation de séquence au cours de l'évolution. L'hypothèse de base est que la conservation de séquence et ou de structure des régions non codantes reflète une contrainte fonctionnelle ; l'étude menée par (Pennacchio et al., 2006) utilise par exemple ce filtre et sur 167 éléments non codant conservés entre l'homme et le takifugu (poisson) ou hautement conservés entre l'homme, la souris et le rat, 75 (45%) ont été identifiés comme CRM tissu-spécifique dans l'embryon de souris… C'est également la base de la méthode de recherche de CRM proposée par (Bejerano et al., 2005). Cependant les avis sont partagés sur cette question parce que la relation conservation-fonction ne révèle parfois qu'une faible corrélation. Ainsi « seulement » 30% des CNE (Conserved Non-coding Element) testés conservés entre l'homme (CNE testés dans des embryons de souris) et le poisson-zèbre (CNE testés dans des embryons de poisson-zèbre) révèlent une conservation de fonction (Ritter et al., 2010). Même si ce critère n'est pas prédictif à 100%, la relation entre conservation de séquence et conservation de fonction est devenue un critère incontournable. Il est important de noter qu'une conclusion issue de ces études est que la conservation du motif de fixation des facteurs de

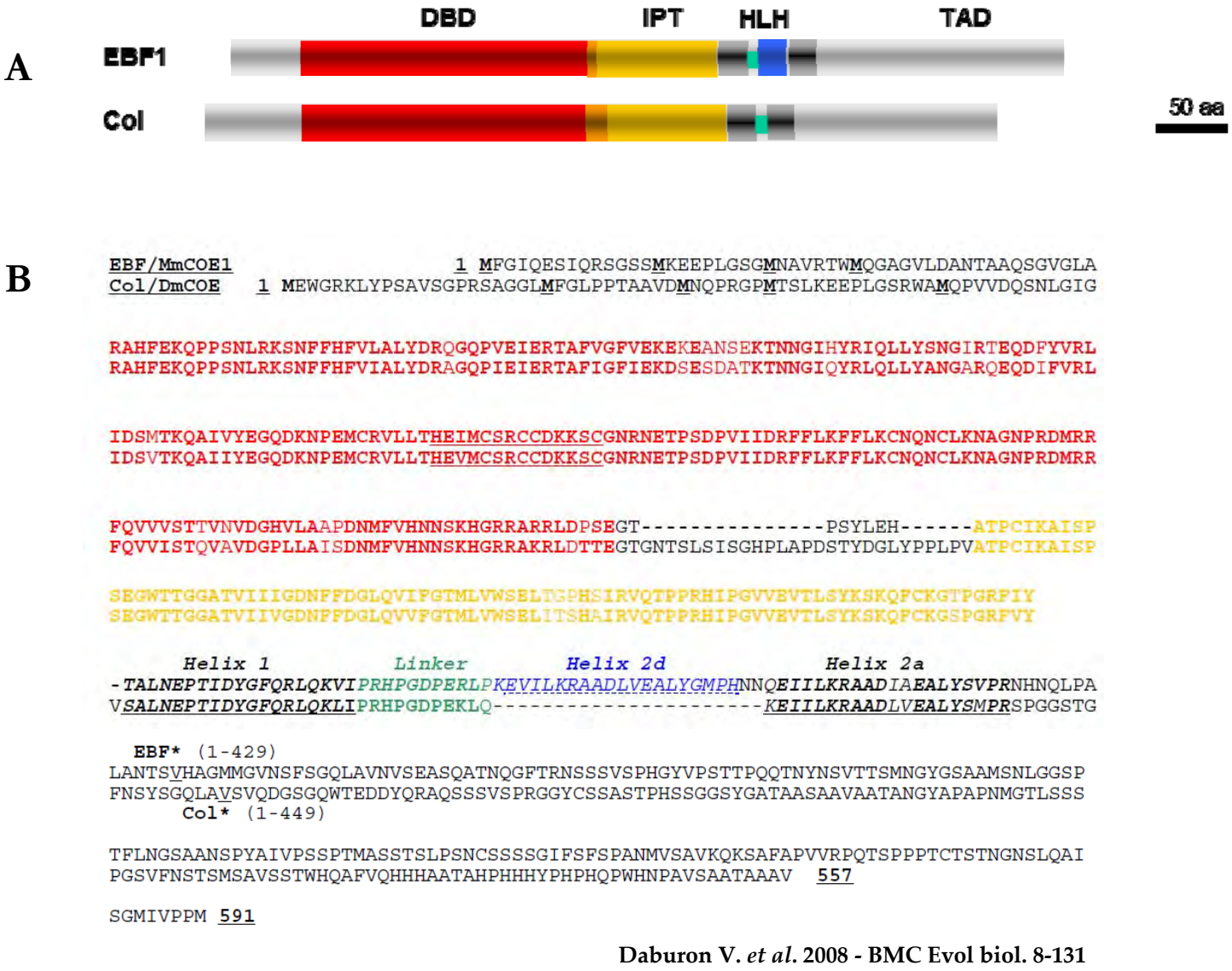
transcription, plus que la conservation de la séquence entière du CRM, joue un rôle important dans la conservation fonctionnelle du CRM. C'est l'analyse faite par (Spivakov et al., 2012) qui ajoute que la mutation de sites peut toutefois être tolérée par des effets « tampons » (motif consensus permissif notamment). Mais là encore, les avis sont partagés et des contre-exemples existent où l'identification de CRM par l'intermédiaire de groupement de sites de fixation conservés entre plusieurs espèces montre finalement que ces motifs ne participent pas à l'activité des fragments définis (Halfon et al., 2011).

L'identification bio-informatique des CRM sous-tend une question centrale : existe-t-il une « grammaire » des CRM, c'est-à-dire une organisation des CRM autour de la nature et de l'agencement des sites de fixation pour des combinatoires définies de FTSS ? Et donc une organisation repérable et identifiable en tant que CRM dans un contexte tissulaire défini. Cette notion de grammaire est suggérée par plusieurs études chez la drosophile dont l'analyse des CRM des gènes de l'immunité (Bettencourt and Ip, 2004; Senger et al., 2004) : l'association d'un site de fixation de type REL-NF κ B avec un site de type GATA (le code REL-GATA) assure une bonne prédiction des CRM « immunitaires » ; ou l'analyse de 100 CRM de gènes du développement embryonnaire précoce de la drosophile (Papatsenko et al., 2009) : différentes combinaisons de sites de fixation des facteurs de transcription Bicoid, Hunchback, Dorsal, Caudal et Twist permettent de répartir les gènes associés à ces CRM dans différents processus : mise en place de l'axe antéro-postérieur, dorso-ventral, segmentation. A contrario, une étude de (Junion et al., 2012) montre cependant qu'une grammaire précise n'est pas forcément aisément repérable du fait du recrutement collectif de certains facteurs de transcription : un facteur peut être à la fois nécessaire et détecté au niveau d'un CRM in *vivo*, sans que son motif de fixation ne soit forcément détectable *in silico*.

La conclusion de ces études montre que l'identification d'un CRM fonctionnel nécessite l'intégration de plusieurs stratégies (expériences de ChIP, recherche *in silico* de motifs, conservation de séquence à travers les espèces…) mais que sa validation ne peut pas se passer d'une validation fonctionnelle *in vivo*.

I.2 – Collier, un facteur de transcription site-spécifique modèle

Au cours de ma thèse, je me suis plus particulièrement intéressée au facteur de transcription Collier et aux modules cis-régulateurs liés par ce facteur. Dans les paragraphes suivants, je décris donc ce FTSS plus en détail ainsi que la famille de FTSS à laquelle il appartient, la famille COE.



Daburon V. *et al.* 2008 - BMC Evol biol. 8-131

Fig. 12 – Structure et conservation des facteurs de transcription COE

A. Structure schématique des protéines EBF1 (vertébrés) et Col (drosophile) montrant les différents domaines caractéristiques de la famille COE conservés au cours de l'évolution. DBD : DNA Binding Domain, IPT : Immunoglobulin-Plexin-Transcription factor ; HLH : Helix-Loop-Helix, TAD : Trans-Activating Domain. On peut noter que le domaine HLH des protéines EBF contient 3 hélices tandis que celui de Col n'en contient que 2. B. Alignement des séquences en acides aminés des protéines EBF1 et Col. Les résidus conservés sont en caractères gras.

I.2.1 – Collier, membre de la famille COE

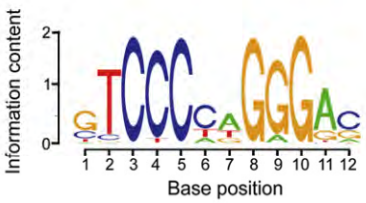
a) EBF/Olf-1 et Collier, les membres fondateurs de la famille de FTSS COE.

En 1993, deux équipes ont cloné, de façon indépendante, des ADNc codant pour des FTSS, qu'ils ont appelé EBF chez la souris pour **E**arly **B**-cell **F**actor (Hagman et al., 1993) et Olf-1 (pour **Olf**actory-1) chez le rat (Wang and Reed, 1993). Bien qu'isolés à partir de tissu différents et postulés intervenant dans des processus différents, la neurogenèse olfactive (Olf-1) et la différenciation lymphocytaire B (EBF), Olf-1 et EBF correspondent au même facteur, désormais appelé EBF1, exprimé dans de multiples types cellulaires. A cette époque, la recherche de similitudes avec les séquences protéiques présentes dans les banques de données n'a donné pour seul résultat qu'un segment d'environ 15 acides aminés correspondant à une hélice alpha retrouvée dans les facteurs de transcription de type b-HLH. Ceci a conduit l'équipe de R. Grosschedl à entreprendre une caractérisation biochimique d'EBF destinée à identifier son domaine de liaison à l'ADN (Hagman et al., 1993; Hagman et al., 1995; Treiber et al., 2010). Ce DBD (pour **D**NA-**B**inding **D**omain) est long d'environ 200 acides aminés et contient un doigt de zinc de type H-X3-C-X2-C-X5-C considéré comme un joint d'articulation entre deux sous-domaines responsables de la spécificité des contacts entre le DBD et l'ADN dans le sillon majeur. Ce doigt de zinc contribue lui-même à l'interaction avec l'ADN en contactant le sillon mineur (Treiber et al., 2010b). Le domaine HLH favorise la formation de dimères d'EBF, même en absence d'ADN. La fonction du troisième domaine hautement conservé, le domaine IPT (**I**mmunoglobulin, **P**lexin, **T**ranscription factor-like), reste mal définie. Il pourrait réguler la stabilité des complexes dimères d'EBF/ADN. Un domaine activateur de la transcription a été localisé en partie C-terminale de la protéine, dans un domaine riche en proline et en sérine.

Le clonage du gène *collier*, nommé ainsi pour son profil d'expression dans un segment gnathal de l'embryon de drosophile, a été réalisé en 1996 (Crozatier et al., 1996). La similitude de séquence entre Collier et EBF a conduit à proposer que ces 2 protéines constituaient les membres pionniers d'une nouvelle famille de facteurs de transcription, la famille COE (**C**ollier – **O**lf – **E**BF) caractérisée par un domaine de liaison à l'ADN unique à cette famille et très conservé au cours de l'évolution (cf. Fig. I2).

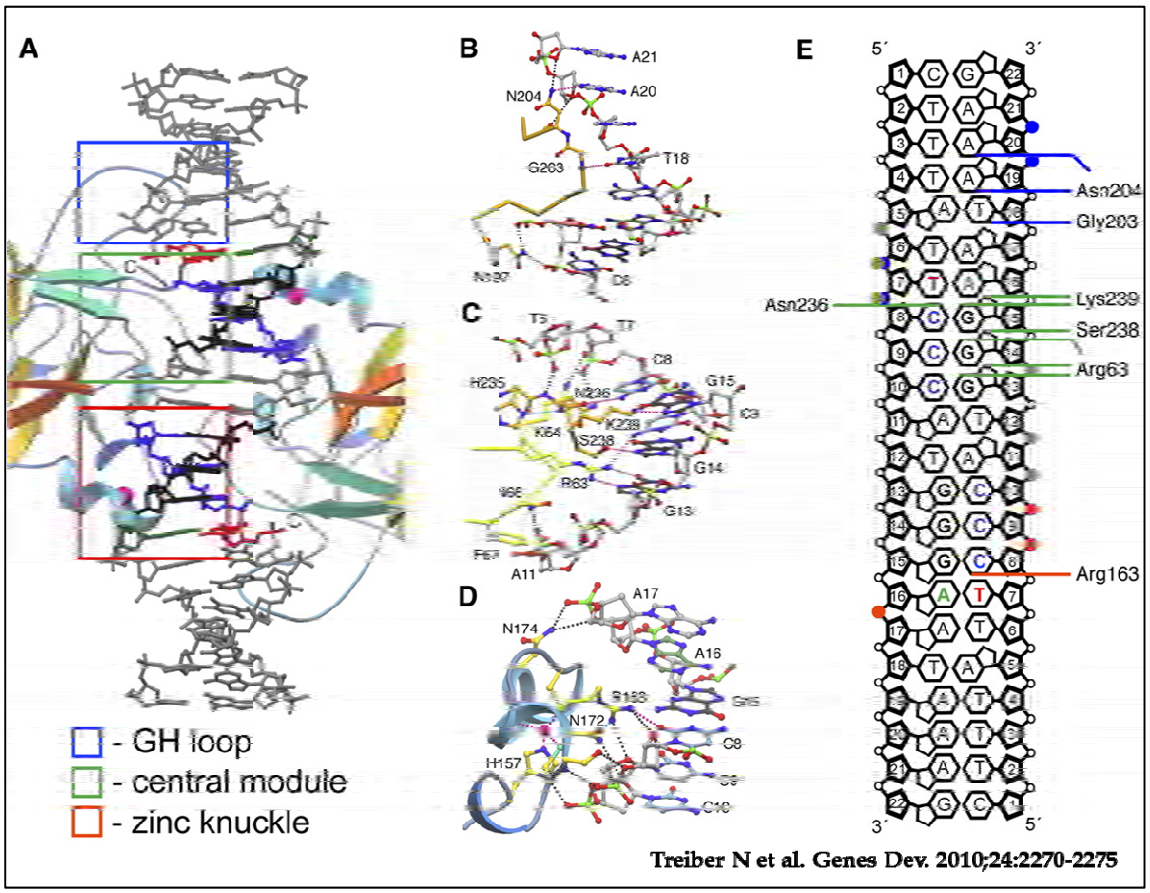
La conservation de séquence des protéines COE a rapidement été exploitée afin de rechercher d'autres membres de cette famille. Un seul membre est détecté chez la drosophile, mais cette recherche a mené à l'identification de EBF2 et EBF3, puis EBF4 chez les rongeurs, et XCOE2/ZCOE2 (EBF2) chez le xénope et le poisson-zèbre (Bally-Cuif et al., 1998; Dubois et

a



Treiber T. *et al.* 2010 - Immunity 32, 714-725

b



(A) Structure model of a DNA strand contacted by an Ebf1 dimer. The conserved nucleotides of the binding motif are colored (A [green], T [red], C [blue],G [black]). The 3 DNA interaction modules of one monomer are highlighted (GH loop [blue], major groove module [green], Zn knuckle [red]). (B–D) Detailed view of DNA contacts made by the GH loop (B), the major groove module (C), and the Zn knuckle (D). (E) Schematic summary of the DNA contacts made by one Ebf1 monomer. Hydrogen bonds to bases are indicated by lines, and backbone phosphates contacted by H bonds are marked by colored spheres. The contact modules are color-coded as in A.

Fig. 13 – Spécificité d'interaction de la protéine EBF1 avec son motif d'ADN.

a. Motif de liaison de la protéine EBF1 *in vivo* dans des lymphocytes pro-B en culture (Treiber et al., 2010) **b**. Interface de liaison à l'ADN d'EBF1. Les différentes structures présentes dans le DBD (dont le doigt de Zinc) permettent une interaction spécifique du facteur EBF avec son motif de reconnaissance sur l'ADN.

al., 1998; Garel et al., 1997; Wang et al., 2002) . Chez le nématode, un seul gène *coe* a été identifié, qui s'est révélé correspondre à *unc-3*, un gène identifié sur la base de son phénotype de mobilité désynchronisée dû à un défaut de guidance axonale de certains motoneurones (Epstein et al., 1974; Prasad et al., 1998). Une différence notable a été repérée entre les protéines des vertébrés et des invertébrés, au niveau de leur domaine de dimérisation qui comporte deux hélices alpha (H1-H2) dans *col* et *unc-3* et trois hélices (H1-H2-H2') dans EBF (Crozatier et al., 1996; Dubois and Vincent, 2001). Il s'est avéré que la structure à deux hélices est une structure ancestrale, commune à tous les métazoaires sauf les vertébrés. Cette différence reflète un évènement de duplication de l'exon codant pour l'hélice H2 à l'avènement des vertébrés (Daburon et al., 2008; Mella, 2004).

La recherche systématique d'orthologues des gènes *coe*, par BLAST sur des génomes entièrement séquencés ou des banques d'EST recouvrant un large spectre de phyla, a ensuite confirmé l'existence d'un seul gène de cette famille chez tous les métazoaires, incluant les cnidaires et les cténaires, hormis les vertébrés qui possèdent 4 gènes *ebf*, avec très probablement une situation intermédiaire de 2 gènes chez les gnathostomes (Daburon et al., 2008; Simionato et al., 2007), et l'absence de gène *coe* chez les plantes et unicellulaires. Cette phylogénie permet de classer les gènes *coe* parmi les gènes spécifiques des métazoaires (Miyata and Suga, 2001).

b) COE, une famille de facteurs de transcription site-spécifique

Les protéines EBF et Olf-1 ont été initialement identifiées sur leur capacité à reconnaitre des motifs spécifiques d'ADN dans les régions régulatrices des gènes *mb-1* (EBF) et *omp* (Olf-1) (Hagman et al., 1993; Hagman et al., 1991; Kudrycki et al., 1993; Wang and Reed, 1993). Le site optimal de liaison à l'ADN d'EBF, ensuite établi par selex, est de structure palindromique –ATTCCCNNGGGAAT, avec 2 nucléotides non contraints séparant les 2 demi-palindromes. Des homodimères EBF ou Col ainsi que des hétérodimères entre les différentes protéines EBF ou EBF/Col lient ce site avec des affinités similaires (Daburon et al., 2008). Une étude de co-cristaux DBD/ADN a confirmé que la liaison d'un dimère EBF à ce site forme un complexe parfaitement symétrique (semblable à une pince). Une analyse détaillée des résidus d'EBF responsables de la spécificité d'interaction avec l'ADN par la technique de gels-retard montre que plusieurs résidus, non essentiels à la liaison d'EBF à un site consensus palindromique parfait (palindrome raccourci –TCCCatGGGA-), sont essentiels pour la liaison à une site naturel non palindromique tel que le site TCCCatGAGA dans le promoteur de mb-1 (cf. Fig. I3).Nous

reviendrons sur les variations naturelles du site de fixation. La confirmation par cristallographie que la liaison d'un dimère d'EBF « couvre » une séquence totale d'environ 18bp sur l'ADN a mené les auteurs à émettre deux hypothèses : Cette liaison n'est possible que dans un contexte de chromatine ouverte. En retour, la liaison d'EBF serait très stable, une propriété en relation avec son rôle pionnier proposé dans la régulation des gènes spécifiques du lignage B (Treiber et al., 2010a; Treiber et al., 2010b).

La caractérisation du site de liaison *in vitro* à l'ADN d'EBF a été suivie d'une caractérisation de cibles génomiques d'EBF1 et EBF2, principalement dans les cellules Pro-B, les adipocytes, les cellules stromales de la moelle osseuse et le SNC par une approche gène-candidat. Un certain nombre de cibles directes ont été identifiées et le site de liaison d'EBF dans leur promoteur validé par des tests cellulaires de transfection transitoire avec un gène rapporteur (Jimenez et al., 2007; Lagergren et al., 2007) ; parmi les gènes candidats d'EBF1, on trouve *ebf1* lui-même pour lequel il a été suggéré un processus d'autorégulation dans les cellules pre-B (Smith et al., 2002). D'autres FTSS qui agissent aussi en collaboration avec EBF dans divers réseaux de régulation figurent parmi les gènes cibles candidats des protéines EBF, parmi lesquels Pax5 et FoxO1 dans le lignage des lymphocytes B, (Zandi et al., 2008), et Pparγ dans les adipocytes bruns. Je reviendrai sur cette question de FTSS cibles et collaborateurs dans la description des fonctions de Collier.

c) Analyse fonctionnelle des gènes ebf/coe dans des espèces modèle.

Des mutants *ebf1*-/-, *ebf2* et *ebf3* ont été générés chez la souris, par invalidation du gène par recombinaison homologue (Corradi et al., 2003; Lin and Grosschedl, 1995; Wang et al., 2004). L'absence totale de lymphocytes B dans les mutants homozygotes *ebf1*-/- a révélé le rôle spécifique de ce gène à une étape précoce de la lymphopoïèse B (revue par (Hagman and Lukin, 2005)). La comparaison des patrons d'expression des gènes *ebf1*, *ebf2* et *ebf3* montre qu'*ebf1* est le seul membre de la famille exprimé dans les lymphocytes B. La situation est plus complexe au niveau d'autres tissus et cellules. Par exemple, les 3 gènes montrent des patrons d'expression largement chevauchants dans le système nerveux central (Garel et al., 1997; Wang et al., 1997) ou encore dans les condensation mésenchymateuses puis les perichondrium dans les bourgeons de membre (Mella et al., 2004). Une analyse focalisée des mutants *ebf1*, *2* et *3* et de doubles mutants *ebf2/3* a identifié de nombreux phénotypes dans le SNC (Corradi et al., 2003; Croci et al., 2006; Garel et al., 2000; Garel et al., 1999). D'autres fonctions des gènes *ebf* ont ensuite été identifiées, notamment pour *ebf2* dans les ostéoblastes immatures formant la niche endostéale dans la moelle

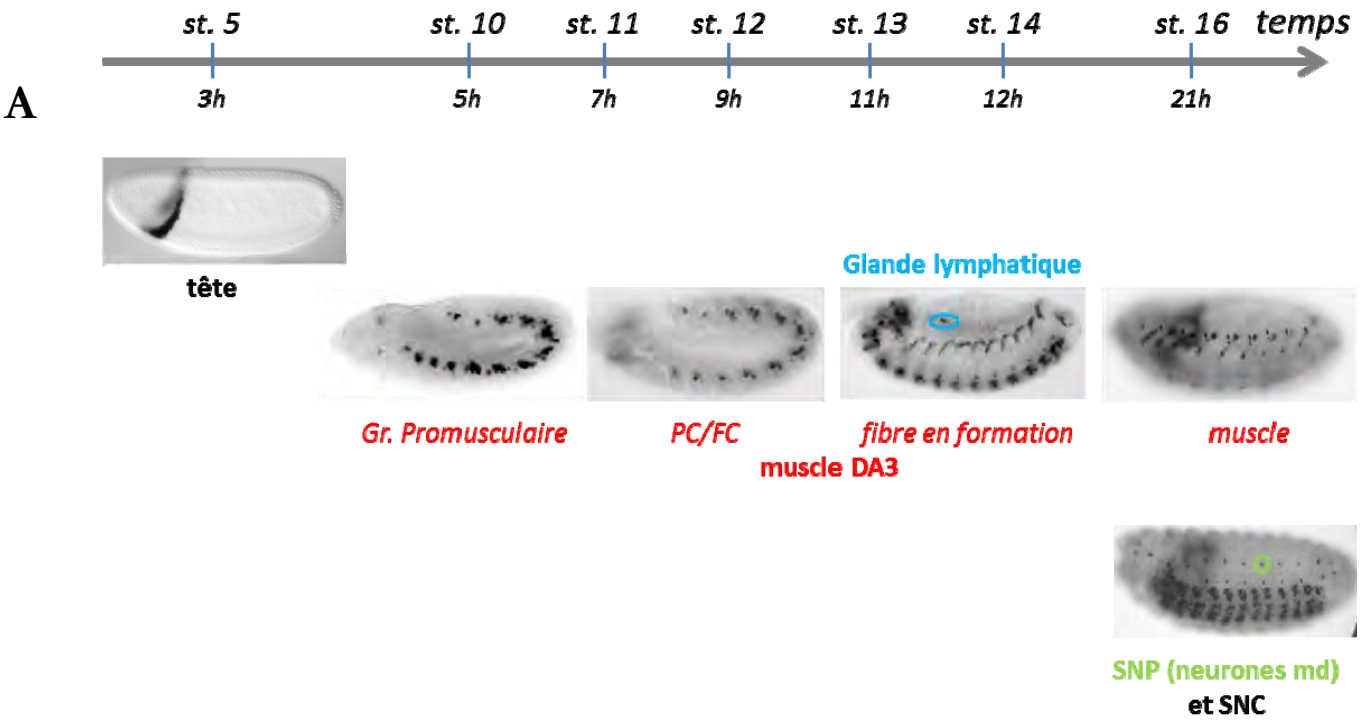
osseuse (Kieslinger et al., 2005; Kieslinger et al., 2010). Une présentation de l'ensemble des données d'analyse phénotypique et moléculaire des mutants *ebf* chez la souris dépasse certainement le cadre de cette introduction. Il est dommage qu'aucune revue générale sur les protéines *coe/ebf* n'ait été publiée depuis 2002.

Chez le Xénope et le poisson zèbre, l'expression de *Xcoe2/Zcoe2* est d'abord détectée dans les cellules précurseurs des neurones primaires. Une double approche de perte et de gain de fonction a révélé le rôle de ce gène dans le processus de transition des neurones primaires entre la compétence et l'engagement irréversible vers le destin neural (Bally-Cuif et al., 1998; Dubois et al., 1998). Un rôle similaire a ensuite été postulé pour *Xebf3* (Pozzoli et al., 2001). Plus récemment, il a été montré un rôle des protéines *Xebf2* et *3* dans la formation des muscles squelettiques (Green and Vetter, 2011). La seule étude à ce jour du rôle des EBF dans la myogenèse, l'observation de l'expression d'*Xebf2* dans le mésoderme pharyngal à l'origine des muscles faciaux (Dubois, 1999; Stolfi et al., 2010), qui peut être corrélée à l'expression de *C-coe* dans le territoire à l'origine des muscles du sillon atrial chez l'ascidie (Stolfi et al., 2010), suggère un rôle spécifique dans des sous-ensemble de muscles mais cette possibilité reste à explorer. Outre chez le nématode C. *elegans* et la drosophile, la fonction des gènes *coe* a été peu explorée en dehors des vertébrés. On peut cependant noter la description des patrons d'expression chez divers phyla, incluant un cténaire, un cnidaire, un mollusque, des annélides et un échinoderme (Demilly et al., 2011; Jackson et al., 2010; Pang et al., 2004), montrant une expression dans des tissus d'origine ectodermique et mésodermique.

I.2.2 – Collier, facteur de transcription multitâche

a) Collier possède différentes fonctions dans le développement de la drosophile

L'analyse des phénotypes mutants *collier* a été faite en plusieurs étapes. Les analyses de perte de fonction induites par l'expression contrôlée d'ARN antisens ont d'abord mis en évidence un rôle clef de ce gène dans la formation du segment intercalaire dans la région gnathale de l'embryon, un phénotype reflétant l'expression de *col* au stade blastoderme, rôle confirmé ensuite par l'analyse des mutants amorphes *col*^{ll} (Crozatier et al., 1996, 1999). L'analyse plus complète des sites d'expression et phénotypes embryonnaires des mutants a ensuite révélé une fonction dans la formation d'un muscle somatique spécifique, le muscle DA3 - ce point sera détaillé dans le chapitre 2 (Crozatier and Vincent, 1999)-, des lignages neuronaux à l'origine de neurones peptidergiques (Baumgardt et al., 2007), les neurones multidendritiques de classe IV (Crozatier



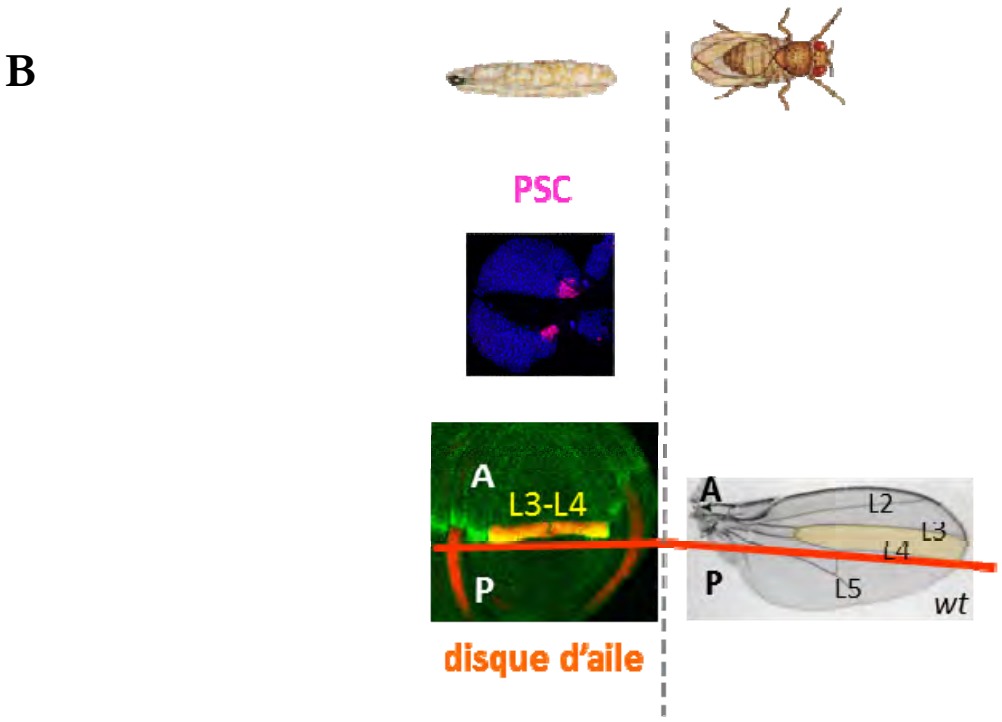


Fig. i4 – Les domaines d'expression de Collier

A. Domaines d'expression de Col au cours de l'embryogenèse. Immuno-colorations avec l'anticorps anti-Col sur des embryons sauvage aux différents stades indiqués en haut de la figure. **B**. Domaines d'expression de Col dans la larve de drosophile. L'expression de Col dans la glande lymphatique définit le PSC (**P**ost Signalling **C**enter ; *en rose* : immuno-coloration avec l'anticorps anti-Col sur une glande lymphatique). L'expression de Col dans le disque d'aile larvaire (*en orange* : immuno-coloration avec l'anticorps anti-Col dans le disque d'aile) contrôle la formation du domaine inter-nervure 3 -4 de l'aile adulte, coloré en jaune.

and Vincent, 2008; Hattori et al., 2007; Jinushi-Nakao et al., 2007). En parallèle, la construction d'un transgène permettant de sauver la létalité embryonnaire des mutants *collier* a permis de révéler des fonctions plus tardives de *collier* au cours du développement larvaire, dans l'organisation spatiale de l'aile, et le contrôle de l'homéostasie cellulaire dans la glande lymphatique, l'organe hématopoïétique (Krzemień et al., 2007; Vervoort et al., 1999) (cf. Fig. I4).

La diversité des sites d'expression et des fonctions de Col souligne l'importance du contrôle de l'expression de *col*, et pose aussi la question de la nature des gènes régulés par Col dans différents lignages cellulaires. La même question peut d'ailleurs être posée pour tous les membres de la famille COE. A ce jour aucune étude comparée des gènes cibles de ces protéines entre deux tissus n'a été réalisée à l'échelle génomique, si l'on excepte différentes lignées pro-B en culture (Treiber et al., 2010b).

b) Mais très peu de ses cibles sont connues…

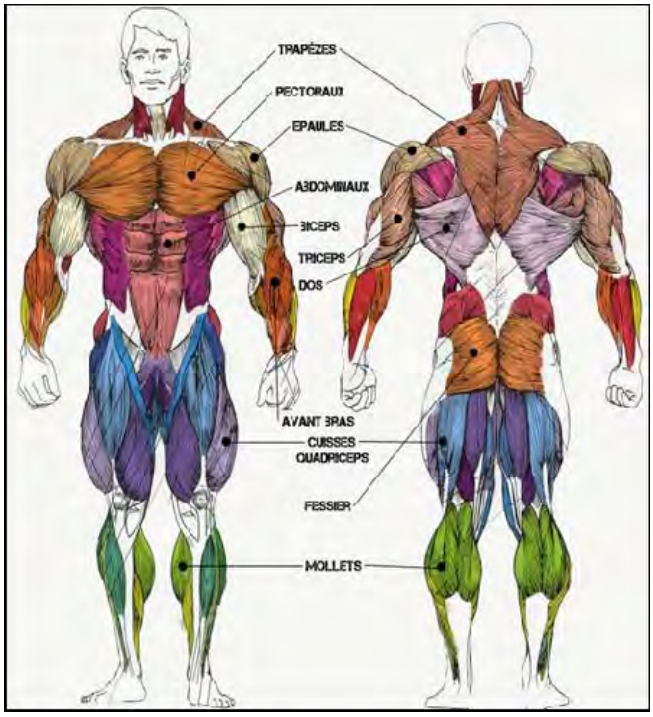
Divers gènes régulés par Col ont été identifiés au cours de l'analyse de mutants perte de fonction. Dans tous les cas, il s'agit de gènes cible lignage-spécifiques (contexte-dépendants). Parmi les gènes régulés au cours de l'embryogenèse, on peut citer *hedgehog* (*hh*) et *cap'n'collar* (*cnc*) dans le segment intercalaire (Crozatier et al., 1999; Ntini and Wimmer, 2011), *col* lui-même dans le muscle DA3 (Dubois et al., 2007), *apterous* (*ap*), *eyes absent* (*eya*) et *dimmed* (dimm) dans des neurones du SNC (Baumgardt et al., 2007) et *pickpocket* (*ppk*) (Crozatier and Vincent, 2008) dans les neurones multidendritiques. Des expériences de gain de fonction montrent que Col a une capacité très limitée à activer ces gènes dans d'autres territoires ou cellules. Plusieurs interprétations peuvent être données. La capacité de Collier à activer la transcription de gènes cibles spécifiques pourrait dépendre de 1) la structure chromatinienne de ces gènes conditionnant la liaison de Collier, et donc de l'histoire transcriptionnelle de la cellule ; 2) de la présence de co-facteurs spécifiques agissant sur les mêmes CRM (cf. Discussion). Une tentative pour répondre à cette question a été la recherche d'interacteurs directs des protéines EBF/Collier par des cribles double-hybride chez la levure.

c) Partenaires des protéines COE/EBF : un seul interacteur direct identifié

À ce jour, le seul interacteur direct identifié d'EBF, sur la base d'un crible double-hybride chez la levure, est ROAZ/Oaz/Zfp423. ROAZ comporte 30 doigts de zinc de type C2H2, et se lie à des séquences spécifiques d'ADN, palindrome parfait ou imparfait de la séquence GCACCC, les 2 demi-palindromes étant séparés par 2 paires de bases [GCACCC(A/T)(A/T)GGGTGC] (Tsai and Reed, 1997, 1998). Les doigts 38-30 sont responsables de l'interaction de ROAZ avec EBF. Bien qu'il ait été postulé, sur la base de tests en culture cellulaire, que ROAZ inhibe les fonctions activatrices d'EBF, ce rôle n'a pas été beaucoup documenté *in vivo*. Un seul article récent montre que le maintien de l'expression de ROAZ au cours de la différenciation des neurones olfactifs (sous le contrôle des régions cis-régulatrices d'*ebf3*) interfère avec l'expression des récepteurs olfactifs (une fonction associée à l'activité d'EBF/Olf-1) et dépend de la présence des doigts 28-30 (Roby et al., 2012). Mais cette évidence reste très indirecte. Une complication de l'interprétation du rôle de ROAZ est que ce facteur se lie lui-même à un motif d'ADN très semblable au site consensus de liaison d'EBF via ses doigt 5 à 7 (Roby et al., 2012; Tsai and Reed, 1998). Un deuxième crible double hybride a été réalisé avec les protéines EBF1 humaines et Collier de drosophile. Un certain nombre d'interacteurs potentiels ont été identifiés mais parmi eux aucun FTSS, hormis EBF lui-même (Formstecher et al., 2005).

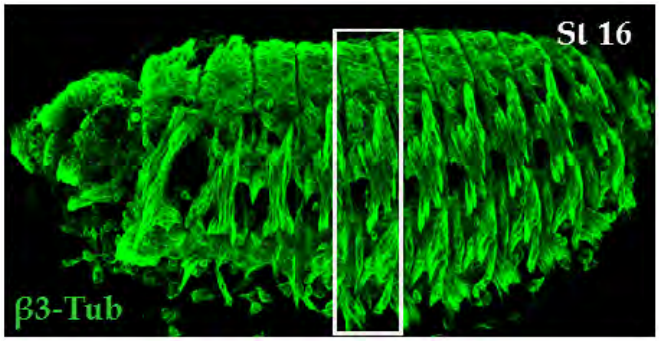
En dehors des interacteurs physiques, les protéines EBF coopèrent avec d'autres facteurs de transcription afin de promouvoir tel ou tel destin cellulaire. Au sein des lymphocytes pro-B par exemple, EBF collabore avec les facteurs E2A/E47 et Pax5 pour permettre la différenciation de ces cellules (O'Riordan and Grosschedl, 1999; Sigvardsson et al., 2002; Sigvardsson et al., 1997). En amont du gène *mb1*, EBF interagit avec E47 pour induire des modifications épigénétiques et favoriser l'ouverture de la chromatine, puis collabore avec Pax5 pour activer *mb1*, cette fonction étant potentialisée par une coopération avec Runx1 (Maier et al., 2004). Chez la drosophile, une interaction de Col avec l'isoforme B de Cnc permet d'activer *hh* dans le segment intercalaire de la tête (Ntini and Wimmer, 2011) tandis que Col et Nautilus (D-MyoD) coopèrent dans le mésoderme pour la formation de plusieurs muscles dorso-latéraux (cf. II.3) (Enriquez et al., 2012).

A



Environ 570 muscles différents…

B



… 30 muscles différents par hemisegment

C

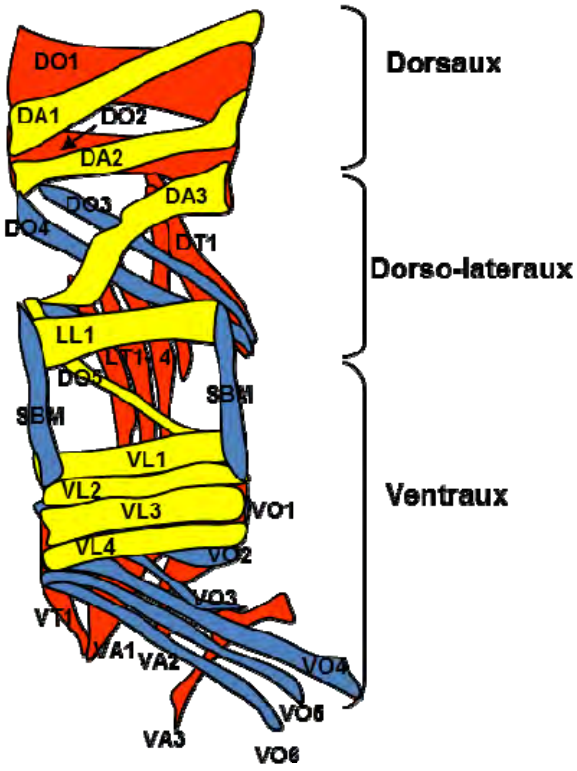


Fig. 15 – Musculatures du corps humain et de la larve de Drosophile

A : Planche d'anatomie de la musculature humaine : un bel exemple de la diversité des identités musculaires. **B** : Musculature somatique de l'embryon de drosophile. Immuno-coloration réalisée avec l'anticorps anti β3-tubulin montrant le patron musculaire d'un embryon au stade 16. **C**. Représentation schématique du patron musculaire d'un segment abdominal. Chaque muscle est identifié par 2 lettres suivies d'un chiffre. D : dorsal, L : latéral, V: ventral. O: oblique, A : aigu, L: Longitudinal, T : transversal. SBM muscle à la frontière segmentaire. En rouge sont représentés les muscles superficiels, en bleu les muscles intermédiaires, et en jaune les muscles profonds.

I.3 – Collier et la myogenèse : rôle dans la mise en place de l'identité musculaire

I.3.1 – La myogenèse : généralités

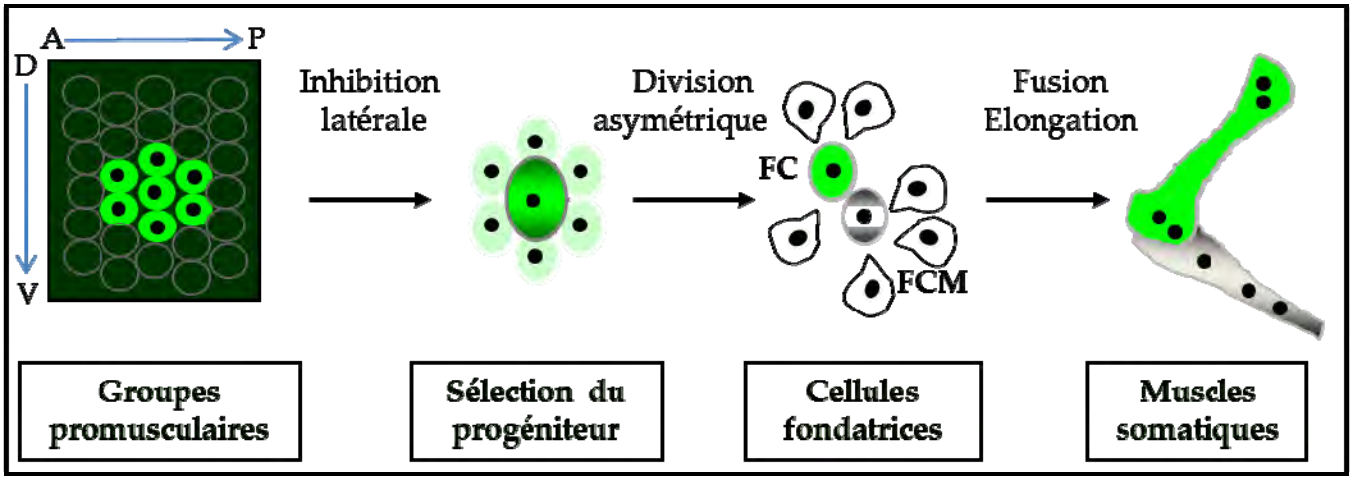
La musculature de chaque animal est composée d'un ensemble stéréotypé de muscles qui partagent des propriétés générales, mais sont morphologiquement différents les uns des autres : l'observation de la musculature humaine en est une belle illustration (Fig. I5). Tous les muscles squelettiques sont composés de fibres syncitiales formées par fusion de myoblastes mononucléés, mais chaque muscle possède une identité propre, c'est-à-dire des caractéristiques d'orientation, taille, forme, connexion au squelette et innervation qui lui sont spécifiques. Chez les vertébrés, le réseau de régulation de la transcription contrôlant la spécification des progéniteurs musculaires a été bien caractérisé. Le rôle primordial des facteurs de régulation myogénique (MRF) Myf5, Mrf4, MyoD et Myogénine, FTSS à domaine bHLH (basic-Helix-Loop-Helix) a particulièrement été décrit dans ce processus (Bryson-Richardson and Currie, 2008; Buckingham, 2006). Des FTSS Mef2 de la famille MADS et les protéines à homéodomaine de la famille Six sont aussi impliquées (Black and Olson, 1998; Richard et al., 2011). Cependant, les mécanismes responsables de l'identité de chaque muscle restent indéfinis.

La musculature somatique larvaire de la drosophile est particulièrement bien adaptée à l'étude du contrôle génétique et moléculaire de l'identité musculaire. Chaque hémisegment thoracique et abdominal de la larve montre un ensemble de 30 muscles somatiques différents connectés à l'exosquelette, un muscle alaire assurant la connexion du tube cardiaque (l'équivalent du cœur) à l'épiderme, et 6 cellules précurseurs des muscles adultes (AMP ; cellules « souches musculaires » adultes) (Baylies et al., 1998). Chaque muscle somatique est composé d'une seule fibre multinucléée (cf. Fig. I5).

I.3.2 – La myogenèse chez la drosophile

a) *Du groupe promusculaire à la fibre musculaire : les différentes étapes de la formation d'un muscle*

La formation des muscles larvaires de la drosophile a lieu durant l'embryogenèse et comprend 4 grandes étapes (Bate and Rushton, 1993) (cf. Fig.I6) : **1**) spécification de groupes de cellules



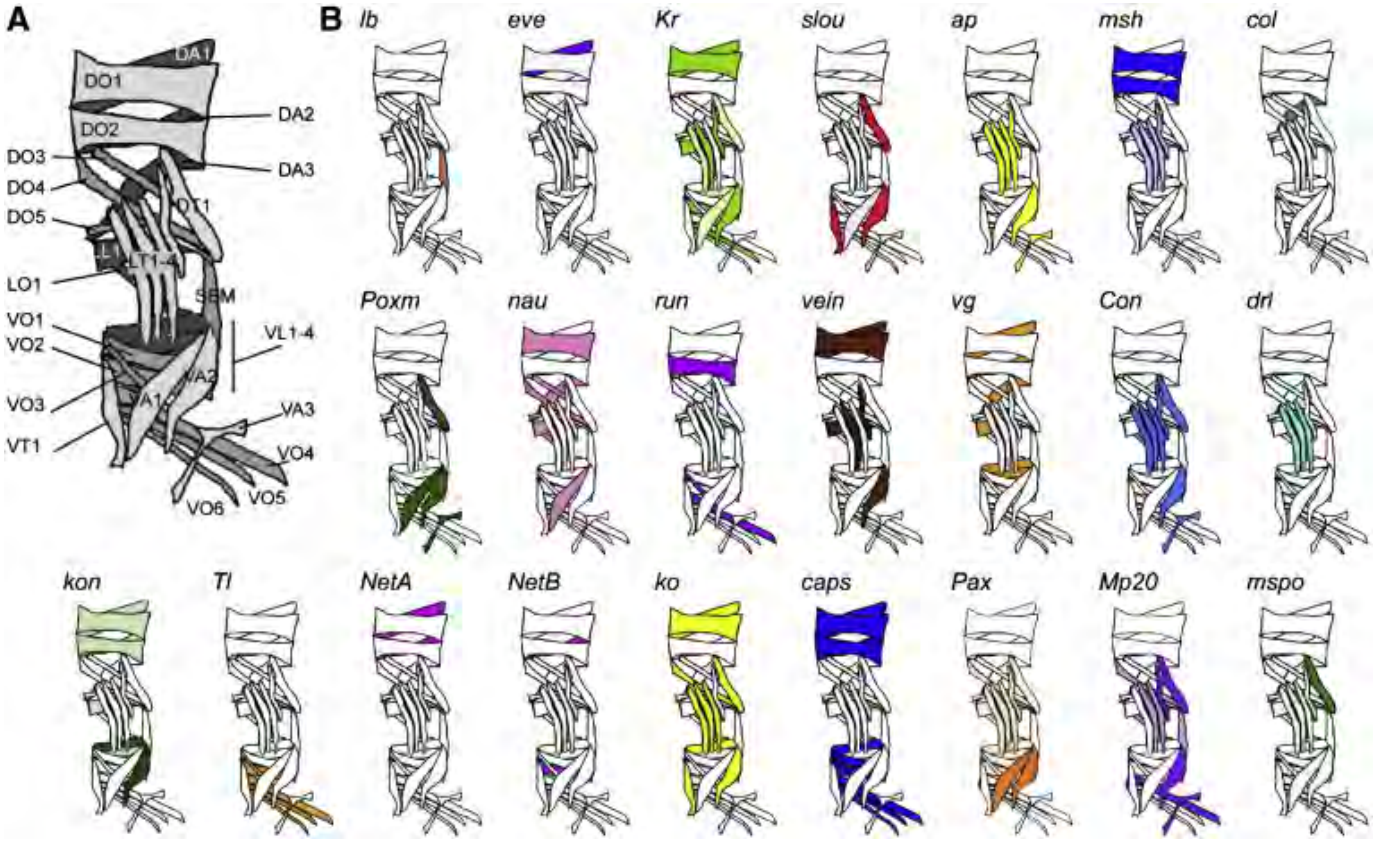
Adapté de Baylies et al., 1998

Fig. I6 – Les principales étapes de la formation des muscles squelettiques chez la drosophile

Représentation schématique des 4 grandes étapes permettant la formation d'un muscle somatique larvaire. Les groupes promusculaires sont sélectionnés à une position donnée au sein du mésoderme selon les axes antéro-postérieur (A-P) et dorso-ventral (D-V). Un processus d'inhibition latérale impliquant la voie Notch/delta permet de sélectionner des progéniteurs au sein de ces groupes promusculaires. Les myoblastes non-sélectionnés deviennent des myoblastes dits « naïfs », ou FCM (**F**usion **C**ompetent **M**yoblast). Chaque progéniteur se divise de manière asymétrique et donne naissance (majoritairement) à deux cellules fondatrices d'identités différentes, qui fusionnent avec des FCM pour former une fibre musculaire. Durant le processus de fusion, la fibre s'allonge pour atteindre ses points d'ancrage à l'épiderme et former ainsi le muscle larvaire.

équivalentes, appelés groupes promusculaires, à des positions précises au sein du mésoderme, en réponse à une information de position issue de l'ectoderme via les signalisations Dpp et Wg ; Ces groupes de cellules sont caractérisés par l'expression de la protéine Lethal of scute (L'sc). **2**) sélection d'une cellule appelée « cellule progéniteur (PC) » à partir de chaque groupe promusculaire, médiée par un processus d'inhibition latérale impliquant la voie Notch. Les cellules non sélectionnées deviennent des myoblastes compétents pour la fusion (FCM). **3**) division asymétrique de chaque cellule progéniteur, donnant naissance à deux cellules fondatrices (FC), ayant la capacité de fusionner avec un nombre fini et déterminé de FCMs ; chaque FC est à l'origine d'un muscle particulier. **4**) Le processus de fusion FC/FCM est suivi de l'élongation de la fibre et de son attachement à l'exosquelette *via* des cellules de tendon. L'ensemble de ces étapes permet d'aboutir à un patron musculaire précis, très stéréotypé au sein de chaque hémisegment. On peut distinguer 3 groupes de muscles, les muscles dorsaux, dorsaux-latéraux et ventraux (cf. Fig.I5). La nomenclature utilisée pour identifier chaque muscle reflète sa position et son orientation : ainsi par exemple le muscle DA3 est le **3**e muscle **D**orsal suivant l'axe dorso-ventral ayant une orientation « accent aigu » ou « **A**cute » (Ruiz-Gómez, 1998). À noter que certaines cellules progénitrices sont à l'origine non pas de deux muscles mais d'un muscle et d'une AMP qui reste quiescente durant les étapes du développement embryonnaire et ne sera réactivée qu'au cours de la métamorphose.

2 facteurs de transcription majeurs assurent la régulation de la myogenèse générale dans l'embryon de drosophile : Twist (Twi), facteur de transcription à domaine b-HLH, et Mef2 (Myocyte enhancer factor 2), facteur de transcription à domaine MADS. Le plus proche des MRF vertébrés en terme de fonction, Twi contrôle la spécification précoce du mésoderme pour permettre la formation des myoblastes. D'abord exprimé dans tout le mésoderme, l'expression de Twi est ensuite restreinte dans chaque segment à un domaine de faible expression à l'origine des muscles cardiaques et viscéraux et un domaine de forte expression à l'origine des muscles somatiques (Baylies and Bate, 1996). C'est à partir de cette spécification du mésoderme somatique que le processus en 4 étapes nécessaire à la formation des muscles (cf. ci-dessus) débute. Mef2 intervient ensuite dans toutes ces étapes, d'abord sous le contrôle de Twi (Nguyen et al., 1994). Mef2 est initialement exprimé dans l'ensemble du mésoderme puis son expression se restreint au mésoderme viscéral et somatique et aux cellules précurseurs du cœur. Dans un mutant *twi* thermosensible (permettant ainsi de sauver la fonction précoce de Twi dans la gastrulation), le processus de myogenèse est abortif, entraînant la perte de tous les muscles somatiques. A l'inverse, l'expression ectopique de Twi dans l'épiderme est capable de reprogrammer des cellules épidermiques en cellules présentant de nombreux marqueurs



Tixier V. *et al.* 2010 – Exp. Cell Res. 316

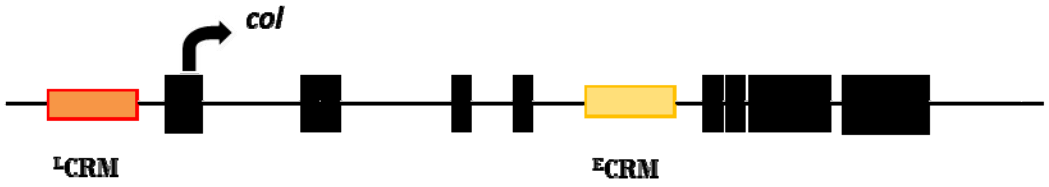
Fig. 17 – Exemples de FTi et leur patron d'expression dans les muscles larvaires

A. Rappel du patron musculaire larvaire observé dans les segments A2-A7 de l'embryon. **B**. Patrons d'expression différentielle de chacun des FTi. Une coloration pâle traduit une expression transitoire ou de faible niveau.

musculaires (myosine par ex.) et capable de fusionner (Baylies and Bate, 1996). Dans un mutant *mef2*, les cellules fondatrices des muscles se forment mais sont incapables de fusionner avec les myoblastes naïfs qui sont alors incompétents pour la fusion, entraînant l'absence de muscles somatiques (et viscéraux) (Ranganayakulu et al., 1995). L'activation ectopique de Mef2 dans l'épiderme permet de réactiver un certain nombre de marqueurs musculaires dans ces cellules sans pour autant les rendre compétentes pour la fusion.(Lin et al., 1997). Des expériences de ChIPchip ont été menées par l'équipe d'E. Furlong pour ces deux facteurs, permettant de dessiner un réseau de régulation myogénique (Sandmann et al., 2007; Sandmann et al., 2006b). Ces analyses ont montré que Twi et Mef2 se fixent en amont de nombreux gènes impliqués dans la myogenèse, plus spécifiquement des gènes de différenciation musculaire pour Mef2 tandis que Twi se fixe à la fois sur des gènes de spécification précoce du mésoderme et des facteurs de transcription responsables de l'identité propre de chaque muscle (facteur de transcription identitaire ou FTi).

b) Acquisition de l'identité musculaire : exemple du muscle DA3

Le modèle actuel, proposé en 1990 (Bate, 1990; Bate and Rushton, 1993; Bourgouin et al., 1992) est que la diversité des identités musculaires reflète l'expression, par chaque PC puis FC, d'une combinatoire spécifique de FTSS appelés facteurs de transcription identitaires (FTi). Une vingtaine de FTi ont déjà été décrits dans différents muscles (pour revue: (Tixier et al., 2010)) (cf. Fig. I7). Even-skipped (Eve) et Tailup/Islet1 (Tup) sont des exemples de FTis impliqués dans la spécification des muscles dorsaux ; Col est un FTi requis dans les muscles dorso-latéraux dont le muscle DA3 (Boukhatmi et al., 2012; Carmena et al., 1998; Dubois et al., 2007). Col est exprimé au cours des 4 étapes successives de formation du muscle DA3, du stade groupe promusculaire au muscle mature, sous le contrôle de 2 CRM : un CRM précoce ou ^{E}CRM (early CRM) du stade groupe promusculaire au stade progéniteur, et un CRM tardif ou ^{L}CRM (late CRM) qui prend le relais à partir du stade progéniteur jusqu'à la fin de la formation de la fibre musculaire (Fig. I8). On peut distinguer deux grandes phases pour la mise en place de l'identité musculaire : 1) spécification et 2) réalisation, connectant information de position et morphologie des muscles. La première phase de « spécification » de l'identité s'étend du stade groupe promusculaire au stade progéniteur. Elle débute par l'activation de FTi spécifiques dans des groupes de myoblastes (groupes promusculaires) à des positions précises au sein du mésoderme. L'activation de cette expression repose sur la lecture de l'information de position par les myoblastes (Carmena et al., 1995). Des études pionnières (Buff et al., 1998; Halfon et al., 2000; Knirr and Frasch, 2001) sur le FTi Eve a montré que l'information de position en provenance de l'ectoderme est transmise aux



Gène *col* : chr2R:10660152-10686459
LCRM (2,6-0,9) : chr2R:10687378-10689157 (Dubois et al., 2007)
ECRM (CRM276) : chr2R:10664068-10666459 (Enriquez et al. 2010)

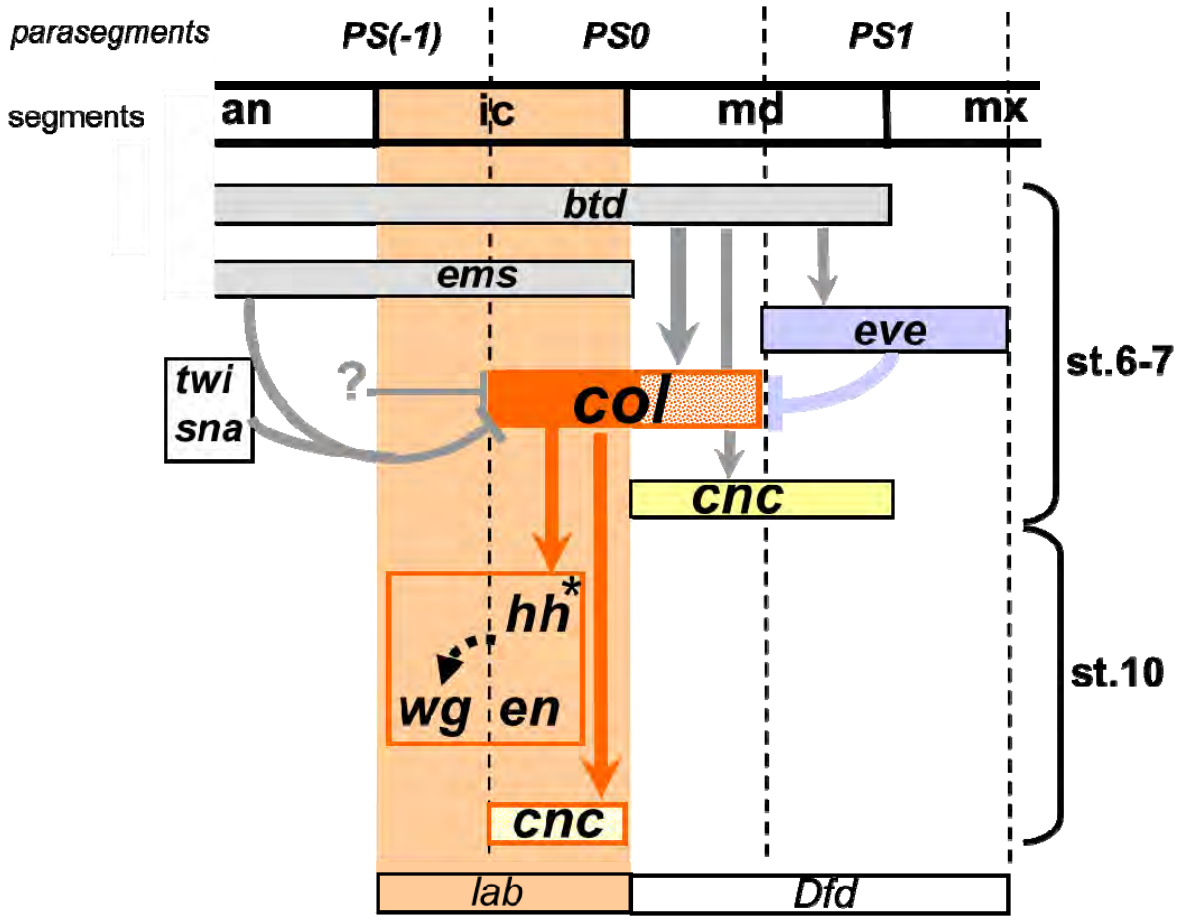
Fig. 18 – Le gène *col* et ses CRM musculaires

L'expression de *col* dans le muscle DA3 est sous le contrôle de 2 CRM distincts : le ECRM ou CRM précoce, situé dans le 5ème intron du gène, qui permet l'expression de col du stade groupe promusculaire au stade progéniteur et le LCRM ou CRM tardif, en mont du site d'initiation de la transcription, qui prend le relais à partir du stade progéniteur jusqu'à la formation de la fibre musculaire DA3.

myoblastes par la liaison des effecteurs des voies de signalisation Decapentaplegic (Dpp) et Wingless (Wg) (respectivement Mad et dTCF) sur des CRM "promusculaires" des FTi. Des interactions croisées entre différents FTis au cours du processus de sélection des PC à partir de ces groupes promusculaires puis dans les FCs issues de la division asymétrique des PCs stabilisent l'identité transcriptionnelle de chaque FC (Jagla et al., 1998; Schaub et al., 2012). A noter, l'action de la voie Notch au cours de la division asymétrique permet de générer deux identités cellulaires à partir de chaque PC (Crozatier and Vincent, 1999; Ruiz Gómez and Bate, 1997). Depuis récemment, on sait que le processus de sélection séquentielle de plusieurs PC à partir d'un même groupe promusculaire participe à la spécification de l'identité de chaque PC, ajoutant une composante temporelle à la composante spatiale de la régulation des FTi (Boukhatmi et al., 2012). La phase de « réalisation de l'identité » est l'activation/modulation différentielle de gènes effecteurs/réalisateurs de l'identité par chaque combinatoire de FTi. Très peu de gènes réalisateurs ont à ce jour été identifiés, en particulier à partir de l'étude des cibles de Lb (Bataillé et al., 2010). Parmi ces gènes, on trouve des gènes de régulation du cytosquelette tels que *mspo*, *mp20* ou *paxillin*. Les profils d'expression des quelques gènes cibles directes ou indirectes identifiés ne permettent pas d'apporter une réponse définitive à la question soulevée par la diversité des FTi et de leurs profils d'expression au cours de la myogenèse : existe-t-il des gènes réalisateurs de l'identité propres à chaque muscle ou l'identité des muscles résulte-t-elle d'une modulation spécifique de l'expression de gènes réalisateurs communs à tous les muscles par chaque combinatoire de FTis?

I.4 – Collier et la segmentation de la tête dans l'embryon.

Le contrôle génétique et moléculaire de la segmentation du tronc et de l'identité de chaque segment de l'embryon de drosophile est devenu un modèle d'école des cascades de régulation génique. Il implique l'action séquentielle de gènes maternels, tel *bicoid* (*bcd*), de gènes de segmentation gap comme *krüppel* (*kr*), de parité segmentaire comme *even-skipped* (*eve*), puis des gènes homéotiques et de polarité segmentaire comme *hedgehog* (*hh*) (Ingham, 1988). Contrairement à la segmentation du tronc, la segmentation de la tête est difficile à interpréter du seul point de vue morphologique. Le laboratoire d'Herbert Jäckle a proposé que l'expression partiellement superposée et combinatoire des gènes "gap-like" de la tête, *orthodenticle* (*otd*), *empty spiracles* (*ems*) et *buttonhead* (*btd*), déterminait à la fois la limite et l'identité des segments céphaliques (Pankratz and Jäckle, 1993). La létalité embryonnaire des mutants *col* est consécutive à la malformation du



Crozatier M. *et al.* 1999 – Dev. 126, 4385-4394

Fig. 19 – Réseau de régulations transcriptionnelles contrôlant la formation du segment intercalaire de la tête.

L'expression de *col* dans le PS0 aux stades 7 à 10 est responsable de l'activation de *hh* et *cnc* dans le segment intercalaire dans l'embryon (Crozatier et al., 1999). Une régulation directe de *hh* par Col (*), a été récemment mise en évidence (Ntini and Wimmer, 2011).

pharynx, un dérivé des segments gnathaux intercalaire et mandibulaire. L'expression de Col au stade blastoderme, uniquement dans les cellules précurseur d'un para-segment (intercalaire postérieur – mandibulaire antérieur) a suggéré un autre mécanisme de segmentation : l'existence de gènes de parité segmentaire, différents de ceux exprimés dans le tronc, agirait en aval des gènes gap-like et en amont des gènes de polarité segmentaire dans les segments gnathaux (Crozatier et al., 1996). Les études réalisées par la suite au laboratoire ont permis de vérifier cette hypothèse et de proposer une nouvelle cascade de régulation de la segmentation dans les segments céphaliques postérieurs (Crozatier et al., 1999) (Fig. I9). À noter dans cette cascade, la régulation locale des gènes *cnc* et *hh* dans les segments gnataux. Je reviendrai sur cette régulation dans le chapitre résultats.

II – Résultats

L'objectif de ma thèse était d'étudier la régulation transcriptionnelle de *collier* (*col*) dans le mésoderme et les gènes cibles de Collier dans le lignage musculaire DA3. Ces travaux sont décrits ci-dessous dans les parties II.1 et II.2. J'ai également participé à une étude sur le contrôle combinatoire de l'identité du muscle DA3 par Collier et Nautilus/MyoD, sous la direction de Jonathan Enriquez, alors en dernière année de thèse. Cette étude, publiée en 2012, est résumée brièvement dans la dernière partie des résultats (II.3).

II.1 – Etude de la régulation transcriptionnelle de Collier : analyse des CRM de *col in vivo*

Au début de mon stage M2R, il était acquis que le contrôle de la transcription de *col* durant la formation du muscle DA3 était assuré par 2 CRM (Modules Cis Régulateurs) distincts. Un CRM précoce, le ^{E}CRM, active la transcription de *col* dans un large groupe promusculaire et la maintient transitoirement, au moins, dans le ou les progéniteurs issus de ce groupe promusculaire (Enriquez et al., 2012). Un CRM tardif, le ^{L}CRM, prend le relais du ^{E}CRM au stade progéniteur et assure la transcription de *col* au cours de la formation du muscle DA3 jusqu'à la fin du processus de myogenèse (Dubois et al., 2007; Enriquez et al., 2010).

Au début de ma thèse, j'ai poursuivi l'étude de cette régulation transcriptionnelle de *col*, avec deux stratégies : une analyse classique de promoteur utilisant des gènes rapporteurs et visant à préciser la position du ^{E}CRM de *col* et l'importance de la fixation de divers FT sur le ^{L}CRM (II1.1 et II1 .2) ; une stratégie à plus long terme, permettant de tester fonctionnellement les rôles respectifs des ^{E}CRM et ^{L}CRM dans l'établissement de l'identité du muscle DA3. La mise en œuvre de cette deuxième stratégie, ralentie par des difficultés imprévues, est encore en cours (voir Discussion). Je me contenterai donc de la décrire dans son principe et d'en relater les premières étapes réalisées (II.1.3).

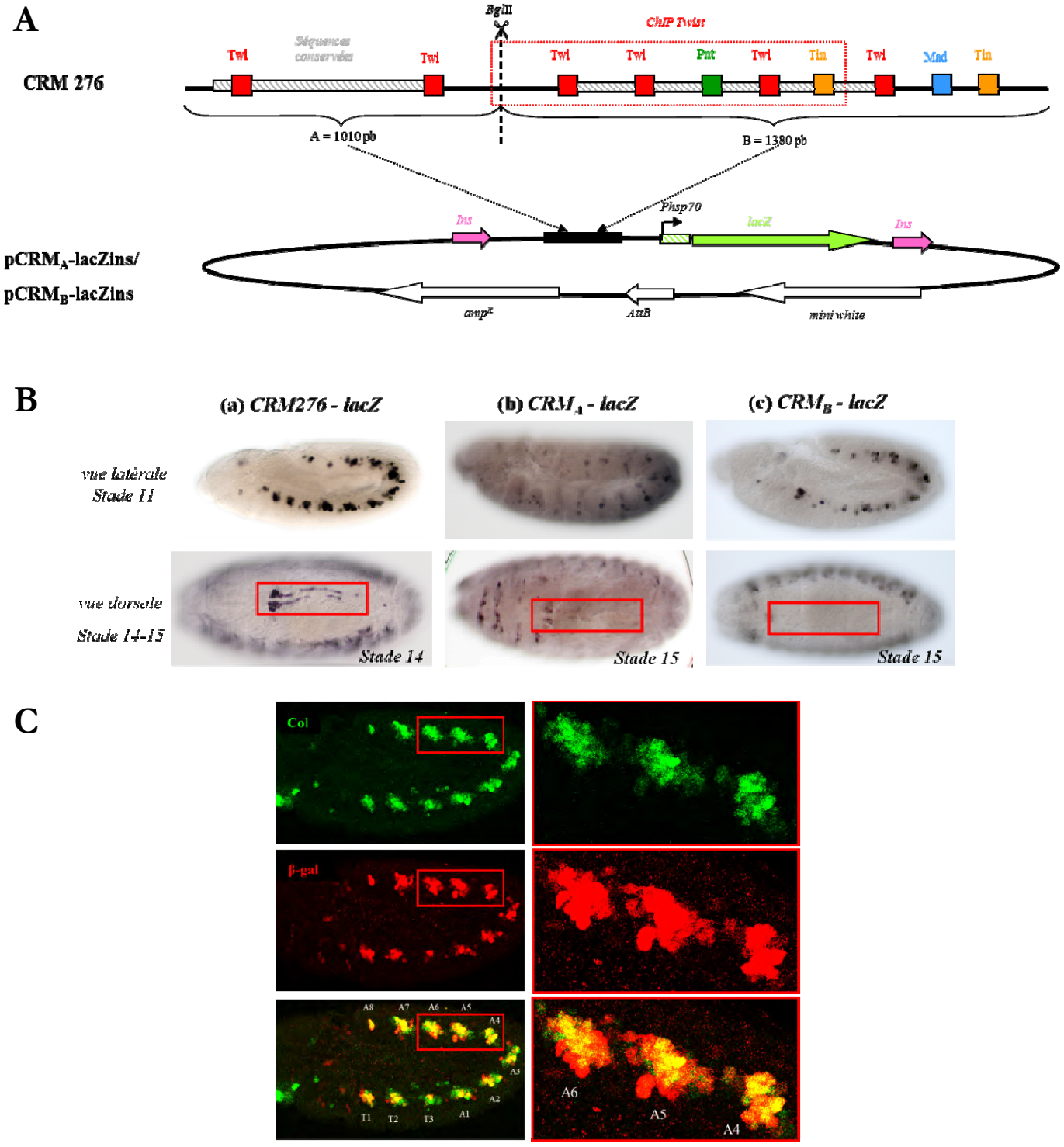


Fig. R1 – Dissection du CRM276 pour la restriction du ^{E}CRM

A. Schéma des transgènes construits pour le découpage du CRM276. **B.** Immuno-coloration d'embryons transgéniques CRM276-lacZins, CRM$_A$-lacZins et CRM$_B$-lacZins avec un anticorps anti-βgalactosidase. La détection du gène rapporteur lacZ met en évidence une expression du CRM276 dans les groupes promusculaires puis les progéniteurs (stade 11) à l'origine du muscle DA3, reproduite par le CRM$_B$ (vues latérales). Le marquage observé dès le stade 14 dans la glande lymphatique à partir du CRM276 n'est reproduit avec aucun des sous-fragments (vues dorsales). L'emplacement de la glande lymphatique est signalé par un rectangle rouge. **C.** Double immuno-coloration d'embryons transgéniques CRMB-lacZins au stade 10 avec des anticorps anti-Col (vert) et anti-βgalactosidase (rouge). Colonne de gauche : Le double marquage (vue du bas) montre un recouvrement presque parfait entre l'expression de Col (vue du haut) et de la β-galactosidase (vue intermédiaire). La position des segments T1 à A8 est indiquée. La région encadrée (segments A4, A5 et A6) est agrandie dans la colonne de droite.

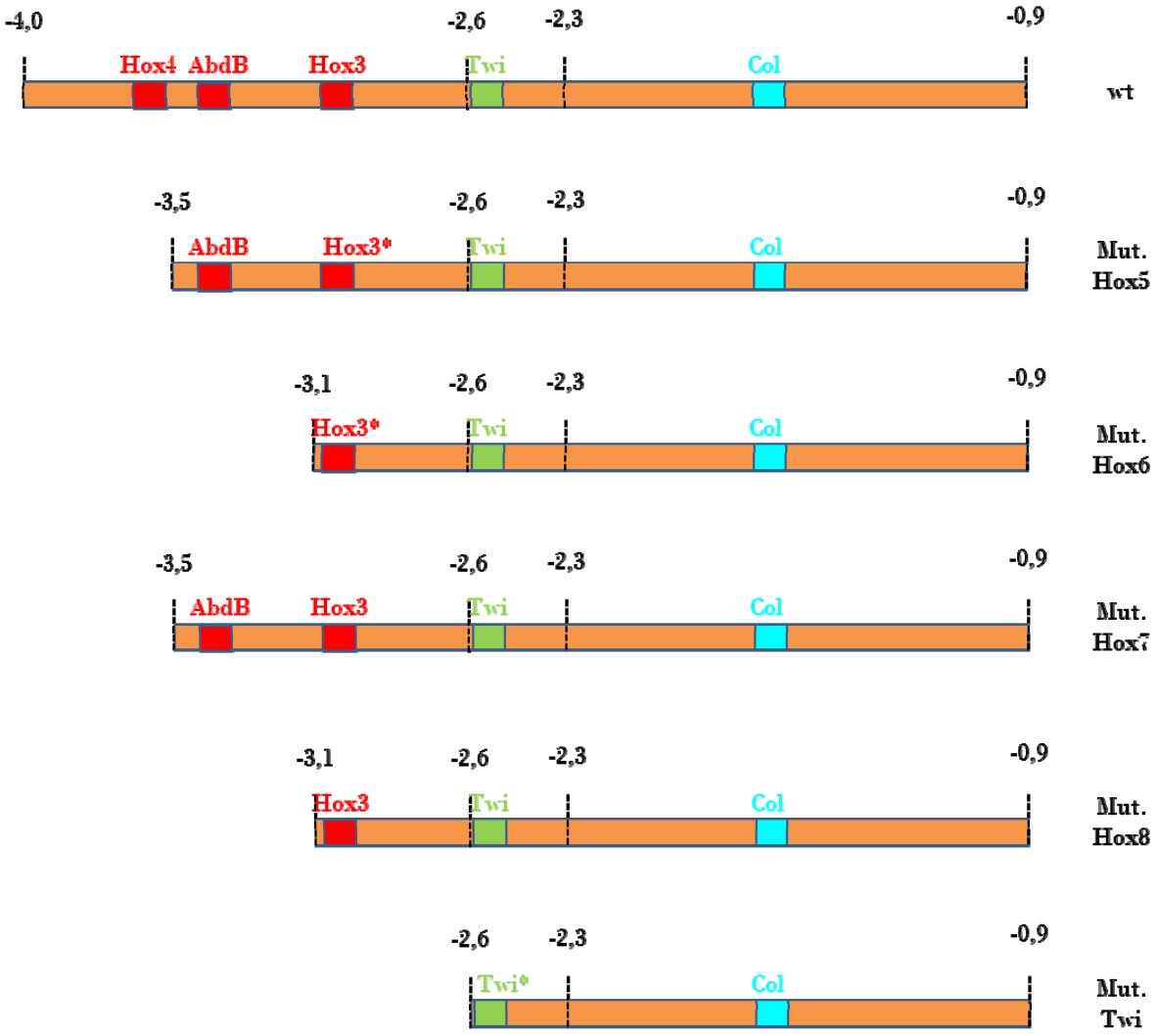
II.1.1 – Dissection du ^{E}CRM de *collier* : quel CRM pour le groupe promusculaire ?

Le ^{E}CRM identifié par H. Boukhatmi (thèse d'université (Boukhatmi, 2013)) correspondait à un fragment de 2.4 kb (appelé CRM276) dirigeant l'expression de *col* dans le groupe promusculaire à l'origine du muscle DA3 et dans les cellules précurseurs de la glande lymphatique. J'ai donc testé l'hypothèse que le CRM276 pouvait contenir 2 éléments de régulation séparés, l'un dirigeant l'expression musculaire de *col*, l'autre l'expression dans la glande lymphatique.

Afin de tester cette hypothèse, j'ai scindé le CRM276 en 2 fragments, en m'appuyant d'une part sur la conservation de la séquence nucléotidique au cours de l'évolution, telle qu'observée en comparant 3 espèces de drosophiles : *D. melanogaster*, *D. pseudoobscura* et *D. virilis*, et d'autre part sur la présence de sites de fixation prédits *in silico* pour les facteurs de transcription Twi et Tin et la position d'un site de fixation *in vivo* de Twi (Sandmann et al., 2007). Le découpage du CRM276 en 2 fragments de 1kb (CRM_A) et 1,4 kb (CRM_B) permet de séparer les 2 blocs de séquences conservées, le fragment immuno-précipité avec Twi étant inclus dans le fragment B. (cf. Fig. R1 et Matériel & méthodes). Les fragments A et B ont été clonés en amont du gène rapporteur *lacZ*. Aucune expression spécifique du transgène CRM_A aux stades précoces du développement embryonnaire n'a pu être détectée. Un faible niveau d'expression de *lacZ* est observé dans la glande lymphatique de quelques embryons, mais uniquement à partir du stade 15 (Fig. R1-B) ; l'activation précoce dans la glande lymphatique observée à partir du stade 14 avec le CRM276 complet n'est donc pas reproduite par le CRM_A. Aucun marquage de la glande lymphatique n'est observé avec le CRM_B. En revanche ce CRM reproduit l'expression musculaire de *col* aux stades groupe promusculaire et progéniteur. Une double immuno-coloration avec des anticorps anti-β-galactosidase et anti-Col montre un recouvrement entre le marquage de la protéine Col (en vert) et celui de la β-galactosidase (en rouge), confirmant que le CRM_B contient les séquences cis-régulatrices nécessaires à l'activation précoce du gène *col* dans le groupe promusculaire à l'origine du muscle DA3 (Fig. R1-C)

Cette étude a donc permis de restreindre le CRM promusculaire (CRM_B = ^{E}CRM, early CRM) à un fragment de 1,4 kb mais n'a pas permis d'identifier le CRM propre à la glande lymphatique au sein du CRM276. La taille du ^{E}CRM est par contre désormais compatible avec une analyse comparée des CRM groupe promusculaire de *eve* (Halfon et al., 2000*)*, *col*, et *tup* (Boukhatmi et al., 2012) pour comprendre les bases moléculaires de l'interprétation de la position au sein du mésoderme (Cf. Discussion).

A



Hox3 : TAATTA ➔ Hox3* : GGGGTA
Twi : CATATG ➔ Twi* : GTTAAC

B

TCTAGAGGATCCCCAATGCTGACACGCTTCTCCAGGTGGCCGAG XbaI(-4 kb)
GAGTGGGAAATGGAAAATGGGCAGGGTTTGCCCCTCGGGGAGCTGCACTAAAACCCCAAA
TGAGATGCATGAATTCCTCTTCGCCGACAAGCAGCTCATAAAAGTTAACATCGAAGTATT
TACAGCATATTTTGGGGAAACTTCAACTTGGAATTTGATTAAATGCAAACTACTTGAAAA
CATACAACTTCAAATCAAGTATGTACAATTTTAAATCAAGTAGATCTCTTTGAATTTATT
GATTTGCTTAACCAACTTAAGTTTCCTTAACTTCTCTTATTTTTGTGTATTTTGGCTAAC
TATTTTTGGCCACAAATCGATCTTTGGTCGCGCCAAAGT**TAATGA**ACCAACCGCACTTTT **Hox4**
CCAATCTCGCATTTGATGCGTTGGATTCCACATATTTGGCCAGACTGAAAAATAATTTCA
TG**TTTATGGCC****ACGTTGTTTTCACCGGCCACAAAAATGCTTATCGAACGCCGGAGATCGC** **AbdB**
CAAGCGAGAGACTTGTAAATATTTATGATTTTTTGCGATTTTTATTCGACTGATAAACGA
GTTTTTCCCATCGCACAGTTGCGATGTGCGAAAATGTTTTGTGGGAAGAAATTGTGCACC
TAGCCCATACGATTCGATTCACCTTCTGTTTGTTGGAATAATAAATCTGTAGAAAGGTTG
TTGTATTTTTAGAAGACTCCATCCCCCTTTGTTATTCTTATGTAAACGAAACCTTTTTCC
TACACACATAGCTGCATAGTAAGTTACCCTCATGAGAACGTTTAGGCCGTGAATATTGGG
TAAATATTTAAACTGGTTGCGATTGTTTTGGCAAAATCGTTGCAGTTGGTCAAACACGTG
TTTGAGATAATT**TAATTA**AATTTACTTGGCTTTTTTAACCCGTGCCACGTGCCACTCGAG **Hox3**
TCCAGTTTGCTGCCCTCGAAATGGAGTGGCAAACACTTGAGCACTTGCTTTAAAAATGCG
GCAAGTGCATTTGGGTTGGGTGGGTAGTTTAGCTTGGTTTTGGCTGGCTGGTGAAAATTC
CTAGGAATCGAAAACTTTGGTGGTCTTCGGACGGCAGACAATGCAACAATGTCATACGTA
CTTGGATGAATCGCTGGCATCTGTCGCAGTCCGTGGAGAATAGACCGAAATTCCCCCGAA
CTCACGAGCATGATGTGGCGATGATGACTTGTGGAAATCAAGCAATGATTTCCACTCCAT

Fig. R2 – Etude du ^{L}CRM : sites Hox et site Twi

A : Constructions pour l'étude de la contribution des sites de liaison des protéines Hox et de la protéine Twist à l'activité du ^{L}CRM de *col*. Tous les fragments ont été placés en amont du gène rapporteur lacZ dans le vecteur attB-inslacZ (Enriquez et al., 2010). **B.** Séquence nucléotidique correspondante au ^{L}CRM [-4/ -0.9], c'est-à-dire au ^{L}CRM entre 0.9 et 4 kb en amont du TSS de *col*, avec les sites de liaison de différentes protéines qui ont été modifiés indiqués en rouge (Hox) et vert (Twi). Les constructions correspondant aux modifications des sites Hox1 et Hox2 sont déjà décrites dans (Enriquez et al., 2010).

II.1.2 – Analyse du ^{L}CRM : Mutations des sites de fixation des facteurs Hox et Twist

Parallèlement à l'étude du ^{E}CRM, j'ai poursuivi la caractérisation du ^{L}CRM. Il avait été préalablement montré que le ^{L}CRM correspond à un fragment de 1 kb (position -2.6 à -1.6 en amont du SIT) responsable de l'expression de *col* du stade PC>FC (Progéniteur>Cellule Fondatrice) à la fin de la myogenèse. Contrairement à un fragment étendu en 5' (fragment -4/ -0.9), un fragment -2.6/-0.9 (contenant le CRM muscle et le CRM tête (Dubois et al., 2007)) ne reproduit pas l'expression musculaire de *col* dans les segments thoraciques T2 et T3 aux stades PC-FC, même si cette expression est observée aux stades plus tardifs de différenciation musculaire. Une étude réalisée par Jonathan Enriquez a montré l'implication des protéines Hox dans cette spécificité segmentaire (Enriquez et al., 2010). Des expériences d'expression pan-mésodermique des protéines Hox ont montré que les protéines Ubx et AbdA agissent sur le fragment -2.6- à 0.9 alors que la protéine Antp agit sur le fragment -4 à -0.9 (Enriquez et al., 2010). La liaison des protéines Ubx et AbdA à un site conservé dans le fragment -2.6/-0.9 (site appelé *hox*^{2}) est requis pour l'activation du ^{L}CRM à la transition PC>FC dans les segments abdominaux. Cette dernière observation a révélé un mécanisme clé de relais entre les ^{E}CRM et ^{L}CRM, dépendant de l'activité des protéines Hox. Cependant, la question restait posée du mécanisme de contrôle de l'expression de *col* dans les muscles thoraciques.

J'ai donc entrepris de tester l'existence d'un site fonctionnel de fixation d'Antp sur le fragment -4/-2.6 en réalisant des mutations/délétions ciblées des sites prédits de liaison de protéines Hox conservés à la même position dans plusieurs espèces de drosophile. J'ai considéré 2 sites potentiels, *hox*^{3} et *hox*^{4}, et un site prédit de fixation d'AbdB que j'ai mutés ou enlevés individuellement ou en combinaison (transgènes *hox*$^{5\text{-}8}$) (cf. Matériel et méthodes et Fig. R2). Malheureusement, l'analyse de ces transgènes a été largement infructueuse, puisqu'aucune des combinaisons de mutations testées n'a induit la perte d'expression du transgène dans les segments thoraciques, comme attendu si Antp se liait directement au fragment -4 /-2.6. Antp pourrait donc agir sur le ^{L}CRM soit de manière indirecte, soit au niveau d'un ou de plusieurs autres site(s) qu'il reste à identifier.

Avec une même logique, j'ai décidé de tester en parallèle la fonction d'un site Twist (Twi) prédit *in silico* et conservé au cours de l'évolution dans le ^{L}CRM, et confirmé *in vivo* par des données de ChIP-on-chip (Sandmann et al., 2007)(cf. Fig.R2). La conjonction de ces données suggérait fortement un rôle de ce site dans la régulation transcriptionnelle de *col* dans le muscle DA3. J'ai donc construit une lignée rapporteur 2.6_0.9 (*twi**)-lacZins (cf. Matériel et méthodes) dans laquelle ce site Twi a été muté. La mutation ciblée de ce site n'affecte pas significativement

l'expression du transgène au stade PC>FC DA3. Notre interprétation est que Twi pourrait conférer de la robustesse à l'expression de Col (par exemple lors du mécanisme de relais entre ^{E}CRM et ^{L}CRM), mais que l'absence de fixation directe de Twi peut être soit compensée par d'autres facteurs fonctionnellement redondants, soit participer d'un phénomène de recrutement collectif tel que proposé par le laboratoire d' E. Furlong (Junion et al., 2012), le site de fixation de Twi facilitant mais n'étant pas déterminant pour sa liaison.

Une étude plus approfondie et quantitative de l'expression de ces transgènes (mutations Hox et Twi) est cependant envisagée en utilisant un transgène rapporteur comportant un intron afin de comparer directement les transcrits primaires transgénique et endogènes et d'explorer un éventuel effet de la mutation de ces sites sur la fenêtre temporelle ou le niveau de transcription de *col*.

La difficulté d'interprétation des données d'analyse du ^{L}CRM par mutation ciblée de sites de fixation pourrait donc refléter une redondance des éléments cis- et trans-activateurs assurant la robustesse de l'expression de *col* dans le muscle DA3, un phénomène probablement artificiellement amplifié par l'utilisation de CRMs dits « insulés » pour mon analyse. En effet, l'isolement du CRM de son contexte génomique et notamment de l'influence de potentiels répresseurs pourrait atténuer l'importance de la fixation des facteurs de transcription testés. Par ailleurs les données de fixation de Twi *in vivo* montrent que ce facteur est déjà positionné sur le ^{L}CRM alors que celui-ci est encore inactif (entre 4 et 6 heures de développement, stades où le ^{E}CRM est alors actif). Plus qu'un rôle d'activateur de la transcription, Twi pourrait alors jouer un rôle de facteur « pionnier » sur le ^{L}CRM en promouvant par exemple le recrutement d'autres facteurs ou de la machinerie de transcription. Et ce rôle est plus difficilement analysable en contexte de gène rapporteur insulé.

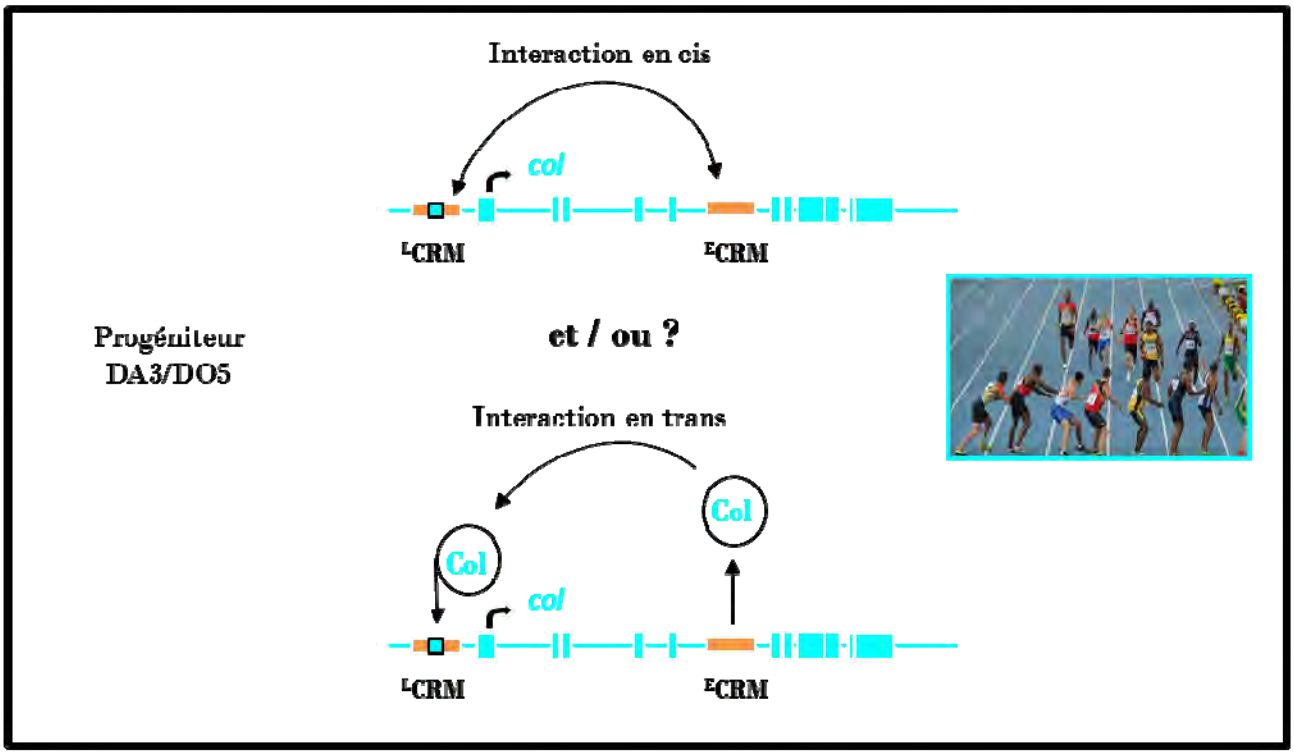
L'autorégulation de *col* dans le lignage DA3 via sa fixation sur le ^{L}CRM -2.6/-1.6 (Dubois et al., 2007) pose la question du mécanisme de « relais » entre les ^{E}CRM et ^{L}CRM. Afin d'aborder l'étude de ce mécanisme, une nouvelle approche a donc été envisagée afin de tester de le rôle de ces CRM dans un contexte génomique reconstitué.

II.1.3 – Fosmide (pFlyFos) pour l'analyse des CRM de *collier* dans leur contexte génomique

a) pFlyFos : un contexte chromosomique reconstitué

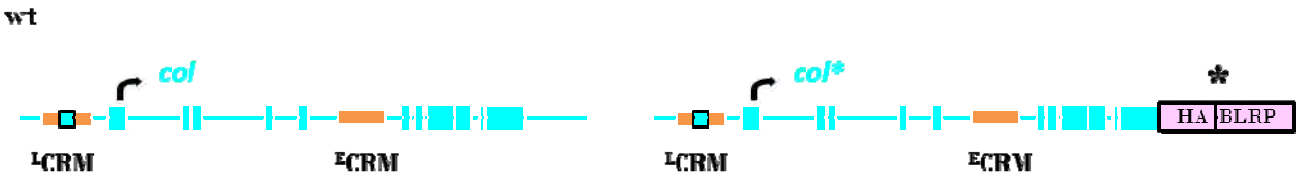
Les hypothèses testées grâce à des modifications du pFlyFos *col**

A



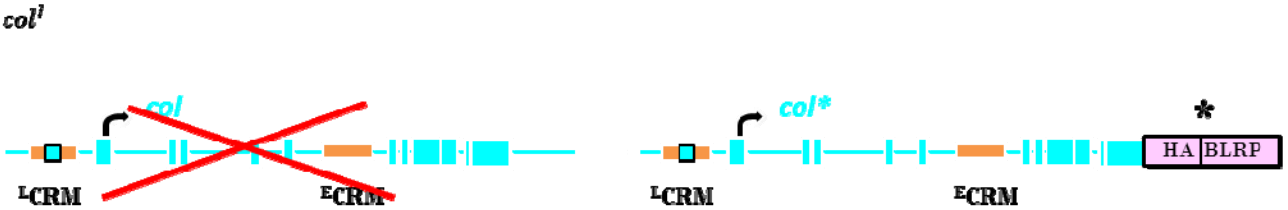
<u>*Copie col* endogène</u> <u>Copie pFlyFos *col*</u>

B



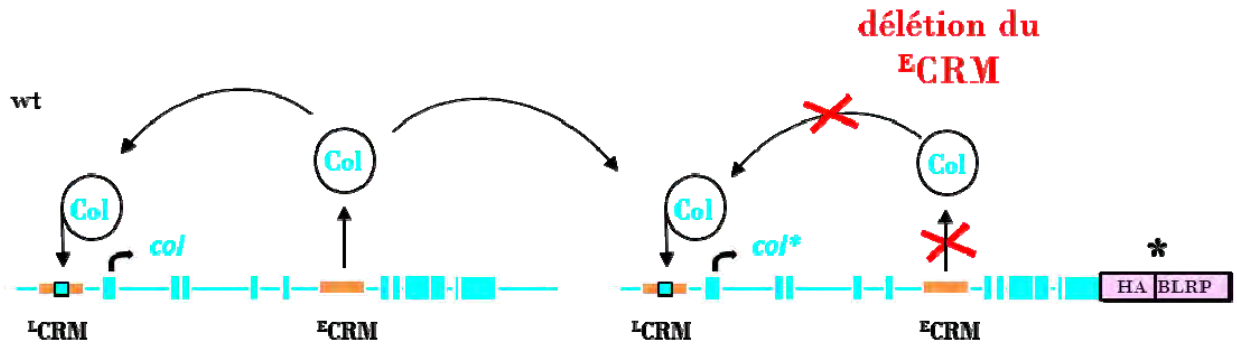
➔ **Contrôle :** analyse du patron d'expression de la protéine Col taggée (Col*-HA.BLRP) en comparaison de Col endogène = **Read-out 1**

C



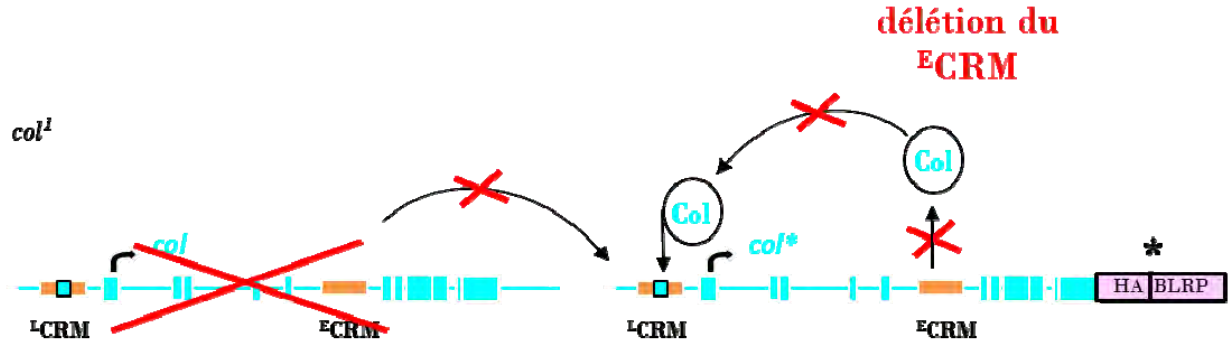
➔ **Contrôle :** Test de sauvetage phénotypique du mutant *col*^{1} par la copie pFlyFos *col*-*ha.blrp*, phénotype musculaire (embryon) et adulte = **Read-out 2**

D



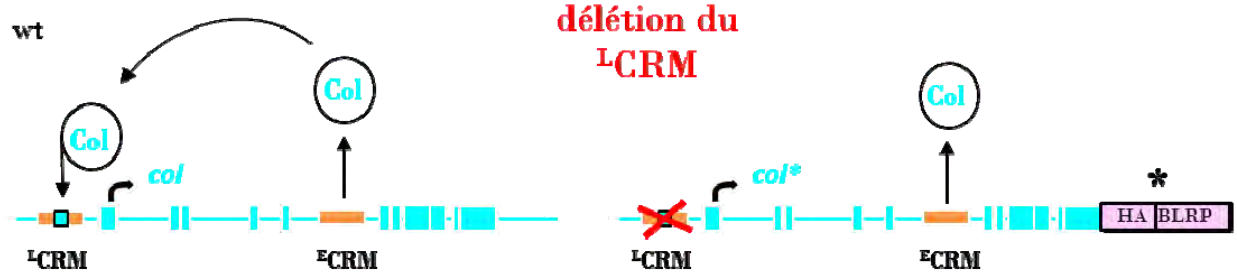
➔ **Question :** Etude de l'interaction en cis du $^{E-}$ et du ^{L}CRM. **Read-out 1**

E



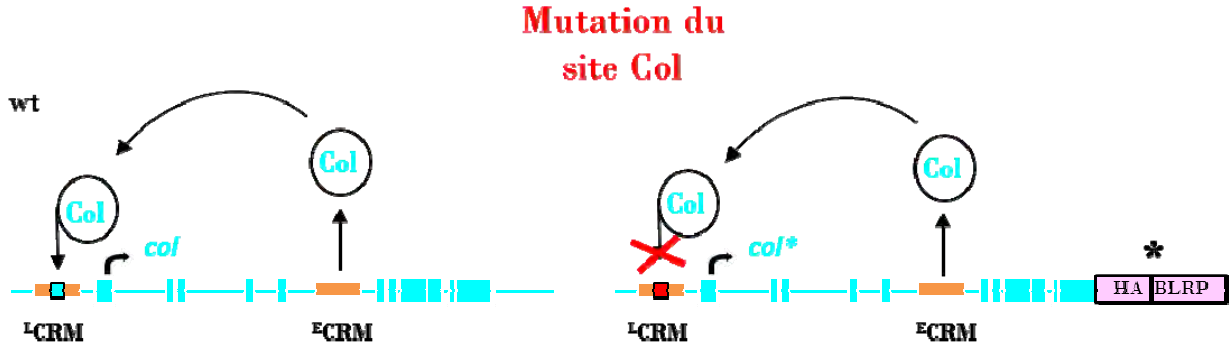
➔ **Question :** Etude de la contribution en trans du ^{E}CRM à l'expression de *col*. **Read-out 1 et 2**

F



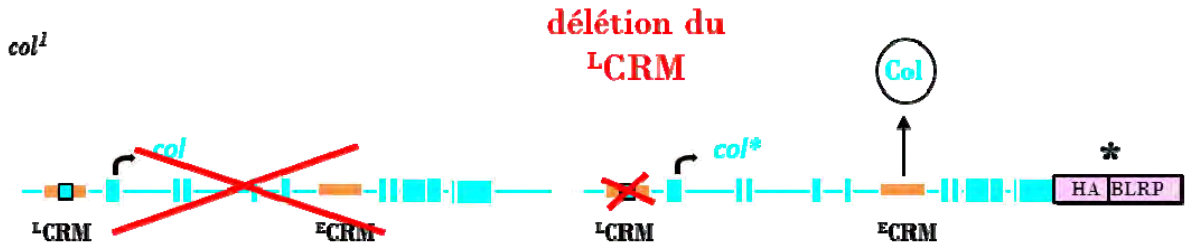
➔ **Contrôle :** Etude de la contribution du ^{L}CRM à l'expression de *col*. **Read-out 1**

G



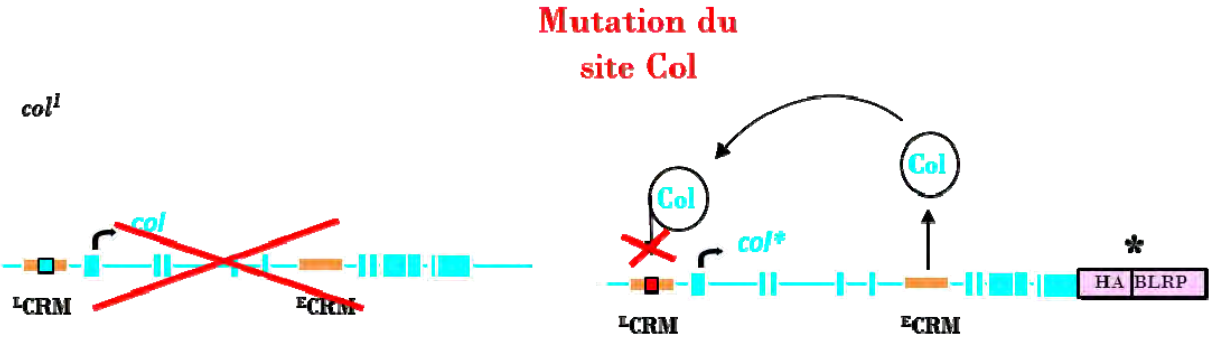
➔ **Question :** Etude de la contribution de l'auto-régulation directe de *col*. **Read-out 1**

H



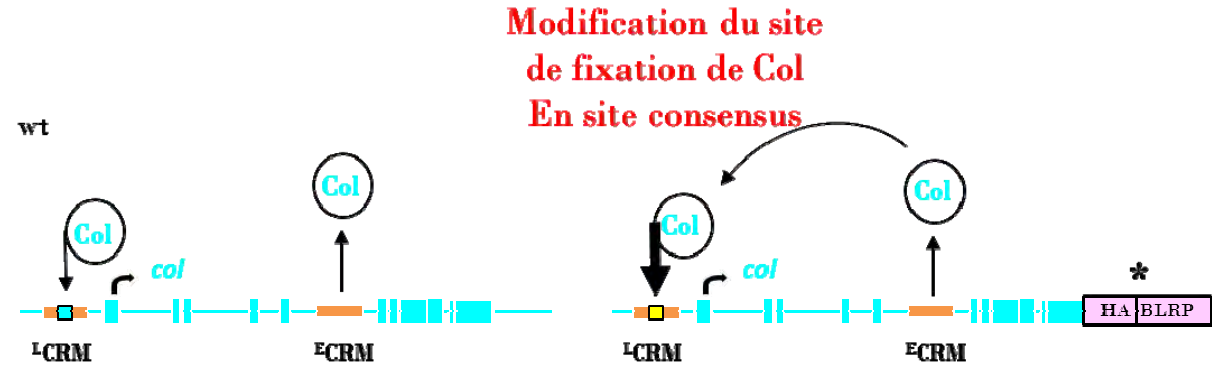
➔ **Contrôle : Etude du phénotype musculaire associé à la perte de l'expression tardive de *col*. Read-out 2**

I



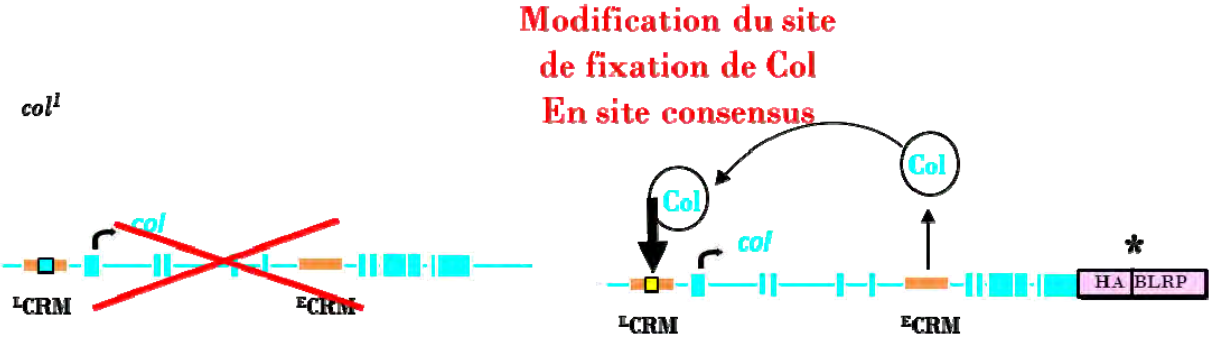
➔ **Etude du phénotype musculaire associé à la perte de la fixation tardive de Col sur son propre CRM. Read-out 2**

J



➔ **Etude du rôle de l'auto-régulation directe de *col* : renforcement du site de fixation. Read-out 1**

K



➔ **Etude du rôle de l'auto-régulation directe de *col ;* renforcement du site de fixation. Read-out 2 (phénotype du muscle DA2 en particulier)**

Fig. R3 – Hypothèses testées grâce à l'utilisation du pFlyFos Col*

L'activité des transgènes ^{E}CRM et ^{L}CRM suggère que le stade progéniteur est le stade clé durant lequel a lieu le « passage de témoin » entre ces 2 CRM permettant de « fixer » l'identité du muscle DA3 à travers le maintien de l'expression de *col* dans son progéniteur. L'activité du ^{L}CRM est soumise à autorégulation directe : Col se lie à un motif de reconnaissance situé 1.7 kb en amont du SIT (cf. Fig. R2), au sein du ^{L}CRM, et la mutation de ce site entraîne la perte d'activité du ^{L}CRM dans le muscle DA3 (Dubois et al., 2007). Bien que chacun des deux CRM puisse fonctionner indépendamment dans un test de transgène rapporteur, la question restait posée du mécanisme de passage de relais entre les deux CRM dans la cellule progéniteur du muscle DA3 *in vivo*.

Deux hypothèses non mutuellement exclusives pouvaient être envisagées (cf. Fig. R3) :

(i) Un mécanisme en *cis* c'est-à-dire une communication directe entre les deux CRM, via un repliement de la chromatine facilitant « l'ouverture » et la liaison directe de Col sur le ^{L}CRM.

(ii) Un mécanisme en *trans*, l'activité du ^{E}CRM étant requise pour l'accumulation de la protéine Col à une concentration suffisante pour sa liaison efficace sur le ^{L}CRM.

Si l'utilisation de gènes rapporteurs permet d'identifier et de caractériser les CRM (cf. ci-dessus), il ne s'agit que d'une analyse approximative puisque les CRM sont testés séparément, indépendamment du contexte chromosomique dans lequel se situe le gène, et donc néglige le rôle de séquences « insulatrices » et les interactions à distance possibles entre CRM. C'est pourquoi il est important de compléter les études réalisées avec des gènes rapporteurs par des modifications de CRM dans un contexte chromosomique normal ou du moins reconstitué, c'est-à-dire conservant l'environnement génomique du gène étudié.

L'utilisation de transgènes basés sur des fosmides contenant de grands fragments d'ADN génomique permet ce type de modifications. Leur structure permet en effet de modifier un gène par recombinaison dans la bactérie (cf. fig. R4 et Matériel et méthodes) puis d'intégrer le gène modifié à un site spécifique dans le génome de la drosophile.

La taille moyenne des fragments d'ADN génomique contenu dans un fosmide pFlyFos étant de 36 kb, la collection génomique de pFlyFos générée par (Ejsmont et al., 2009) contient un très grand nombre de gènes complets, y compris leurs régions régulatrices. Dans le cas du pFlyFos #022589, il contient l'intégralité de la région transcrite de *col*, ainsi que 10.5 kb en amont et 3.8 kb en aval.

Pour déterminer quel mécanisme assure le passage de témoin entre les ^{E}CRM et ^{L}CRM, j'ai entrepris de modifier chacun de ces CRM au sein du pFlyFos *col* #022589 (Fig.R3). Ces

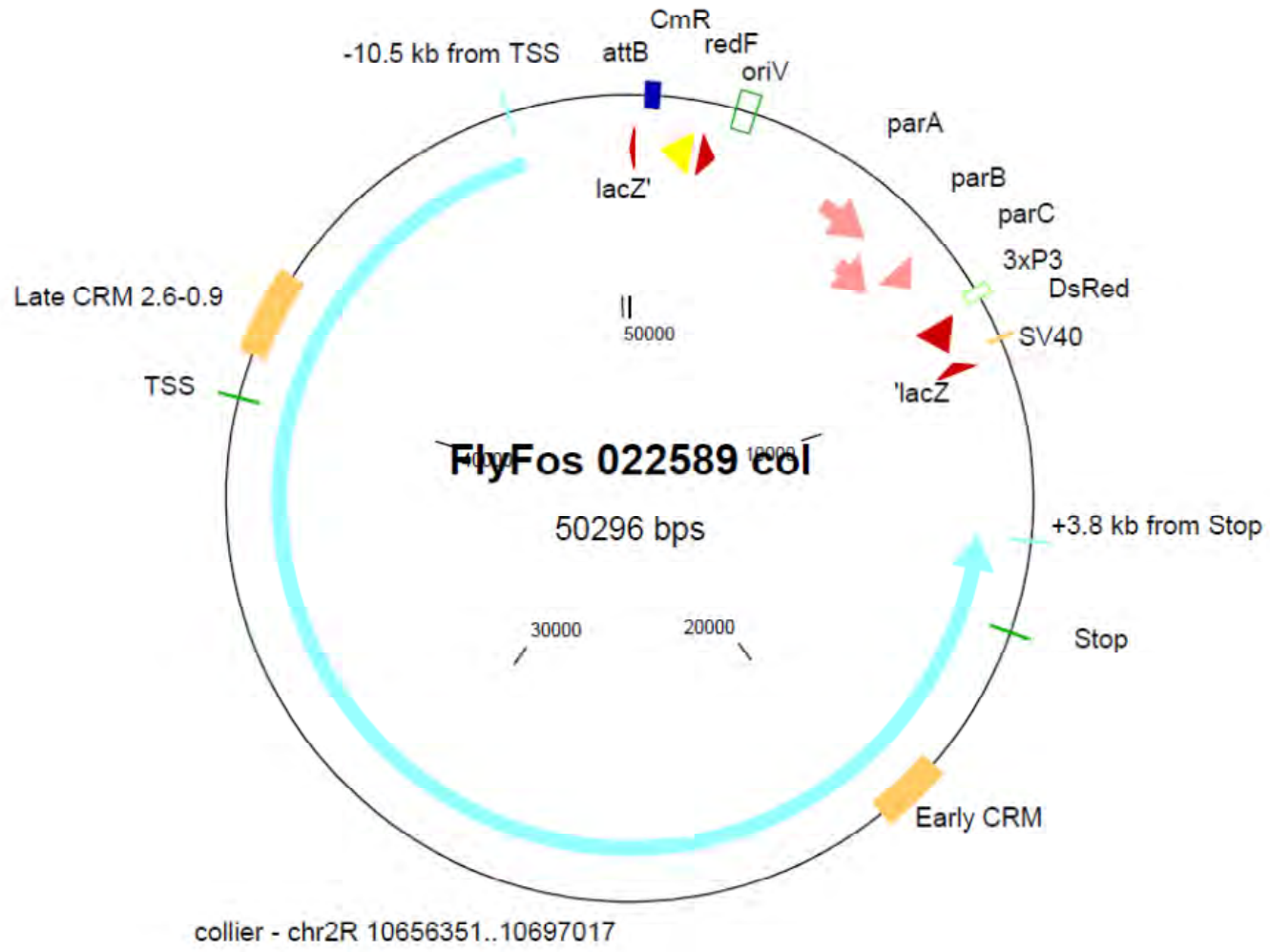


Fig. R4 – Carte du fosmide pFlyFos 022589 *col* contenant la région transcrite de *col*

Le fosmide pFlyFos 022589 *col* contient l'intégralité de la région transcrite de *col*, 10.5 kb en amont et 3.8 kb en aval. Les 2 CRM contrôlant l'expression de *col* au cours de la formation du muscle DA3 (Early et Late CRM) sont inclus. Le pFlyFos *col* porte un site AttB permettant son intégration dans le génome de la drosophile par recombinaison site-spécifique AttP-AttB, le gène dsRed sous le contrôle d'un promoteur actif dans l'œil de drosophile (3xP3) pour la sélection des intégrations dans la drosophile, un gène de résistance au chloramphénicol pour la sélection et la maintenance du fosmide dans la bactérie et une origine de réplication OriV inductible à l'arabinose pour permettre sa réplication à un haut niveau de copie dans E.*coli*.

expériences sont en cours et je décris ici la stratégie choisie et les résultats attendus. Afin de pouvoir distinguer la copie transgénique de la copie *col* endogène, la première modification à réaliser est d'étiqueter la protéine du pFlyFos (Col*) en position C-terminale (cf. Matériel & méthodes). Nous avons opté pour une double étiquette : HA pour la détection par immuno-coloration et BLRP (Biotin Ligase Recognition Peptide) (Col*-HA.BLRP). Cette deuxième étiquette a pour objectif de permettre l'identification de gènes cibles directes de Col tissu-spécifiques par chromatographie d'affinité (voir ci-dessous II.2.1).

Dans un premier temps, la fonctionnalité du pFlyFos *col** doit être vérifiée en utilisant 2 critères (read-out) : 1) l'expression de la protéine Col* (Read-out 1), comparée à Col endogène et 2) le sauvetage du phénotypique musculaire embryonnaire du mutant *col*^{1} (Read-out 2) (Fig.R3, B et C). Ces deux *read-out* seront ensuite utilisés pour déterminer le rôle de chacun des 2 CRM musculaires. La délétion du ^{E}CRM de *col** permettra, dans un contexte sauvage, de tester le rôle possible de l'interaction des 2 CRM en *cis* (read out-1), et dans un contexte *col*^{1}, la contribution en *trans* du ^{E}CRM à l'activation du ^{L}CRM (read-out 1 et 2) (Fig.R3, D et E). La délétion du ^{E}CRM en contexte mutant permettra par ailleurs de préciser le phénotype musculaire associé à la perte précoce de l'expression de Col (Fig. R3, H), une question toujours en suspens.

La délétion du ^{L}CRM en contexte sauvage (Fig.R3, F) ou mutant (Fig.R3, H) servira ensuite de contrôle pour déterminer précisément le rôle de l'autorégulation et l'importance du site de liaison de Col dans le ^{L}CRM, un site significativement divergent du site consensus (Fig. R3, G, I, J, K). La délétion du ^{L}CRM en contexte mutant permettra par ailleurs de préciser le phénotype musculaire associé à la perte tardive de l'expression de Col (Fig. R3, H).

b) Etapes préliminaires : la copie pFlyFos col est fonctionnelle.

Avant de modifier le pFlyFos *col** #022589, j'ai voulu m'assurer qu'il était fonctionnel, c'est-à-dire qu'il permettait de sauver la létalité embryonnaire associée à la mutation du gène *col* (mutant *col*^{1} (Crozatier et al., 1999)). Le pFlyFos *col* * a donc été intégré sans modification dans le génome de la drosophile, au niveau de la plateforme AttP2 (chromosome III), site d'insertion de toutes les constructions prévues. Les résultats montrent qu'une copie du pFlyFos *col** sauve complètement la létalité embryonnaire due à la mutation *col*^{1}. Elle atténue aussi fortement le phénotype observé dans les ailes adultes (perte de la région centrale de l'aile et de la nervure 4). Ce phénotype, qui reflète l'expression de Col en réponse à Hedgehog dans le disque d'aile, est observé lorsque la létalité embryonnaire de *col*^{1} est sauvée par expression d'un ADNc sous le contrôle de régions

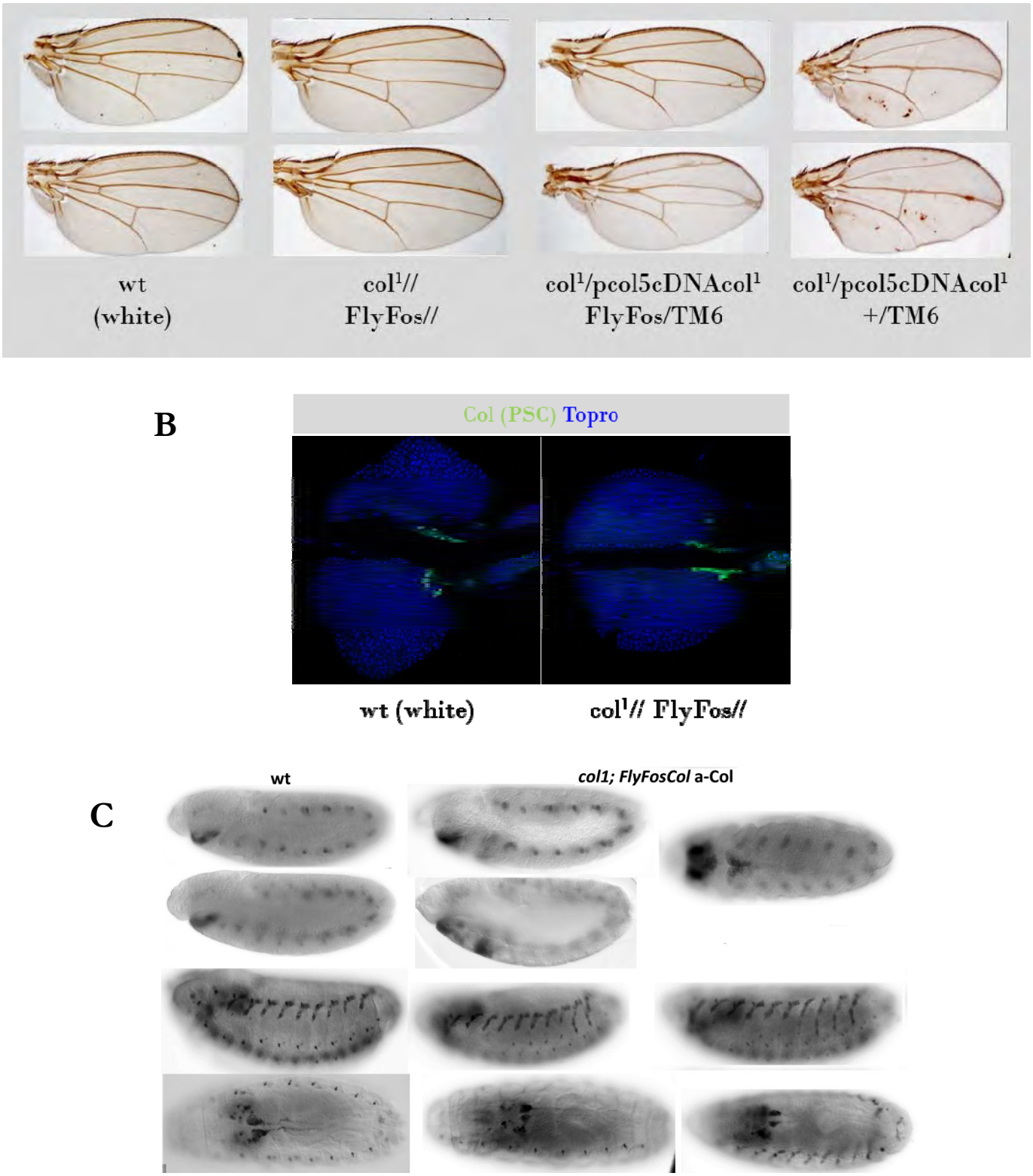


Fig. R5 – le pFlyFos *col* permet de sauver le phénotype mutant *col*^{1}

A : Sauvetage du phénotype d'aile induit par le mutant *col*^{1} dans l'aile. Une seule copie du pFlyFos (3^{e} colonne) permet de réduire considérablement le phénotype observé dans un mutant collier (4^{e} colonne) et deux copies (2^{e} colonne) conduisent à un phénotype sauvage (1ère colonne). **B** : La protéine Col produite à partir du pFlyFos (droite) présente un patron similaire à la protéine Col endogène (gauche) dans la glande lymphatique larvaire ainsi que durant toute l'embryogenèse (**C**) et n'induit pas de phénotype particulier.

régulatrices actives dans l'embryon (*pcol5cDNA)* (Vervoort et al., 1999) (Fig. R5). Deux copies du pFlyFos *col** sauvent complètement ce phénotype d'aile. Un seul CRM de *col* actif dans le disque d'aile avait jusqu'à présent été décrit et cartographié, à environ 15 kb en amont du SIT de transcription de *col* (Hersh and Carroll, 2005), en dehors de la région clonée dans le pFlyFos *col** #022589. Le sauvetage observé suggère donc qu'il existe au moins un autre élément du même type (« shadow enhancer » ?) au sein du pFlyFos *col** #022589 qu'il reste à localiser (cf. Discussion). Un immuno-marquage de la protéine Col* en mutant col^{1} montre un patron similaire à la protéine Col endogène dans l'embryon ainsi que dans la glande lymphatique larvaire (Fig. R5). Aucun phénotype particulier n'est observé ni sur l'embryon, ni sur la larve, ni sur l'adulte lorsque le *pFlyFos col* est surnuméraire ou, à l'inverse, lorsqu'il remplace la copie endogène (col^{1}*// ; pFlyFos col//*). Le pFlyFos *col* #022589 contient donc un gène *col* fonctionnel et peut donc être utilisé pour les expériences de modifications des CRM.

*c) Etiquetage de la protéine Col**

L'étape suivante a consisté à étiqueter la protéine Col dans le pFlyFos (Col*-HA.BLRP). Il existe deux isoformes de Col issues d'un évènement d'épissage alternatif et qui semblent actives d'un point de régulation de la transcription, du moins dans le cas de la régulation de l'expression de *hedgehog* dans le segment intercalaire de la tête (Ntini and Wimmer, 2011). J'ai choisi de positionner l'étiquette HA.BLRP en position C-terminale de l'isoforme B de la protéine Col, l'isoforme précédemment utilisée pour les expériences de sauvetage de la létalité embryonnaire induite par la mutation col^{1} (Crozatier and Vincent, 1999). L'étiquette a été placée immédiatement en amont du codon stop, afin d'éviter des recombinaisons induites par interférence de sites FRT lors des recombinaisons suivantes (cf. Matériel & méthodes).

Une fois l'étiquetage validé par séquençage de l'ADN recombinant, le *pFlyFos col*-HA.BLRP* a été inséré sur le chromosome III du génome de drosophile. J'ai vérifié qu'il permet de sauver la létalité embryonnaire induite par la mutation col^{1} suggérant la formation d'une protéine fonctionnelle… mais qui s'est avérée non détectable ! (voir Discussion).

II.2 – Recherche des cibles directes du facteur de transcription Collier dans le muscle DA3

Une question majeure concernant le contrôle de l'identité musculaire reste la nature des effecteurs de cette identité. En effet, s'il est établi que l'identité de chaque muscle -sa forme, sa taille, son orientation, ses sites d'attachement à l'épiderme-, est dictée par la combinatoire de facteurs de transcription dit identitaires exprimés au niveau de sa cellule fondatrice, les gènes cibles de ces facteurs restent à identifier. Le cœur de mon travail de thèse a consisté à développer une stratégie d'identification des gènes cibles de Collier dans le muscle DA3.

II.2.1 – Objectif initial : Recherche des cibles directes de Collier par une stratégie d'immunoprécipitation tissu-spécifique : « ChAPseq »

Au vu de la complexité des sites d'expression embryonnaire de Col, mon premier choix était une recherche tissu-spécifique des cibles de Col, centrée sur le muscle DA3. La stratégie choisie visait à exprimer une protéine Col étiquetée BLRP (Biotin Ligase Recognition Peptide) sous le contrôle de l'ensemble des séquences cis-régulatrices de *col*. La biotinylation de Col^{BLRP} par expression ciblée de la Biotin ligase bactérienne BirA (Beckett et al., 1999; Rodriguez et al., 2006) spécifiquement dans le mésoderme (pilote Mef2-gal4xUAS-BirA) ou le muscle DA3 (pilote pcol85-gal4) devait permettre ensuite la sélection par chromatographie d'affinité (ChAP : **Ch**romatin **A**ffinity **P**urification) des complexes ADN–Col spécifiquement formés dans ces tissus. La méthode a été d'abord validée en culture cellulaire par J. Oyallon, alors post-doc dans l'équipe. L'objectif était théoriquement (voir ci-dessous) de croiser ces données de ChAP avec les données issues de l'étude à grande échelle des éléments cis-régulateurs chez la drosophile menée par le consortium ModEncode (Nègre et al., 2011). Cette étude réalisée sur une collection d'embryons entre 0 et 12 heures de développement comprenait en effet les résultats de ChIPseq réalisés avec un anticorps anti-Col. Le croisement de ces données avec nos données de ChAP devait permettre de préciser quelles cibles directes de Collier sont DA3-spécifiques.

Pour mettre en œuvre la stratégie de ChAP, j'ai d'abord choisi de travailler avec un BAC de type P[acman] (Venken et al., 2009), contenant environ 80 kb de région génomique de drosophile. Le P[acman] CH321 – 64I16 inclut le gène *collier*, 30 kb de séquences en amont et 20 kb en aval. Malheureusement, il s'est avéré que ce BAC ne permet pas de sauvetage phénotypique

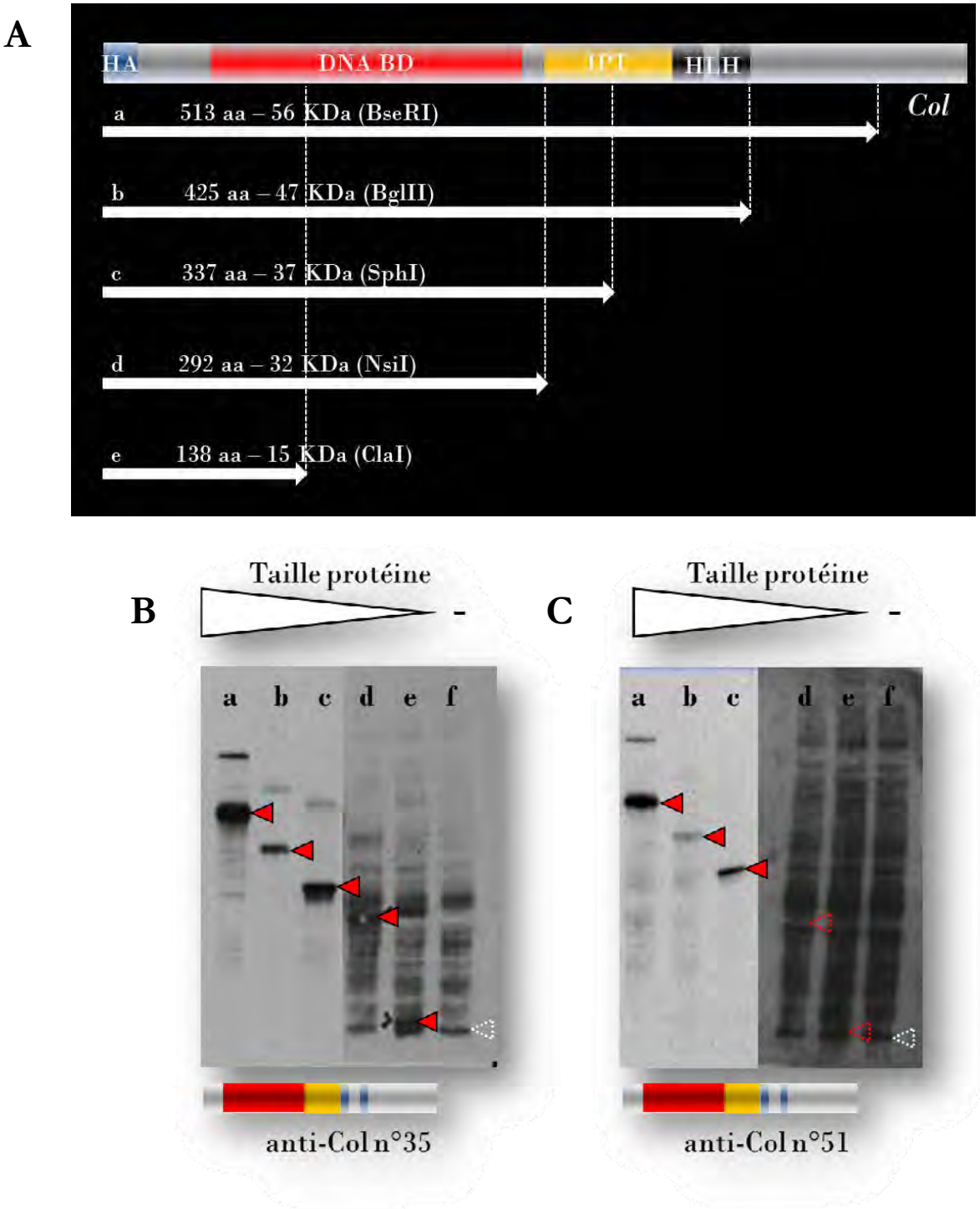


Fig. R6 - Domaines reconnus par différents anticorps monoclonaux anti-Collier.

A. Schéma de la protéine Collier avec les différents domaines (DNA BD : Domaine de fixation à l'ADN ; IPT : Immunoglobulin-Plexin-Transcription factor ; HLH : Helix-Loop-Helix). Les flèches blanches indiquent la fraction de protéine Col résiduelle après troncature (A à E). Les tailles des protéines tronquées et les sites enzymatiques utilisés (entre parenthèses) sont indiqués au-dessus des flèches. **B.C.** Western-blots réalisés avec les anticorps monoclonaux anti-Col n°35 (1er groupe) (B) et n°51 (2e groupe) (C). La piste f ne comporte pas de protéine Col. Les pointes rouges pleines indiquent les protéines Col reconnues par l'anticorps, les pointes en pointillés les protéines non reconnues. (*nota* : d, e, f : surexposition)

du mutant *col*^{1}. Des expériences complémentaires m'ont amenée à conclure que la protéine Collier n'est pas synthétisée à partir de ce transgène, d'où le choix récent du pFlyFos *col**.

Cette déconvenue m'a conduit à revenir à une stratégie « classique » de ChIPseq à partir d'extraits nucléaires d'embryons entiers (cf. II.2.2). La première option envisagée était d'utiliser les mêmes anticorps que le consortium ModEncode (Nègre et al., 2011) en concentrant mon analyse sur des embryons aux stades 13-14, stades d'initiation de la phase de « réalisation » de l'identité musculaire. Cependant, il s'est avéré que les données ModEncode obtenues pour Collier méritent d'être considérées avec extrême précaution (voir Discussion)…

Suite à cette double déconvenue, nous avons donc décidé :

(i) D'opter pour un fosmide de type « pFlyFos » pour la construction des outils nécessaires au ChAPseq (en cours de validation, cf. ci-dessus).

(ii) D'utiliser des anticorps monoclonaux générés et validés par notre laboratoire pour des expériences de ChIPseq. Cette stratégie et les résultats obtenus sont détaillés dans les chapitres suivants.

II.2.2 – Identification de gènes cibles directs de Collier à partir d'embryons entiers par des expériences de ChIPseq

a) Caractérisation et sélection d'anticorps monoclonaux anti-Collier pour les expériences de ChIP (avec Yannick Carrier, AI-CNRS)

La validité des expériences de ChIP dépend de la qualité, et notamment de la spécificité, des anticorps utilisés pour immuno-précipiter la chromatine.

Nous avions à notre disposition 5 lignées d'anticorps monoclonaux de souris différents dirigés contre la protéine Collier, sélectionnés par tests Elisa puis immuno-marquages d'embryons. Pour nos expériences de ChIPseq, nous avons opté pour l'utilisation d'un mélange d'anticorps reconnaissant différents épitopes de la protéine Collier.

Afin de localiser l'épitope de la protéine Collier reconnu par chacun des anticorps monoclonaux disponibles, nous avons réalisé des expériences de Western-blot en utilisant des protéines Collier pleine taille (62 kDa) ou tronquées (56 kDa, 47 kDa, 37 kDa, 32 kDa et 15 kDa – cf. Fig.R6) par délétions séquentielles de la partie C-terminale de la protéine. Les résultats ont

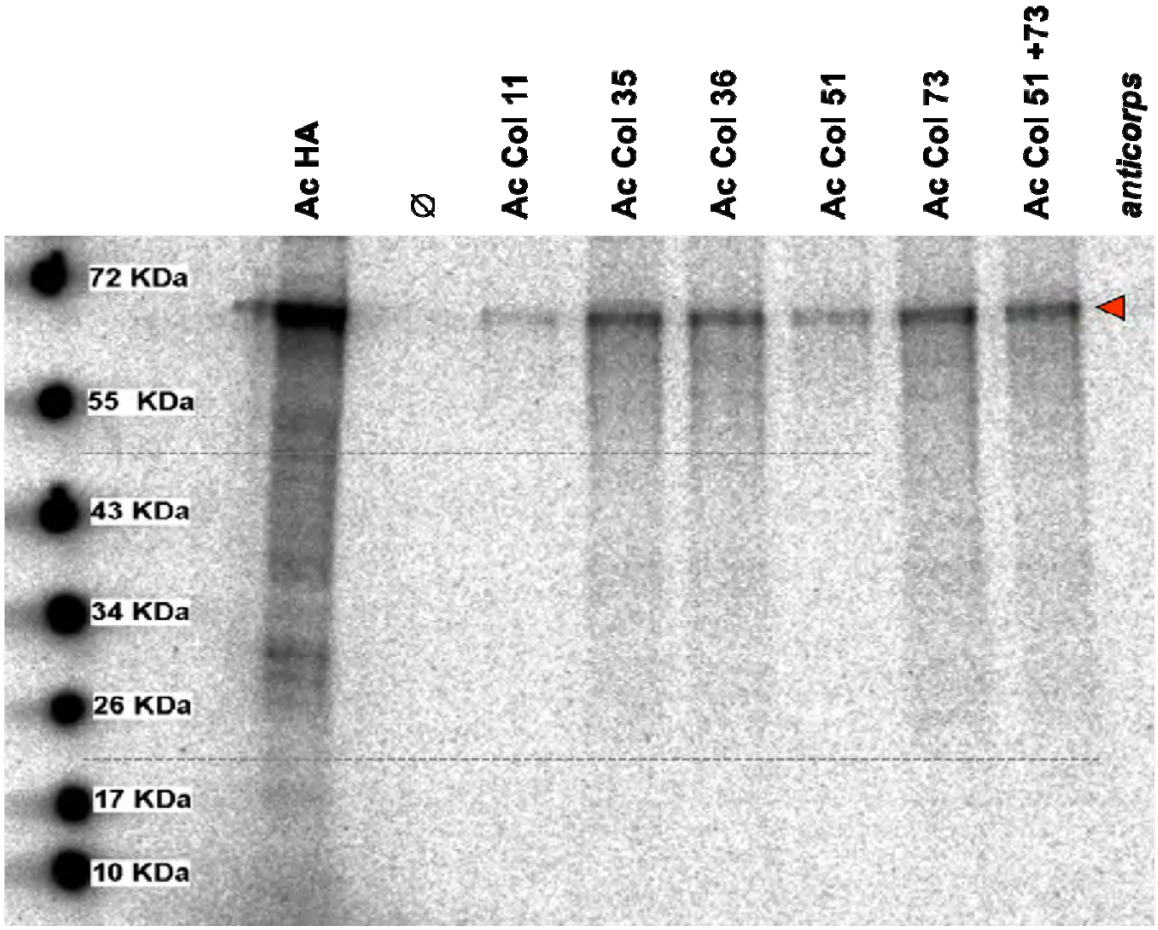


Fig. R7 – Les anticorps anti-Col immuno-précipitent la protéine Col synthétisée *in vitro*.

Test d'immuno-précipitation (IP) de la protéine HA::Col* (Met[35S]) synthétisée *in vitro* par différents anticorps anti-Col indiqués au-dessus de chaque piste. L'IP contrôle réalisée avec l'anticorps anti-HA permet de visualiser la protéine HA::Col pleine taille synthétisée (forme majoritaire) ainsi que l'ensemble des formes tronquées correspondant à des arrêts de traduction *in vitro*. Chacun des 5 anticorps anti-Col permet d'immuno-précipiter la protéine Col (pointe rouge), avec une meilleure efficacité pour les anticorps 35, 36 et 73. L'observation de la limite de détection des formes tronquées (lignes pointillées) confirme que les anticorps 35, 36 et 73 reconnaissent la partie N-terminale de la protéine (environ 20 kDa) tandis que les anticorps 11 et 51 reconnaissent un domaine plus en aval de la protéine (à priori l'IPT ou le domaine HLH).

mis en évidence 2 groupes d'anticorps : un premier groupe composé des anticorps 35, 36 et 73 réagissant avec toutes les formes tronquées de la protéine Collier et localisant ainsi l'épitope reconnu dans 15 kDa en position N-terminale de la protéine ; un second groupe composé des anticorps 11 et 51, reconnaissant spécifiquement le domaine IPT de Col (Fig.R6). Ce groupement est en accord avec une première étude réalisée au laboratoire par Virginie Daburon (diplôme d'ingénieur du CNAM) qui avait montré que les anticorps 35, 36 et 73 sont de type IgG1 et les anticorps 11 et 51 de type IgG2b.

Nous avons ensuite testé la capacité de ces anticorps à immuno-précipiter la protéine Collier. Des tests réalisés avec des protéines Collier synthétisées *in vitro* (Fig. R7) et des extraits nucléaires d'embryons entiers (Fig. R8) montrent que chacun des 5 anticorps étudiés est capable de fixer et retenir la protéine Collier, avec cependant plus d'efficacité pour les anticorps du premier groupe (épitope N-terminal) : environ 10% de la quantité initiale de protéine Col recombinante est retenue par ces anticorps contre 2% pour les anticorps du second groupe. Un test d'immuno-précipitation sur extraits nucléaires d'embryons non fixés a ensuite permis de confirmer que les anticorps du premier groupe ont une meilleure affinité pour la protéine Col. Le niveau d'expression de la protéine Collier dans l'embryon ne permet pas de la détecter directement en Western blot à partir d'un extrait total (ou très faiblement), mais elle est détectée après enrichissement par immuno-précipitation avec un anticorps du premier groupe. On peut noter un enrichissement légèrement supérieur lorsqu'on associe des anticorps du premier et second groupe (51 et 73) (Fig.R8). Un mélange d'anticorps des deux groupes permet donc de diminuer le seuil de réaction avec Col. Nous avons donc choisi de travailler avec un mélange des anticorps 51 et 73 pour les expériences de ChIP. Pour les expériences contrôle, un anticorps monoclonal de souris dirigé contre le peptide HA a été choisi.

b) Préparation de la chromatine, immuno-précipitation et tests de validité

Mon objectif étant d'identifier les cibles de Col au cours de la myogenèse, nous avons travaillé avec une population d'embryons majoritairement aux stades 13-14 (cf. Fig. R9-A), stades du développement durant lesquels les FCM commencent à fusionner avec les FC pour former la fibre musculaire, stades correspondant à l'initiation de la phase de « réalisation » de l'identité musculaire. Il est à noter qu'à ce stade Collier s'exprime également, et majoritairement, dans le système nerveux central de l'embryon (environ 50 cellules par hémisegment au stade 14 (Demilly et al., 2011), contre 2 noyaux en moyenne par hémisegment pour le muscle). Collier est également

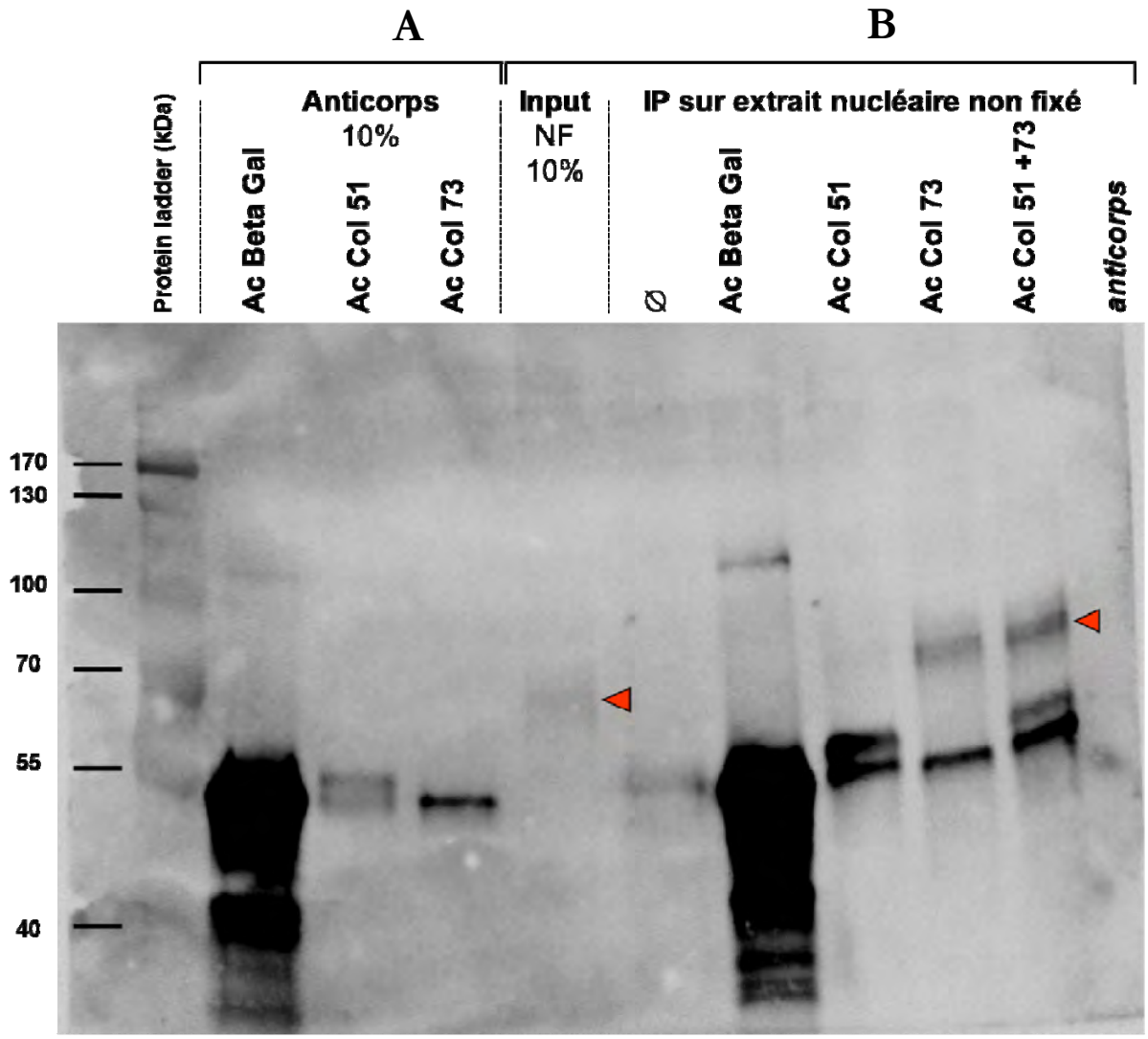
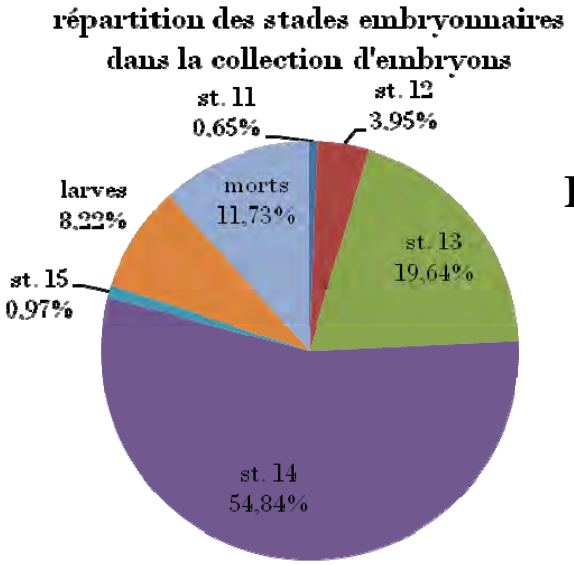


Fig. R8 – Les anticorps anti-Collier sont capables d'immuno-précipiter la protéine Col à partir d'extraits nucléaires d'embryons.

Détection de la protéine Col immuno-précipitée à partir d'un extrait nucléaire embryonnaire. **A :** détection des anticorps primaires utilisés pour l'IP. **B** : Détection de la protéine Col dans les immuno-précipitats. **Input** : extrait nucléaire avant IP. La protéine Col y est faiblement détectable (pointe rouge) – **IP sur extrait nucléaire non fixé** : La protéine Col est très légèrement enrichie dans l'IP avec l'anticorps 51, l'anti-Col 73 permet un bon enrichissement de la protéine Col et l'immuno-précipitation est améliorée par la combinaison des 2 anticorps 51+73.

A

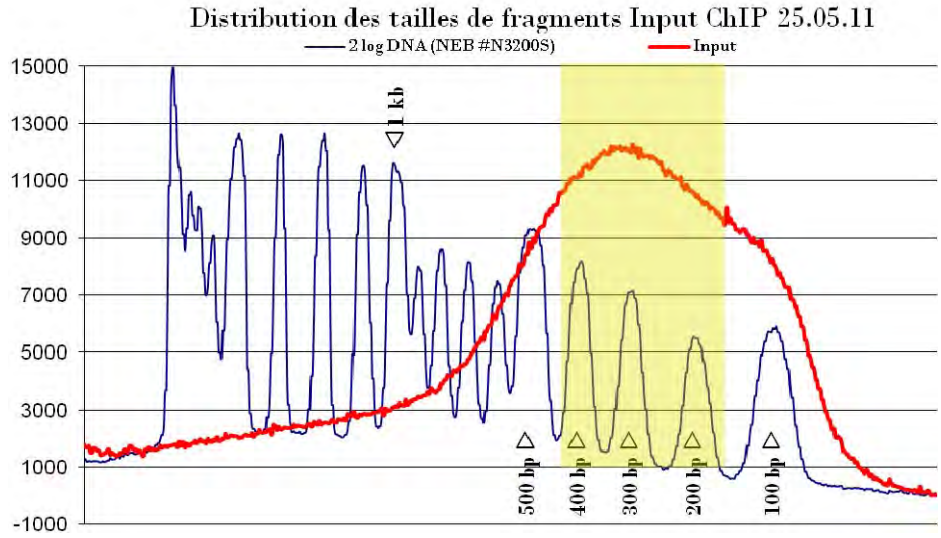


B

	Volume	Concentration	Quantité
Input	10 µL	98 ng/µL	980 ng
ChIP Col	18,5 µL	0,35 ng/µL	6,5 ng
ChIP HA (Mock)	22 µL	0,33 ng/µL	7,3 ng

Dosage avec Quant-iT dsDNA (Invitrogen #Q32851)

C



D

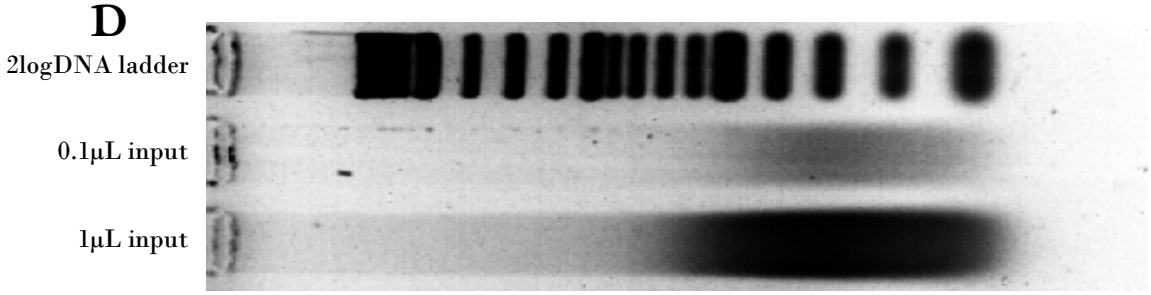
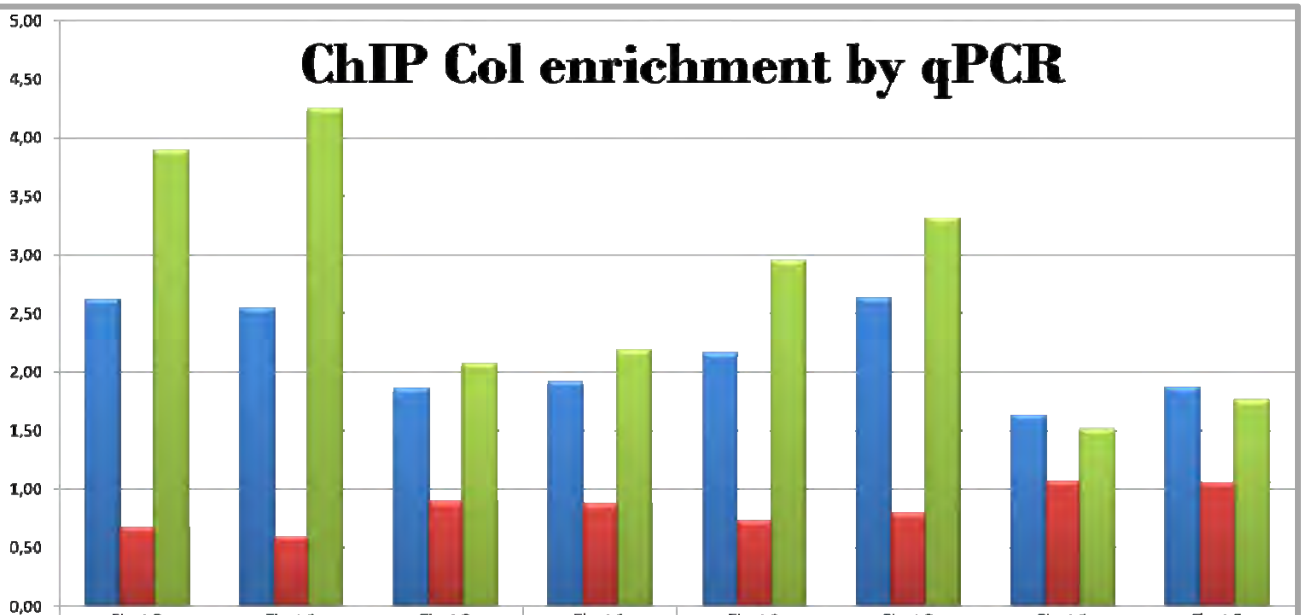


Fig. R9 – Caractéristiques de l'ADN des échantillons ChIP séquencés.

A. Diagramme de répartition des embryons utilisés pour les expériences de ChIP selon leur stade de développement. **B.** Dosage de l'ADN contenu dans les échantillons avant (Input) et après immuno-précipitation avec les anticorps Col ou HA. **C.** Diagramme de répartition de tailles des fragments après sonication de l'extrait nucléaire avant immuno-précipitation (Input) et **D.** Gel d'électrophorèse correspondant : la position des fragments de taille définie (2logDNA ladder) est représentée par la courbe bleue et la répartition de taille de fragments avant IP par la courbe rouge (unités arbitraires). Le rectangle jaune indique la fraction des fragments sélectionnés pour le séquençage (entre 200 et 400 pb).

A



Occupancy ChIP Col	2,63	2,55	1,86	1,92	2,17	2,64	1,63	1,87
Occupancy ChIP HA (mock)	0,67	0,60	0,90	0,88	0,73	0,80	1,07	1,06
ChIP Col enrichment	3,90	4,26	2,08	2,19	2,96	3,32	1,52	1,77

B

$$ChIP\ Col\ enrichment = \frac{Occupancy\ ChIP\ Col}{Occupancy\ ChIP\ HA\ (mock)} = \frac{Normalized\ Fold\ Change\ ChIP\ Col}{Normalized\ Fold\ Change\ ChIP\ HA\ (mock)}$$

$$Normalized\ Fold\ Change = \frac{(E_{primers\ target})^{\Delta Ct\ target}}{(E_{primers\ ref.})^{\Delta Ct\ ref.}}$$

Occupancy = mesure de l'occupation d'un site donné sur le génome par la protéine d'intérêt.

$E_{primers}$ = Efficacité d'amplification des paires de primers : une efficacité de 100%, c'est-à-dire permettant de doubler la quantité de produits à chaque cycle, est égale à 2, une efficacité de 90% sera égale à 1.9…

Target = site connu comme cible de la protéine d'intérêt, c'est-à-dire de Col ; 2 sites ont été testés : un site d'autorégulation présent sur le promoteur du gène *col* (« *col wt* »), et un site Col consensus (« Col Cons »)à la même position dans un transgène.

Ref. = site « neutre » pris comme référence sur le génome. 2 sites ont été choisis : un fragment en amont du gène CG11964, et un en amont du gène rp49.

Δ**Ct** = variation de Ct pour le site considéré (target ou ref.) entre la condition ChIP (Col ou HA) et la condition contrôle (input, avant IP) = Ct Input – Ct ChIP

Fig. R10 – Analyse qualitative de l'immuno-précipitation de fragments spécifiques d'ADN par qPCR

A. Histogramme des enrichissements relatifs des fragments d'ADN couvrant 2 sites de fixation connus de Col : site Col wt et site Col consensus inclus dans le CRM musculaire tardif, par rapport à 2 fragments de référence (CG11964 et rp49). Les IP ont été réalisées en duplicat expérimental (Éluats 1 et 2). **B.** Détails de la méthode de calcul utilisée, d'après (Fraga, 2008).

faiblement exprimé dans le segment intercalaire de la tête dans les embryons aux stades 11-12, soit 4.6% des embryons de la collection.

Pour l'immuno-précipitation de la chromatine (ChIP), nous avons adapté le protocole du laboratoire d'E. Furlong (Sandmann et al., 2006a) aux caractéristiques propres à notre expérience : prise en compte de l'utilisation d'anticorps monoclonaux plutôt que polyclonaux, optimisation du temps de fixation pour la protéine Col, paramétrage du sonicateur pour le fractionnement de la chromatine (Figure. R9 – C et D) et optimisation de l'élution pour essayer de pallier le faible nombre de noyaux exprimant Col dans l'embryon entier. (Voir Matériel et Méthodes).

Enfin, nous avons choisi de réaliser nos expériences de ChIP à partir d'embryons de drosophiles transgéniques portant un transgène rapporteur ^{L}CRM-Col^{Cons}-lacZ. Dans ce transgène, lacZ est placé sous le contrôle du ^{L}CRM tardif de Col modifié par substitution du site normal de fixation Collier par un site « consensus », de très haute affinité pour les protéines COE, tel que défini par selex (Daburon et al., 2008; Hagman et al., 1995). L'expression de LacZ à partir de ce transgène est identique au transgène modifié, excepté pour son niveau beaucoup plus élevé dans le muscle DA3 (L. Dubois). La présence de ce gène rapporteur pouvait nous servir de contrôle interne, en plus du site Col endogène présent au niveau du CRM tardif de *col* (site Col wt), seul site de fixation direct de Collier identifié dans le lignage DA3.

Pour contrôler la qualité de l'immuno-précipitation, des expériences de qPCR ont été réalisées afin de mesurer l'enrichissement relatif des fragments contenant les sites de fixation de Col dans le ^{L}CRM (site Col wt et site Col consensus) par rapport à 2 séquences de référence (situées en amont des gènes *cg11964*, référence utilisée par [(Sandmann et al., 2006a)] et *rp49*, gène dit « de ménage »)(Fraga, 2008). 2 expériences de ChIP Col ont été réalisées en parallèle ainsi que 2 expériences de ChIP HA comme contrôle, afin de s'assurer de la reproductibilité de la procédure expérimentale. Pour ces expériences, nous avons obtenu un enrichissement d'environ 4 (3.90 et 4.26) pour l'immuno-précipitation du site Col consensus par rapport au locus *cg11964* et de 2 (2.08 et 2.19) pour l'enrichissement du site Col wt par rapport à cette même référence. Cet enrichissement est légèrement inférieur lorsqu'on prend comme référence le locus *rp49* : autour de 3 pour le site Col consensus (2.96 et 3.32) et de 1.5 pour le site Col wt (1.52 et 1.77) (cf. Figure. R10). Etant donné la reproductibilité de ces résultats, nous avons regroupé les 2 expériences de ChIP Col ainsi que les 2 expériences contrôle avant de faire séquencer les fragments immuno-précipités (Fig. R9-B).

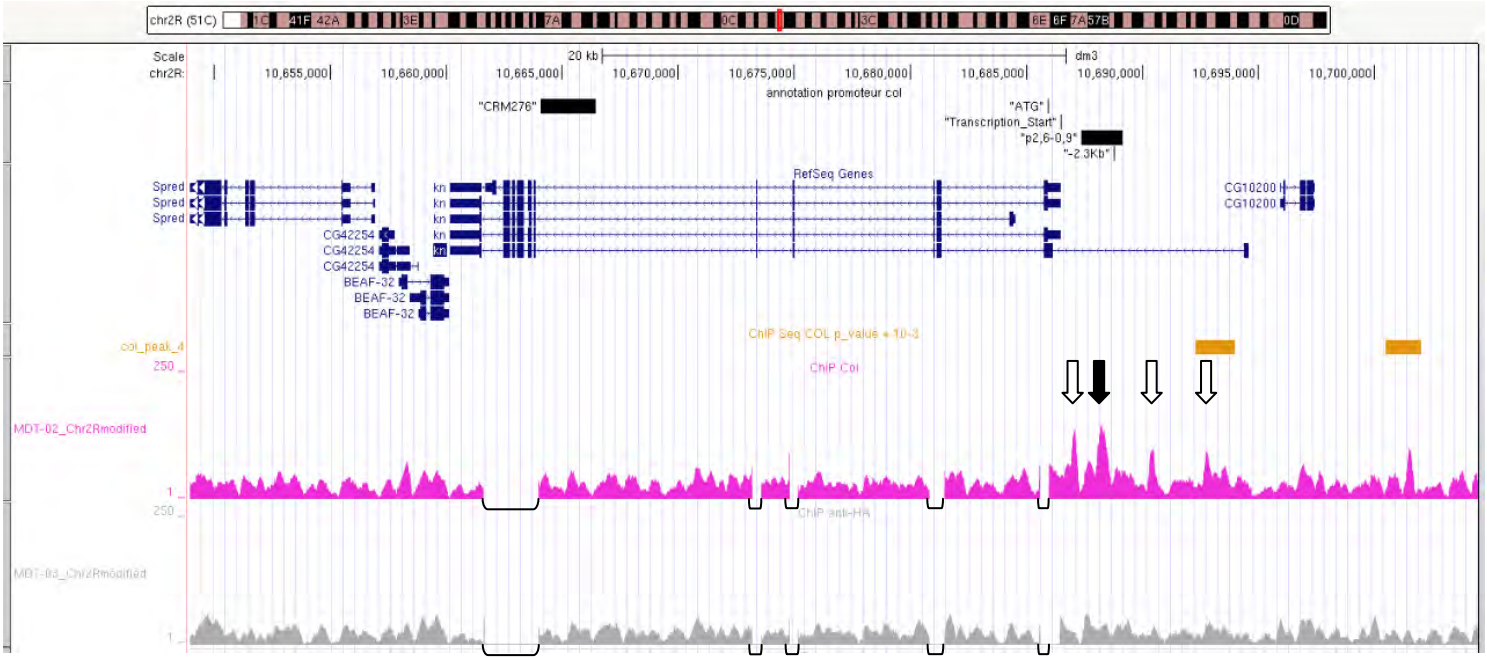


Fig. R11 – Annotation de la région génomique de *col*

Visualisation des données de ChIPseq pour la région génomique de *col* (*kn*) sur le UCSC Genome Browser (genome.ucsc.edu/). La comparaison entre la piste contrôle (ChIP avec un anticorps anti-HA, piste grise), et la piste ChIP Col (piste rose) montre 4 pics en amont du gène *col* (flèches verticales). Le pic le plus élevé (flèche noire) recouvre le site d'autorégulation directe précédemment validé fonctionnellement (Dubois et al., 2007) au sein du ^{L}CRM (2,6-0,9). Les positions du ^{E}CRM (CRM276) et du ^{L}CRM sont indiquées par des rectangles noirs sur la piste supérieure. Les données de ChIPseq correspondant aux exons de *col* (crochets noirs) ont volontairement été retirées pour permettre une meilleure visualisation des données correspondant aux introns (cf. Discussion).

c) *Détection des pics de fixation de Collier sur la chromatine (peak calling) (avec Bernard Jost et Stéphanie Le Gras – IGBMC Strasbourg)*

La synthèse de la librairie (amplification, ligation des adaptateurs…) a été réalisée par la plateforme de séquençage de l'IGBMC – Strasbourg -, ainsi que le séquençage des échantillons avec un séquenceur Illumina GA2X (méthode Solexa). Les fragments séquencés ont été alignés sur le génome de la drosophile (Dm3 – R5.12). Ils sont visualisables sur le UCSC Genome Browser (genome.ucsc.edu/) sous format .wig (fichier signal). Au total, 16 138 952 fragments séquencés d'une taille moyenne de 190 pb ont pu être alignés sur le génome de la drosophile pour l'échantillon ChIP Col et 13 135 866 fragments pour l'échantillon contrôle ChIP HA, soit environ 25 et 22 fois respectivement la taille du génome de la drosophile. Malheureusement l'automatisation de la procédure a écarté les séquences redondantes dont la région en amont de *col* dupliquée dans le gène rapporteur lacZ et la région codante de *col* en raison d'une contamination par un plasmide d'expression au moment de l'élution des fragments (cf. Discussion et Matériel & méthodes). La position des fragments séquencés dans ces régions n'était donc pas visualisable sur le *genome browser*. Néanmoins, un traitement *ad hoc* des données brutes de séquences, en retirant manuellement les valeurs aberrantes correspondant aux exons de *col*, a permis de reconstituer le profil des fragments immuno-précipités dans la région génomique de *col*, hors exons. Ce profil montre 4 sites potentiels de fixation de Col en amont du site d'initiation de la transcription (Fig. R11). Le plus élevé contient le site d'autorégulation directe précédemment validé fonctionnellement (Dubois et al., 2007) au sein du ^{L}CRM.

La détection des pics (peak calling) a été effectuée en utilisant le logiciel Macs (Pepke et al., 2009), particulièrement adapté à la détection de pics de fixation de facteurs de transcription, supposés étroits. Les pics significatifs spécifiques de ChIP Col sont détectés par comparaison de des pistes ChIP Col et ChIP HA, (p-value $\leq$ 10-3). N'ont été retenus que les pics pour lesquels l'enrichissement (« fold enrichment ») entre le ChIP Col et le ChIP HA était supérieur ou égal à 2.45, soit 413 pics associés à la fixation de Col sur le génome de la drosophile, avec un enrichissement compris entre 2.45 et 22. Le choix du seuil de 2.45 ainsi que la p-value ont été retenus avec l'expertise de la plateforme de l'IGBMC : la position des pics détectés avec ces seuils sur un échantillon aléatoire de loci permet de visualiser l'ensemble des pics significatifs en évitant au maximum les faux positifs. Le seuil d'enrichissement fixé à 2.45 est relativement haut, surtout si on prend en considération que l'enrichissement des seules cibles connues de Col est plutôt proche de 2 dans les expériences de qPCR (cf. ci-dessus). J'ai sciemment choisi de restreindre

A

Détection des pics (MACS) :

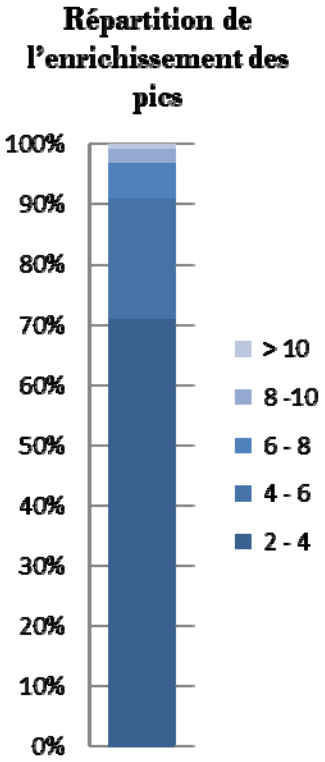
- *P-value ≤ 10^{-3}*
- *Fold enrichment (ChIP/mock) ≥2.45*

⇨ ***413** pics détectés (1600 pb en moyenne)*

Annotation des pics (PeakAnalyzer)

- ***330** gènes associés aux pics (Near Downstream Gene)*
- ***279** gènes associés à un pic unique (85%)*
  ***51** gènes associés à des pics multiples.*

B



C

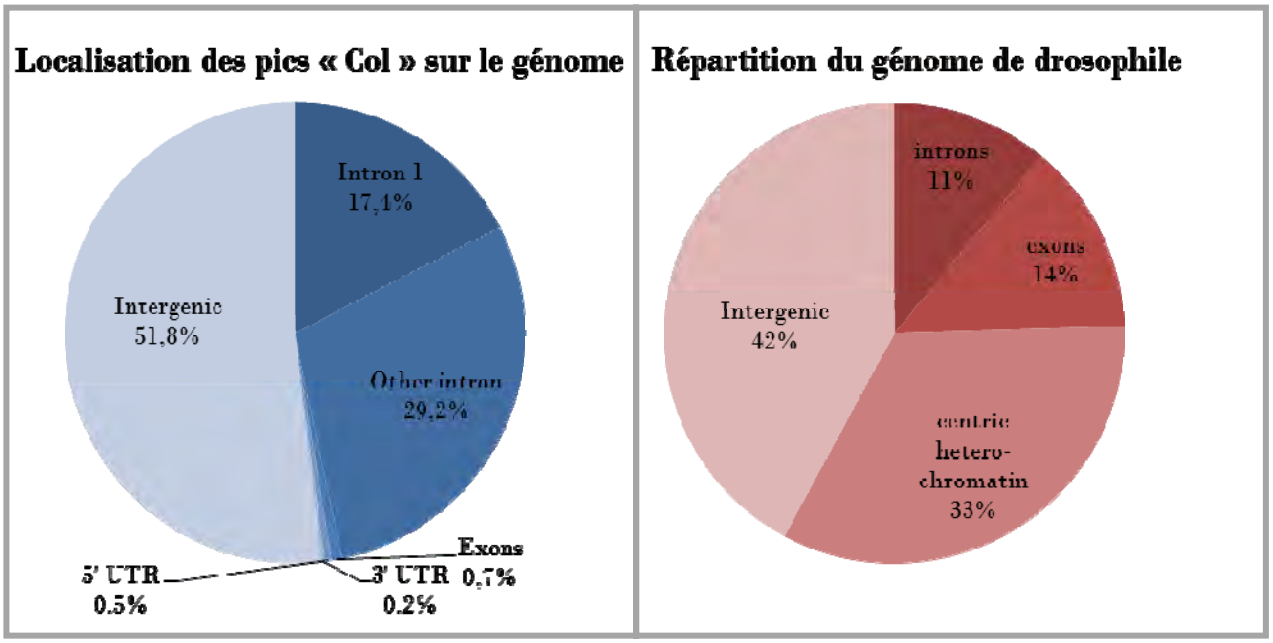


Fig. R12 – Détection et annotation des pics issus du ChIP Col

A. Paramètres utilisés et résultats obtenus pour la détection et l'annotation des pics associés à la liaison de Col. **B.** Diagramme de répartition de l'enrichissement (ratio ChIP Col/ChIP HA) des 413 pics détectés. **C.** Diagramme de répartition de la localisation des sites de fixation de Col *in vivo* en comparaison de la distribution globale du génome de la drosophile entre introns, exons et régions intergéniques.

mon analyse à ces 413 pics de plus fort enrichissement, en sachant que les données brutes restent disponibles pour une analyse des pics d'enrichissement moindre, le cas échéant.

d) Analyse bio-informatique globale des résulats de ChiP SEQ Col (avec Anaïs Painset – INSA Lyon Bioinformatique et Modélisation)

51.8% des 413 pics de fixation de Col retenus sont localisés dans une région intergénique (dont 35% à moins de 5 kb en amont d'un gène) et 46.6% dans un intron, ce qui, par rapport à une répartition globale du génome de la drosophile, représente un enrichissement en régions introniques (Fig. R12). Cette préférence de fixation de Col dans les introns est en accord avec la présence de nombreux CRM putatifs dans les introns (Roy et al., 2010).

Les pics ont été annotés en utilisant l'option *Near Downstream Gene* de PeakAnalyzer (Salmon-Divon et al., 2010) : chaque pic est donc associé au gène en aval le plus proche ou, dans le cas de la localisation dans un intron, au gène correspondant. Plusieurs pics pouvant être associés au même gène, 330 gènes ont ainsi été définis comme cibles potentielles de Col (Fig. R12).

Une analyse de la fonction prédite de ces gènes sur la base de données bibliographiques via le logiciel DAVID (http://david.abcc.ncifcrf.gov/ ; (Huang da et al., 2009)) montre une grande diversité de catégories fonctionnelles avec un léger enrichissement, peu significatif, pour les facteurs de transcription et les protéines transmembranaires. Cette analyse ne permet donc pas d'identifier un réseau spécifique de gènes/fonction en aval de Collier (Fig. R13). Le réseau de régulation génique dans lequel évolue Col ne semble donc pas restreint à une catégorie fonctionnelle donnée, une conclusion en accord avec la diversité d'expression, de fonctions embryonnaires de ce facteur et des quelques cibles connues.

La recherche de motifs *de novo* sur les séquences des pics via la suite MEME (MEME, MEME-ChIP et TOMTOM http://meme.nbcr.net/meme/ ; (Bailey et al., 2009)) a permis de mettre en évidence la présence d'un site apparenté au motif EBF (TCCCnnGGGA), conforme au motif consensus préalablement défini par selex, sur 90% des pics (373 motifs/413 pics ; analyse à +/-100 pb autour du sommet du pic). L'excellente E-value associée (= 1,7 e-252), leur détection sur la grande majorité des pics issus du ChIP Col et leur position - ces motifs sont majoritairement retrouvés au sommet des pics – permettent de conclure que les expériences de ChIP Col ont permis d'identifier des sites de fixation de Col *in vivo* sur la chromatine. La matrice PWM associée permet de définir le motif consensus de fixation de Col *in vivo* (TCCCnnGGGA) similaire au site défini pour EBF par des expériences de ChIP-chip à partir de cellules lymphoïdes (Treiber et al., 2010b) (cf. Fig.R13). Pour chaque motif Col identifié par cette analyse, j'ai recherché une corrélation éventuelle entre la nature du site (palindrome étendu :

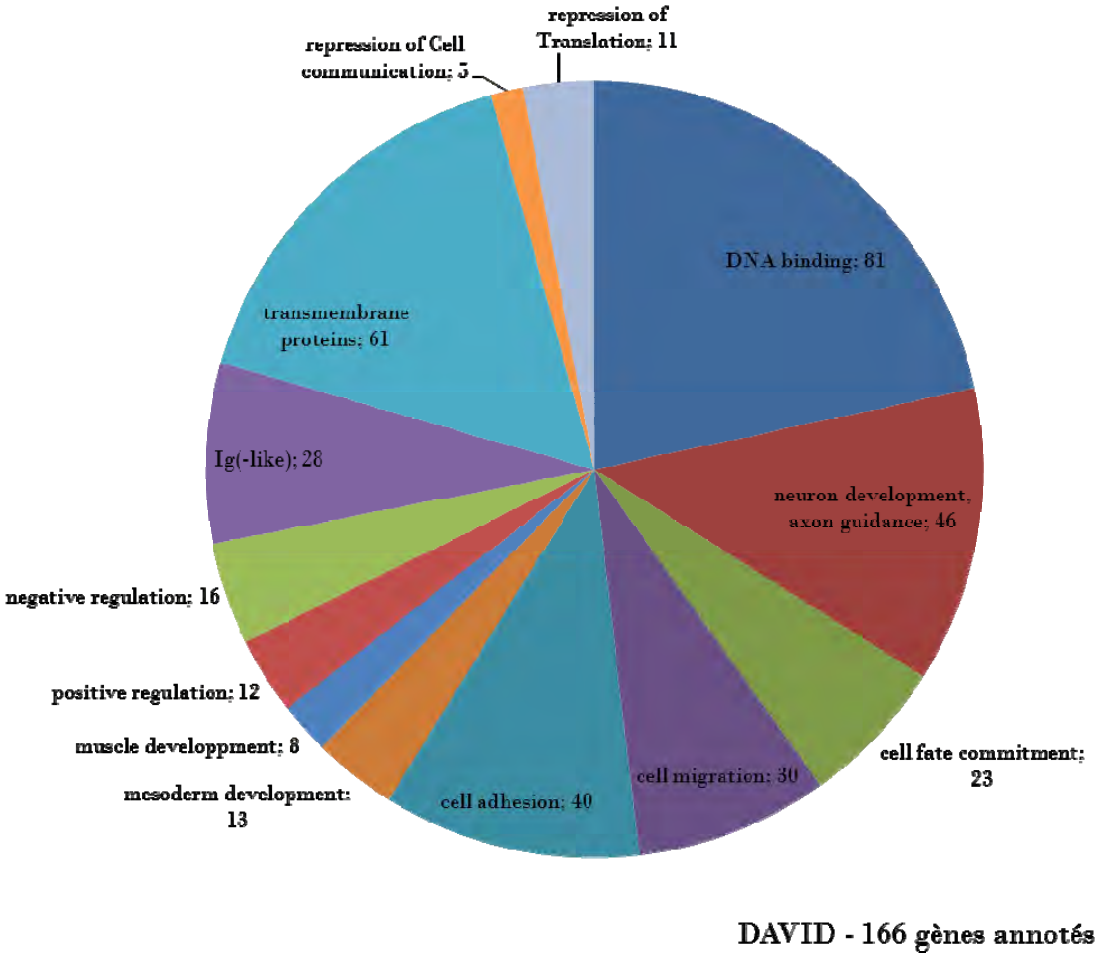
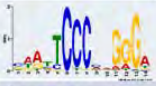


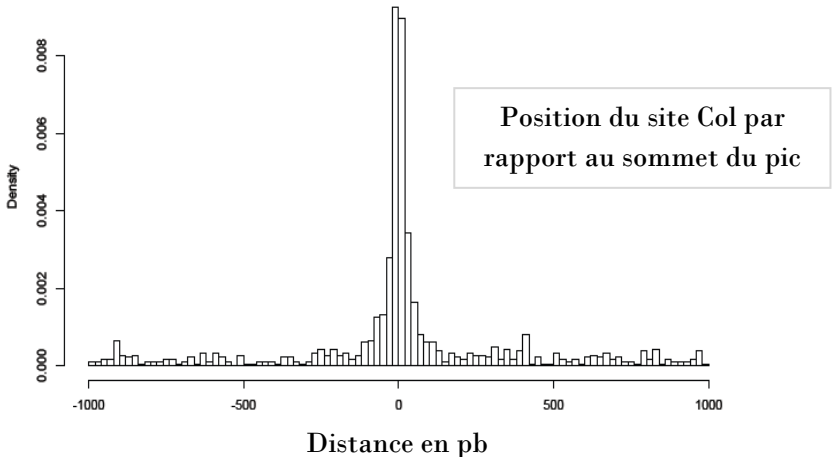
Fig. R13 – Fonctions prédites des gènes cibles potentiels de Col (DAVID)

Diagramme des différentes fonctions représentées par les gènes cibles potentiels de Col, d'après le logiciel DAVID (http://david.abcc.ncifcrf.gov/). Les bases de données utilisées pour cette annotation sont : COG-ontology, SP-PIR (**P**rotein **I**nformation **R**essource) keywords, UPseq features (catégories fonctionnelles), GOTERM (ontologie), Interpro, pir-superfamily, smart (domaines protéiques), et Kegg (voies de signalisation).

A

De novo discovery	MEME analysis			TOMTOM analysis	
	motif	Number of sites	E-value	Associated factor	P-value
+/- 100 pb around summit		373 (90%)	1,7 e-252	EBF-1	1,78 e-06
+/- 500 pb around summit		324 (78%)	2,6 e-107	EBF-1	1,89 e-06
+/- 100 pb – 100 best fold enrich.		100 (100%)	1,1 e-92	EBF-1	3,42 e-06
+/- 100 pb around summit Markov Model 6		396 (96%)	2,8 e-158	EBF-1	2,40 e-06

B



C

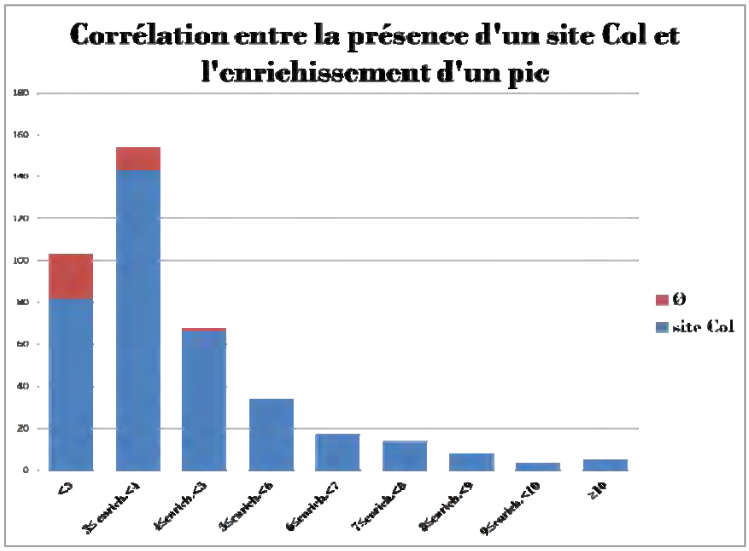


Fig. R14 – Les pics Col *in vivo* sont enrichis en motifs de reconnaissance pour les protéines COE

A. Recherche *de novo* de motifs sur les séquences ADN des pics Col, via MEME, et assignation de ces motifs à un facteur connu via TOMTOM. La matrice la plus enrichie correspond à une matrice de type EBF-1 (Travis et al., 1993; Treiber et al., 2010). **B.** Localisation des motifs Col/EBF-1 montrant un fort enrichissement au centre du pic. **C.** Répartition des sites Col en fonction de la hauteur du pic. Tous les pics de plus fort enrichissement (<5) contiennent un site Col.

TCCCnnGGGA, palindrome restreint : CCCnnGGG ou site non palindromique) et la hauteur du pic associé (Fig. R15). Les sites étendus sont majoritairement retrouvés dans les pics de plus haute valeur, suggérant une plus haute affinité de Col pour ces motifs.

La composition en nucléotides et les fréquences d'enchaînements propres au génome de la drosophile sont des paramètres qui doivent être pris en compte dans notre analyse. Cette prise en compte nécessite une modélisation de ce biais de composition par l'intermédiaire d'une chaîne de Markov. Avec une chaîne de Markov de niveau 6 (c'est-à-dire prenant en compte les probabilités d'occurrence d'un nucléotide par rapport aux 6 nucléotides précédents), le motif Col est retrouvé dans la séquence de la grande majorité des pics (396 motifs/413 pics ; analyse à +/- 100 pb autour du sommet du pic) avec une très bonne E-value (=2,8.e-158) (Fig. R14). Les 2 premiers G de la matrice PWM correspondante ont cependant un poids moindre dans le motif. A l'inverse, si on effectue la recherche sur le jeu des séquences des 100 meilleurs scores d'enrichissement, la recherche de motifs *de novo* permet de dégager un motif Col dont les T et A flanquants prennent davantage d'importance (voir Discussion). Dans tous les cas, le motif Col est retrouvé de manière prédominante ; aucun autre motif significativement enrichi n'a pu être dégagé de cette analyse MEME, si ce n'est une séquence de 7 pb (matrice étendue de 15 pb : TTTT[C/T]TCTC[A/T]GTG[C/T]A) présente dans une centaine de séquences et qui ne correspond à aucun site de facteur de transcription déjà connu. La matrice la plus proche associée correspond au site de fixation pour le facteur vertébré Nkx3.2, mais avec une p-value de 3 e-03, qui est peu significative.

La recherche de motif *de novo* a également été réalisée par l'intermédiaire de la suite RSAT (**R**egulatory **S**equence **A**nalysis **T**ools, http://rsat.ulb.ac.be/rsat/ ; (Thomas-Chollier et al., 2012)). Cette suite donne la possibilité de sélectionner directement un « background » spécifique du génome de la drosophile qui est ajusté en fonction des séquences fournies en entrée et de rechercher des motifs enrichis sur ces séquences par différents algorithmes : analyse de nucléotides (surreprésentation de motifs de 6-7 nucléotides sur les séquences), biais de position (surreprésentation par rapport à un point donné sur la séquence, comme le centre du pic par exemple), analyse par dyad (pour des motifs palindromiques, recherche de paires de tri-nucléotides séparées par un nombre fixe de nucléotides aléatoires, ce qui donne davantage de flexibilité qu'une analyse de nucléotides simple)… Comme pour la suite MEME, les motifs découverts peuvent ensuite être associés à des matrices connues pour des sites de fixation de facteurs répertoriés dans différentes banques de données. Quel que soit l'algorithme utilisé, le motif Col a été retrouvé en tête de liste, la recherche par dyad ayant été particulièrement performante puisque le motif Col possède une structure adaptée à ce type de recherche : une

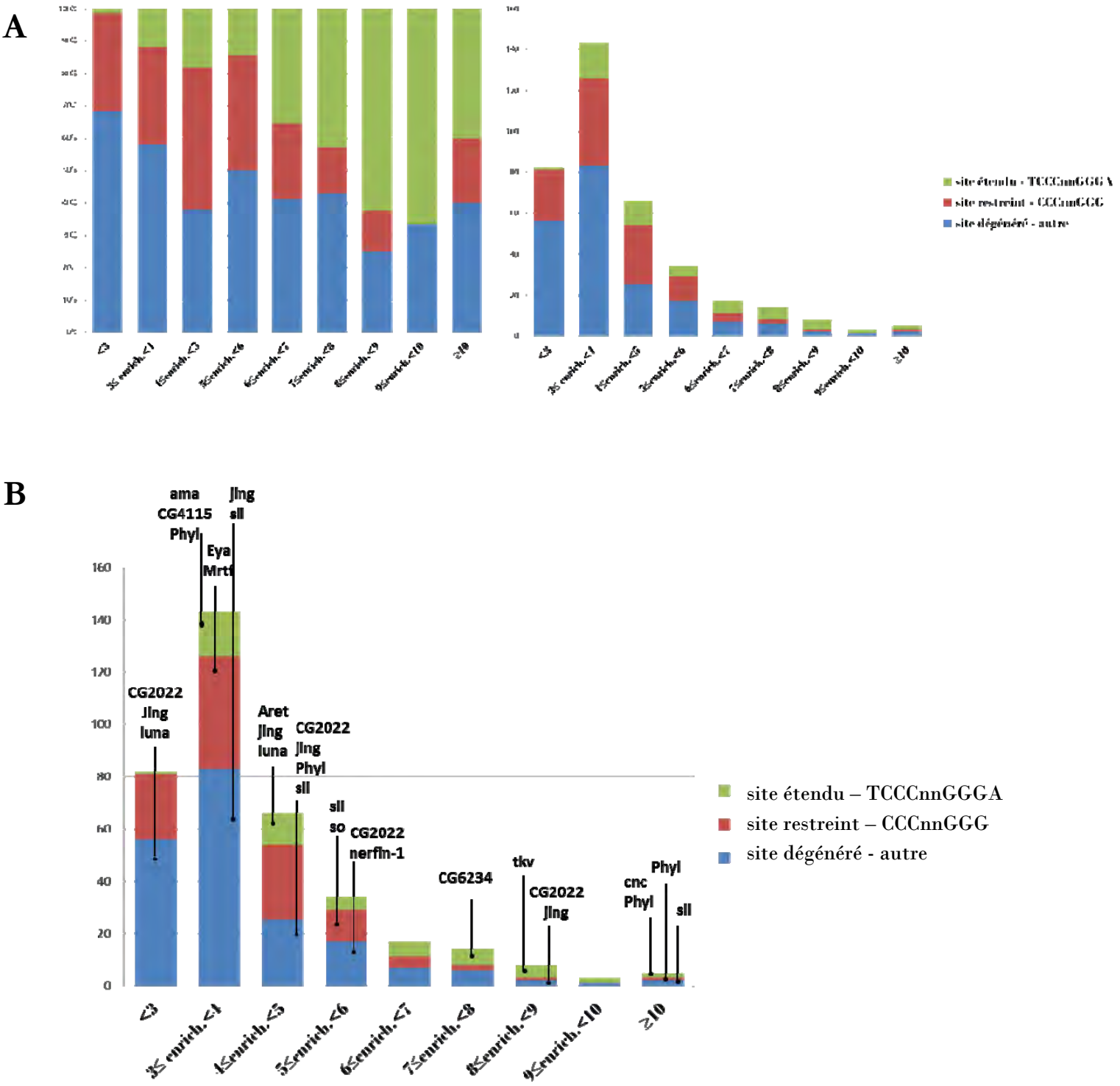
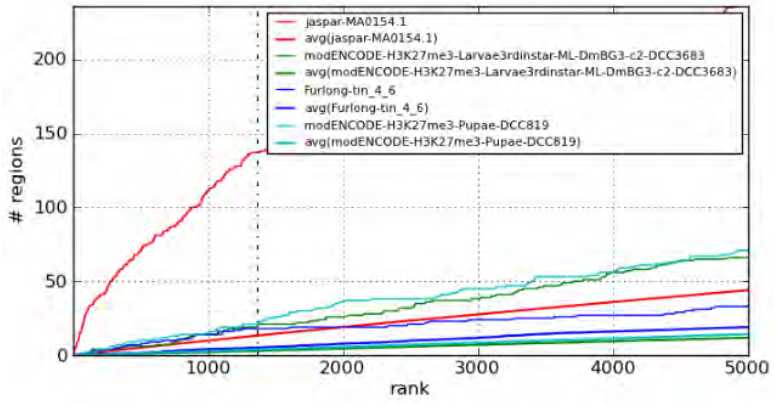


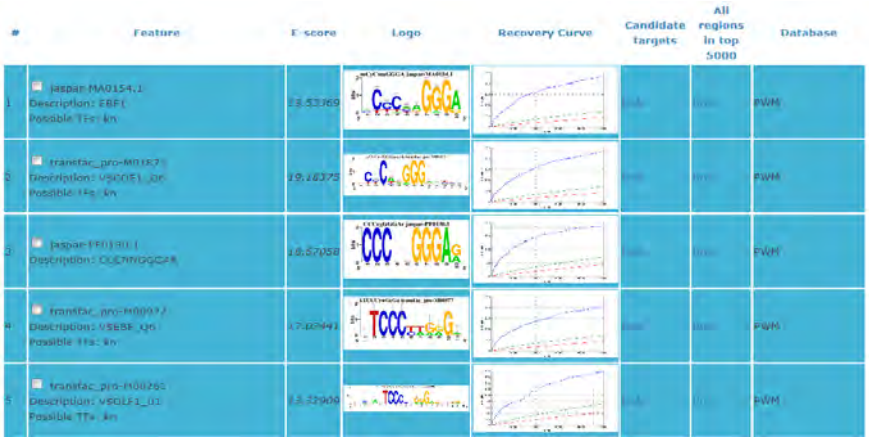
Fig. R15 – Relation entre la force du site de liaison et l'enrichissement du pic

A. Répartition des sites Col (étendu, restreint ou dégénéré) relative à l'enrichissement des pics. A gauche : pourcentages ; à droite : valeurs absolues. **B.** Classification des motifs Col pour 15 des gènes candidats étudiés (*nota :* certains gènes possèdent plusieurs pics, chacun ayant un motif Col).

A



B



C

i-cisTarget

An integrative genomics method for the prediction of regulatory features and cis-regulatory modules in *Drosophila*

Overview of significantly high ranked regions

236 significantly highly ranked regions (higher than 4891 in 136353) for jaspar-MA0154.1.

Rank	Region ID	Target present in input query?	Associated genes
1	chr2L-reg7201	True	
2	chr3L-reg11794	False	klu
3	chr2R-reg11222	True	Oaz
4	chr3R-reg20589	False	
5	chr3L-reg24219	False	CG14459
6	chr2R-reg12147	False	trpm
7	chrX-reg9561	True	Lim1
8	chr3L-reg15830	False	NHP2
9	chr2L-reg15434	False	CG9014
10	chr3R-reg13433	True	CG42342

Fig. R16 – L'analyse Cis-Target des pics met nettement en avant la présence du site Col

A. Recovery Curve of best features. La courbe de découverte du motif jaspar MA0154.1 (motif Col – courbe rouge fine) est nettement supérieure à la moyenne de découverte de ce motif sur un set aléatoire de séquences (avg jaspar MA0154.1 – courbe rouge épaisse), et ce motif est clairement prédominant devant les autres motifs découverts (site Tinman et marque d'histones H3K27me3). **B.** Les 5 premiers motifs découverts sur les séquences des pics ChIP Col sont les motifs EBF/COE. **C.** Exemple de liste de régions cibles potentielles portant la même « signature » en terme de motif découvert. Ici les 10 premières régions associées à la matrice Jaspar MA0154.1 (« Candidate targets »), rangées par ordre de significativité. Les régions signalées par « true » dans la 3e colonne appartiennent également au set de départ, i.e. les séquences des pics ChIP Col.

paire de tri-nucléotides séparée par deux bases quelconques. La recherche par biais de position a, quant à elle, confirmé la présence des motifs Col à proximité du sommet du pic de ChIP. Cependant, comme avec MEME, la recherche *de novo* n'a mis en évidence aucun autre motif surreprésenté dans les séquences analysées.

Enfin nous avons soumis notre jeu de séquences au logiciel i-CisTarget (anciennement CisTargetX - http://med.kuleuven.be/lcb/i-cisTarget/ - (Herrmann et al., 2012)). À partir d'une liste de gènes régulés conjointement (ici notre liste des 330 gènes cibles potentielles de Col), i-CisTarget recherche les motifs enrichis en amont de ces gènes et dans leur premier intron afin d'identifier de potentiels sites de fixation pour des facteurs de transcription. Pour chacun des motifs découverts, une liste de régions génomiques portant le même motif est ensuite proposée. Ces régions sont classées selon leur probabilité d'être régulées par le facteur de transcription associé. Le classement fait intervenir la conservation entre 12 espèces de drosophiles. En sélectionnant plusieurs des motifs découverts, i-CisTarget peut également définir et rechercher des CRM, c'est-à-dire des clusters de sites de fixation, sur l'ensemble du génome de la drosophile et ainsi mettre en évidence des combinatoires de facteurs de transcription nécessaires à la régulation d'un réseau de gènes. i-CisTarget peut également prendre en charge directement les séquences de pics issus de ChIP et appliquer les algorithmes de recherche de motifs *de novo* sur ce set. L'analyse *de novo* sur les séquences des pics du ChIP Col (413 séquences) met à nouveau clairement en avant la présence du motif Col sur ces séquences, devant tout autre motif avec un E-score entre 13 et 19 pour les motifs Col, et autour de 9 pour le premier motif suivant, le site de fixation d'un facteur de transcription à doigt de zinc chez *S. cerevisiae* (Fig. R16). Par contre, l'analyse à partir de la liste de gènes cibles potentiels de Col (330 CG), c'est-à-dire sans identification de la région de fixation potentielle de Col, fait davantage ressortir les marques d'histones (H2b…) ou les facteurs associés à l'ouverture de la chromatine (Pc, dRing, Ez…)… que le site Col qui n'est trouvé qu'en 39^{e} position ! Les E-scores associés à cette analyse sont également plus faibles (autour de 7 pour les meilleurs scores). L'identification du site Col sans connaissance préalable de sa position n'est donc pas si aisée et le motif est facilement « noyé » lorsqu'un ensemble trop important de séquences adjacentes est pris en compte.

Un des objectifs de l'analyse bio-informatique globale des données issues du ChIP Col était d'essayer de déterminer une « grammaire » tissu-spécifique des CRM contrôlant les gènes cibles de Col dans le muscle versus dans le système nerveux. Puisqu'aucun enrichissement significatif n'a été trouvé sur le jeu global de données, nous avons sélectionné, via l'annotation fonctionnelle de DAVID un set de gènes annotés « muscle » (23 gènes) et un set de gènes annotés « système nerveux » (47 gènes) et soumis les séquences des pics associés à CisTarget afin de détecter des

motifs de facteurs de transcription spécifiques à chacun de ses groupes. Malheureusement les quelques motifs sélectionnés ne permettent pas par la suite de réassigner chacun des gènes selon leur catégorie initiale et ne peuvent donc pas être considérés comme spécifiques de CRM musculaires ou du système nerveux.

Les premières analyses bio-informatiques globales effectuées permettent donc essentiellement de mettre en évidence la présence d'un site Col sur la majorité des séquences issues du ChIP, et ce avec une forte significativité, puisque ce site est retrouvé quel que soit le logiciel utilisé. Nos efforts pour découvrir une grammaire spécifique aux CRM liés par Col sont cependant restés infructueux.

e) Sélection de gènes candidats - Systems biology.

Un jeu de 330 gènes cibles potentiels de Col est trop large pour que chaque gène puisse être analysé en détail. Une sélection d'environ 30 candidats semblait un bon compromis entre qualité de l'échantillonnage et la faisabilité de plusieurs analyses complémentaires. Cette sélection a été réalisée sur la base de plusieurs critères expérimentaux indépendants, issus d'analyses à l'échelle génomique systématique, détaillés ci-dessous : hauteur du pic, présence de sites de fixation *in vivo* de facteurs mésodermiques, de marques d'ouverture de la chromatine, corrélation avec des analyses transcriptomiques. Un critère supplémentaire est celui de la « séquence » du motif de fixation Col identifié par MEME.

Les 30 gènes candidats retenus sur la base d'un ou plusieurs de ces critères sont : *ama (amalgam), aret (arrest/bruno), cg12484, cg2022, cg34371, cg4115, cg4161, cg6234, cnc (cap-n-collar) , Dys (Dystrophin), eya (eyes absent), jing, kuz (kuzbanian), luna, mbl (muscleblind), Mrtf (Myocardin related transcription factor), nerfin-1 (nervous finger 1), numb, phyl (phyllopod), pum (pumillo), px (plexus), salr (spalt related), sens-2 (senseless 2), sli (slit), smr (smrter), so (sine oculis), ten-m (tenascin major), tkv (thickvein), tl (toll)* et *unc-5*.

Un premier critère absolu (sans préjugé des cellules dans lesquelles Col pourrait se fixer) est la hauteur des pics. J'ai donc sélectionné sur cette base les gènes *phyl* (enrichissement de 22.41 et 10.54), *smr* (13.6), *cnc* (12.62), *sli* (11.3), *px* (9.71), *cg2022* (8.68) et *jing* (8.57). À noter, l'existence de pics multiples pour 5 de ces 7 gènes: *jing* (5 pics), *phyl* (4), *sli* (4), *px* (3) et *cg2022* (3).

Un deuxième critère est la probabilité que ce gène soit exprimé dans le mésoderme à l'origine du muscle DA3. Pour satisfaire ce critère, j'ai choisi d'utiliser les données de ChIP générées pour les facteurs de transcription mésodermiques Mef2 et Twist (Twi) (Sandmann et al., 2007; Sandmann et al., 2006b). L'hypothèse est que des CRM liés à la fois par Col et Mef2 et/ou Twi

correspondent à des cibles de Col dans le muscle DA3. Cette hypothèse s'appuie largement sur le modèle de régulation de *col* lui-même, puisque le ^{E}CRM tardif qui est sujet à auto-régulation, est un site de liaison *in vivo* de Twi et Mef-2 (Dubois et al., 2007; Sandmann et al., 2007; Sandmann et al., 2006b). La fenêtre temporelle de notre étude favorisant plutôt un rôle de Mef2, j'ai sélectionné plusieurs gènes comportant un site de fixation pour ce facteur (et éventuellement un site de fixation pour Twi) situé à moins de 0.5 kb du site Col, la taille moyenne d'un CRM chez la drosophile ayant été défini à moins de 1kb. C'est le cas des gènes *cg2022, cg6234, Dys, Mrtf, pum, px, smr, tkv* et *tl*. En complément des données de fixation de Mef2 et Twi, j'ai utilisé des données issues de la même équipe concernant l'ouverture (l'accessibilité à l'ARN polymérase) de la chromatine dans le mésoderme (Bonn et al., 2012). Une hypothèse raisonnable est que les gènes présentant une « signature » de chromatine ouverte - fixation de la polymérase PolII et acétylation de l'histone H3K27 comme marques d'un promoteur actif, tri-méthylation de l'histone H3K79 comme marque d'une région génomique transcrite -, soient transcrits dans le mésoderme. Les gènes *ama*, *cg34371, cg6234*, *eya*, *jing*, *numb*, *phyl, px, ten-m, tkv* et *unc-5* répondent à ce critère. À noter cependant que ces données ont été générées pour des temps de développement antérieurs (st.10-11) aux stades considérés pour nos expériences de ChIP (st.13-14), c'est pourquoi ce critère n'a pas été considéré avec autant de poids que les 2 précédents.

Un troisième critère est la variation d'expression des gènes dans des embryons mutants pour *col* ou surexprimant *col* dans le mésoderme (Twi-gal4 x UAS-col). Ces données de transcriptomes ont été générées par Laetitia Bataillé et Jean-Louis Frendo dans notre laboratoire. Certains candidats présentant des variations d'expression significatives en conditions de mutation pour *col* (> -1.5 x ou $< +1.5$ x, p-value<0.05), et donc susceptibles d'être des cibles directes ou indirectes de Col dans les tissus où Col est exprimé, ont été sélectionnés : *ama, cg4115, cg6234, Mrtf, Dys, nerfin-1*et *ten-m*.

Un quatrième critère auquel je me suis intéressé est la nature, - séquence et conservation évolutive-, du site prédit de fixation de Col. Le motif optimal *in vitro* de fixation défini par selex pour les protéines EBF/Col est le motif ATTCCCnnGGGAAT (Travis et al., 1993). Sur 8 motifs de ce type présents dans le génome de la drosophile, on en retrouve un parmi les 390 motifs présents dans les fragments immuno-précipités, associé au gène *cg12484*. Le pic correspondant à une hauteur de 3.4. Il faut noter que les motifs correspondant aux pics de plus fortes hauteurs (*phyl*, *smr*, *cnc*) sont tous différents entre eux et également du consensus parfait, indiquant que la séquence nucléotidique du site reconnu *in vivo* par la protéine Col tolère un nombre important de variations et n'est pas déterminante pour la hauteur du pic. Je reviendrai sur ce point dans la discussion. Une étude de la conservation évolutive de ces motifs entre 12 espèces de drosophiles

Candidat	nombre de pics associés	hauteur du pic (enrichissement)	nature du motif Col	conservation du motif	site Mef2 <0,5b	transcriptome WT vs mutant col	ouverture de la chromatine
ama	1	3.25	étendu	ananassae		x	très bonne
aret	2	4,76 - 4,09	dégénéré - étendu	pseudo-obscura			fermée
cg12484	1	3,4	selex	grimshawi			fermée
cg2022	3	8,68 - 5,68 - 2,86	dégénérés	yakuba	x		fermée
cg34371	2	4,91 - 3,03	étendu - dégénéré	sechellia - erecta			très bonne
cg4115	1	3,38	étendu	erecta		x	fermée
cg4161	2	7,11 - 3,42	étendu - dégénéré	grimshawi			fermée
cg6234	1	7,45	étendu	willistoni	x	x	très bonne
cnc	1	12,62	étendu	persimilis			fermée
Dys	1	3.73	dégénéré	willistoni	x	x	fermée
eya	1	3,91	restreint	mojavensis			très bonne
jing	5	8,57 - 4,69 - 4,56 - 3,33 - 2,52	2 dégénérés - 1 étendu - 2 dégénérés.	ananassae			très bonne
kuz	1	2,92	dégénéré	sechellia			bonne
luna	2	4,37 - 2,65	étendu - dégénéré	ananassae			faible
mbl	2	6,41 - 3,1	étendu - dégénéré	ananassae		x	faible
Mrtf	1	3,35	restreint	sechellia	x	x	fermée
nerfin-1	1	5,31	dégénéré	erecta		x	fermée
numb	1	4,62	étendu	sechellia			très bonne
phyl	4	22,41 - 10,54 - 4,43 - 3,34	étendu-restreint-dégénéré -étendu	persimilis - erecta			très bonne
pum	4	6,27 - 5,92 - 4,83 - 4,72	dégénéré - restreint - étendu - dégénéré	persimilis - erecta	x		faible
px	3	9,71 - 2,97 - 2,57	étendu - dégénéré - Ø	erecta	x		très bonne
salr	1	4,87	restreint	persimilis			fermée
sens-2	1 (+1 : Rcal)	3,71 (-6,31)	Ø (- dégénéré)	(grimshawi)			fermée
sli	4	11,3 - 5,04 - 4,1 - 3,98	dégénéré - restreint - 2 dégénérés	ananassae - erecta			bonne
smr	1	13,6	dégénéré	ananassae	x		bonne
so	1	5,75	restreint	erecta			bonne
ten-m	1	2,79	restreint	ananassae		x	très bonne
tkv	1	8,27	selex	erecta	x		très bonne
tl	1	8,36	étendu	ananassae	x		bonne
unc-5	2	6,21 - 4,93	dégénéré - restreint	grimshawi	x		très bonne

Tableau. R17 – Liste des 30 candidats étudiés et leurs caractéristiques moléculaires

(de la plus proche à la plus éloignée du point de vue évolutif : *D. simulans, D. sechellia, D. melanogaster, D. yakuba, D.erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. virilis, D. mojavensis, D. grimshawi*) m'a conduit à sélectionner des gènes pour lesquels le motif reconnu par Col est conservé à la même position dans ces espèces, ou un sous-ensemble, indiquant que ces motifs sont soumis à une pression de sélection qui suggère un rôle fonctionnel (Stark et al., 2007). C'est par exemple le cas du motif associé aux gènes *cg12484, cg4161, sens-2* et *unc-5* (conservé entre *D. melanogaster* et *D. grimshawi*) ou à celui du gène *eya* (conservé entre *D. melanogaster* et *D. mojavensis*).

L'ensemble des 30 gènes sélectionnés avec leurs caractéristiques principales est présenté dans le tableau R17.

f) Validation d'une sélection de 30 gènes candidats

Plusieurs niveaux de validation des candidats sélectionnés en tant que gènes cibles directes de Col sont nécessaires afin d'entreprendre l'étude de leur fonction au cours de la myogenèse: un patron de transcription conforme à l'expression de Col et une activité cis-régulatrice du fragment d'ADN contenant le site de fixation *in vivo* de Collier. Enfin, une modification de cette activité cis-régulatrice est attendue si on mute le site de fixation prédit de Collier.

Expression des gènes candidats

Mon premier niveau de validation a été d'établir le profil de transcription des gènes candidats afin de déterminer s'il recouvre celui de Col. J'ai choisi comme méthode l'hybridation *in situ* avec des sondes fluorescentes, si possible introniques (pour 25 candidats sur les 30), doublée d'une immuno-coloration avec les anticorps anti-Col et/ou anti-Mef2 (permettant de localiser les noyaux des cellules mésodermiques).

Aucun des candidats retenus n'est exprimé spécifiquement dans le muscle DA3 mais 13 sont exprimés dans ce muscle et d'autres tissus. Il s'agit des gènes *ama, aret, cg34371, cg4161, cg6234, Dys, eya, jing, luna, mbl, Mrtf, nerfin-1,* et *so*. La transcription de ces gènes est détectée soit au stade progéniteur (*aret, eya, jing, so*) soit dans le muscle en formation (st. 14) (*ama, cg34371, cg4161, cg6234, Dys, luna, mbl, Mrtf, nerfin-1*). Les gènes *cg12484* et *smr* pourraient être également exprimés au cours de la formation du muscle DA3 mais le patron d'expression est confus et nécessite d'être étudié plus en détail.

Parmi les candidats restants, *cnc* était déjà connu comme étant régulé par Col dans le segment intercalaire de la tête. L'hybridation *in situ* avec une sonde intronique a confirmé sa transcription dans un domaine recouvrant partiellement celui de *col* dans la tête mais aussi l'absence d'expression de ce gène dans le muscle DA3. De manière intéressante, sont également exprimés dans la tête, dans un domaine recouvrant partiellement le domaine Col, les gènes *ama* et *nerfin-1*.

10 autres gènes candidats (*cg2022, cg4115, kuz, numb, phyl, px, salr, ten-m, tl* et *tkv*) sont très faiblement exprimés dans l'embryon, et la faiblesse du signal détecté ne permet pas de conclure quant à leur domaine d'expression.

Le gène *sli* n'est pas exprimé dans le muscle DA3, cependant j'ai pu observer son expression dans d'autres muscles dorso-latéraux, probablement issus du groupe promusculaire Col positif (Enriquez et al., 2012).

Enfin plusieurs candidats sont exprimés dans le système nerveux central (SNC) de l'embryon, y compris dans des neurones exprimant Col, leur régulation par Col pouvant avoir lieu dans ces neurones. Les gènes *sli*, *pum*, *unc-5* répondent à ce critère. Le gène *pum* est également exprimé dans les neurones multidendritiques de classe IV (qui expriment spécifiquement Col), dans les rangées dorsales et intermédiaires. On observe aussi des transcrits *mbl* et *sens-2* dans ces neurones multidendritiques.

Les patrons d'expression de 8 de ces candidats sont présentés figures R18 à R25, avec la visualisation de la région génomique associée sur le *genome browser* UCSC et un alignement de séquence présentant la conservation du motif prédit de fixation de Col localisé par (MEME). La figure R26 présente les fenêtres d'expression de 15 des gènes candidats telles que visualisées en hybridations *in situ* dans les domaines d'expression de Col. Le bilan de ces analyses est présenté dans le tableau en Annexe 1.

Analyse des CRM

Le niveau d'analyse suivant a concerné l'activité cis-régulatrice du fragment contenant le site de fixation de Col. Cette analyse a été réalisée pour une sélection de 15 gènes candidats : *ama, aret, cg2022, cg4115, cg6234, cnc, eya, jing, luna, Mrtf, nerfin-1, phyl, sli, so et tkv*. Cette deuxième sélection « à priori » a été réalisée, en parallèle à l'analyse des patrons d'expression par *in situ*, afin d'avoir un échantillonnage représentatif des critères de sélection appliqués (hauteur du pic, présence de pics multiples, présence de sites Mef2 à proximité etc.).

Une première approche, opportuniste et donc non systématique, a d'abord été exploitée : l'utilisation de la collection de lignées transgéniques GMR-gal4 réalisée par l'équipe de G.M.

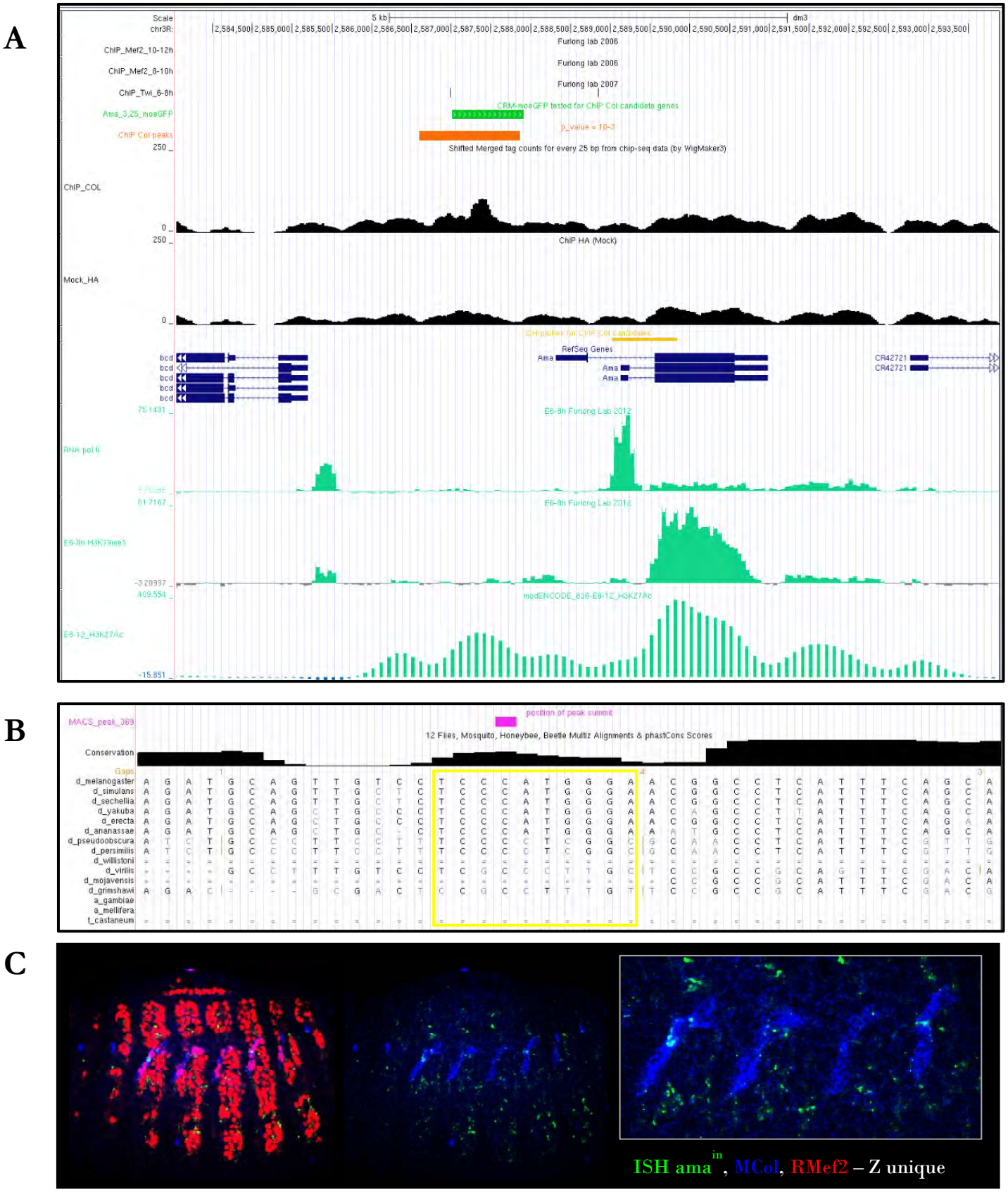


Fig. R18 – amalgam (ama)

A. Annotation de la région génomique d'*ama*. La position du pic Col est notée (orange), ainsi que la position des sites de fixation de Twi (6-8h de développement) et Mef2 (8-10h et 10-12h) d'après des expériences de ChIP (E. Furlong), les profils des marques d'histones (E. Furlong : 6-8h et ModEncode 8-12h de développement) et la position du fragment (jaune) utilisé comme sonde en ISH. **B.** Conservation de la séquence nucléotidique du pic *ama* entre 12 espèces de drosophiles et position du motif Col (encadré jaune) relativement au sommet du pic (magenta). **C.** Hybridation *in situ* avec la sonde *ama* (vert). Embryon st. 14. Le muscle DA3 est marqué par immuno-coloration avec l'anticorps anti-Col et l'ensemble des noyaux musculaires avec l'anticorps anti-Mef2.

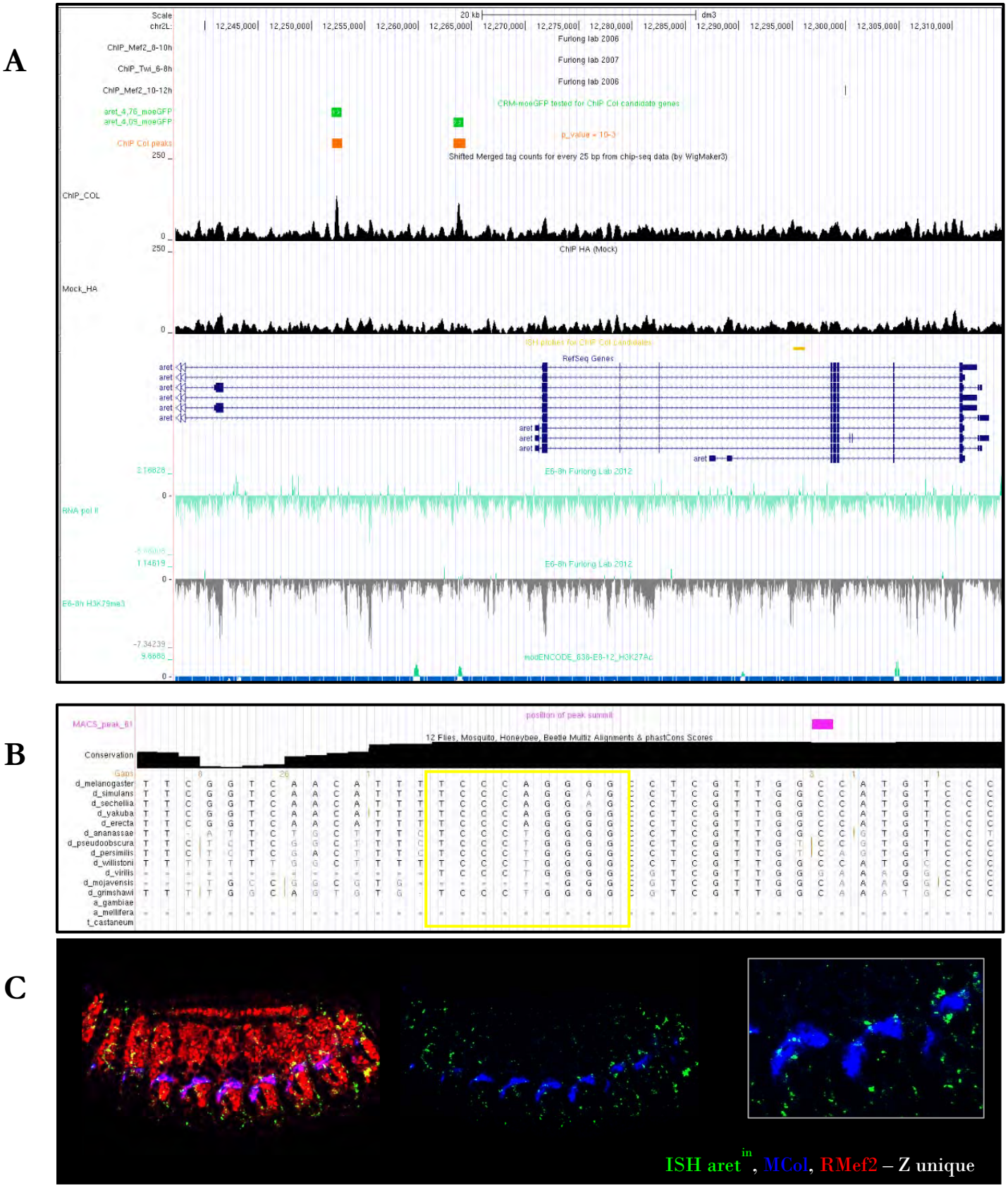


Fig. R19 – arrest/bruno (aret)

A. Annotation de la région génomique d'*aret*. La position du pic Col est notée (orange), ainsi que la position des sites de fixation de Twi (6-8h de développement) et Mef2 (8-10h et 10-12h) d'après des expériences de ChIP (E. Furlong), les profils des marques d'histones (E. Furlong : 6-8h et ModEncode 8-12h de développement) et la position du fragment (jaune) utilisé comme sonde en ISH. **B.** Conservation de la séquence nucléotidique du pic *aret* entre 12 espèces de drosophiles et position du motif Col (encadré jaune) relativement au sommet du pic (magenta). **C.** Hybridation *in situ* avec la sonde intronique *aret* (vert). Embryon st. 14. Le muscle DA3 est marqué par immuno-coloration avec l'anticorps anti-Col et l'ensemble des noyaux musculaires avec l'anticorps anti-Mef2.

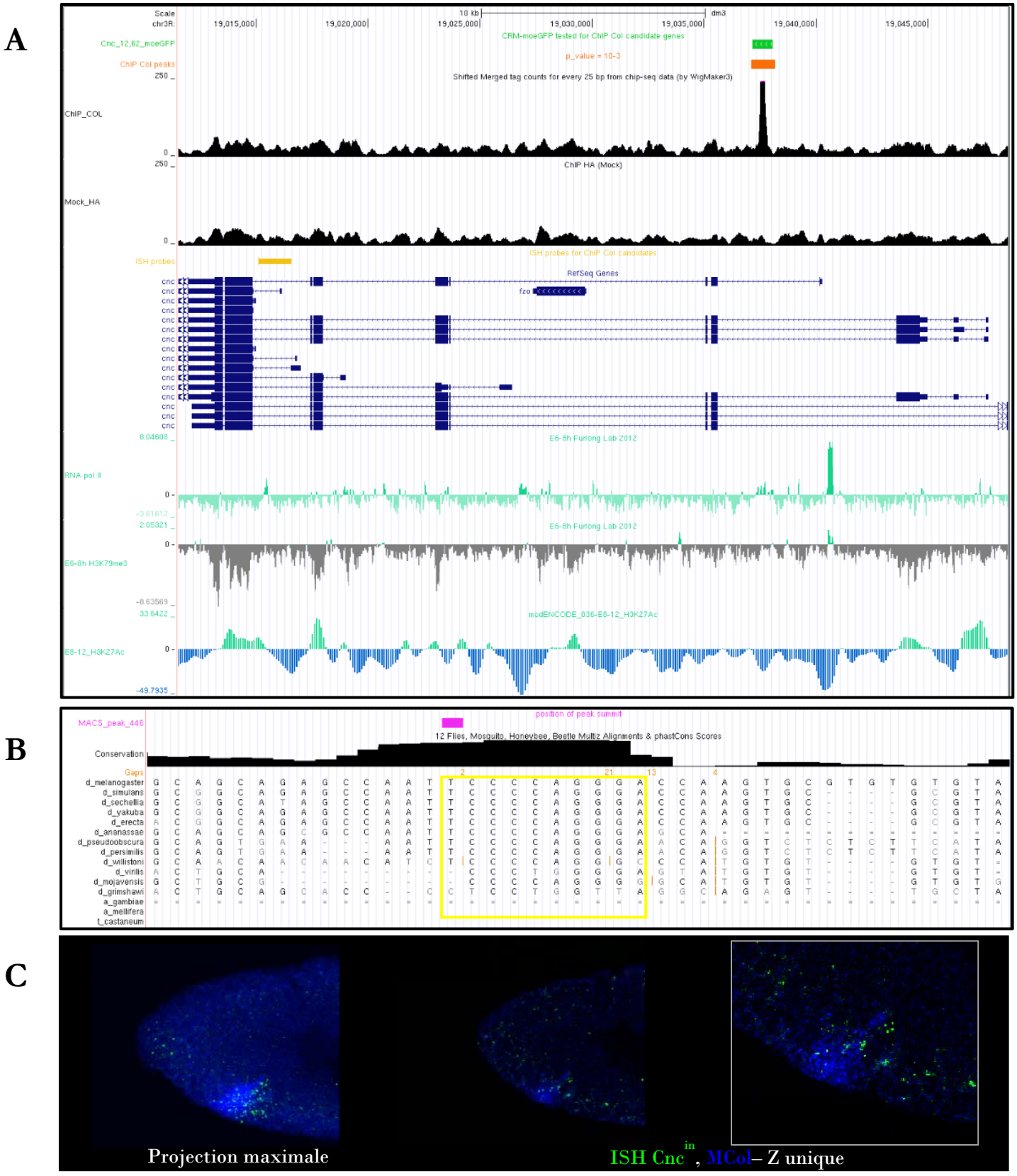


Fig. R20 – cap-n-collar (cnc)

A. Annotation de la région génomique de *cnc*. La position du pic Col est notée (orange), ainsi que la position des sites de fixation de Twi (6-8h de développement) et Mef2 (8-10h et 10-12h) d'après des expériences de ChIP (E. Furlong), les profils des marques d'histones (E. Furlong : 6-8h et ModEncode 8-12h de développement) et la position du fragment (jaune) utilisé comme sonde en ISH. **B.** Conservation de la séquence nucléotidique du pic *cnc* entre 12 espèces de drosophiles et position du motif Col (encadré jaune) relativement au sommet du pic (magenta). **C.** Hybridation *in situ* avec une sonde intronique *cnc* (vert). Embryon st. 10. L'immuno-coloration avec l'anticorps anti-Col permet de visualiser le patron d'expression de Col dans la tête.

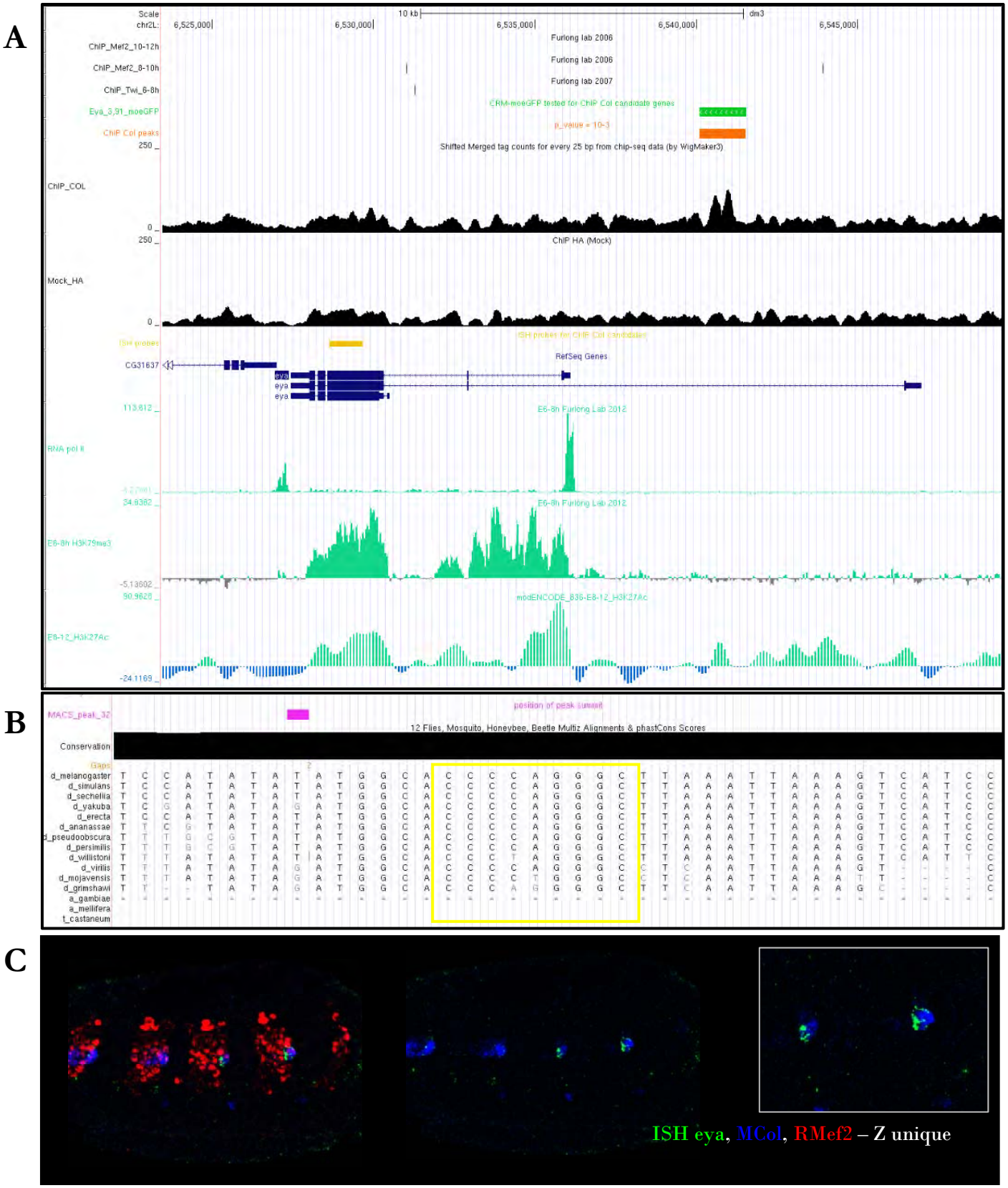


Fig. R21 – eye absent (eya)

A. Annotation de la région génomique d'*eya*. La position du pic Col est notée (orange), ainsi que la position des sites de fixation de Twi (6-8h de développement) et Mef2 (8-10h et 10-12h) d'après des expériences de ChIP (E. Furlong), les profils des marques d'histones (E. Furlong : 6-8h et ModEncode 8-12h de développement) et la position du fragment (jaune) utilisé comme sonde en ISH. **B.** Conservation de la séquence nucléotidique du pic *eya* entre 12 espèces de drosophiles et position du motif Col (encadré jaune) relativement au sommet du pic (magenta). **C.** Hybridation *in situ* avec la sonde exonique *eya* (vert). Embryon st. 11. Le progéniteur du muscle DA3 est marqué par immuno-coloration avec l'anticorps anti-Col et l'ensemble des noyaux musculaires avec l'anticorps anti-Mef2.

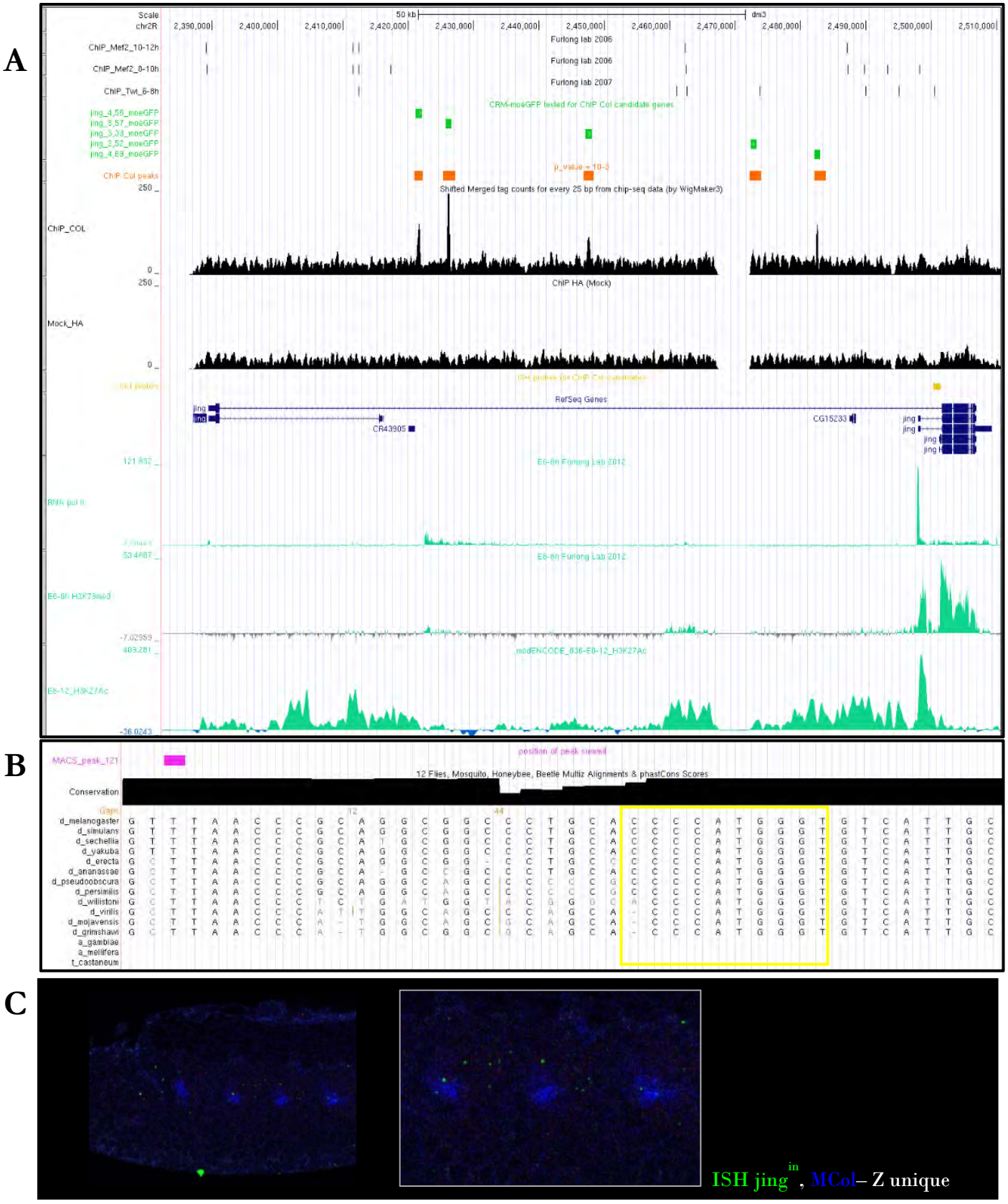


Fig. R22 – jing

A. Annotation de la région génomique de *jing*. La position du pic Col est notée (orange), ainsi que la position des sites de fixation de Twi (6-8h de développement) et Mef2 (8-10h et 10-12h) d'après des expériences de ChIP (E. Furlong), les profils des marques d'histones (E. Furlong : 6-8h et ModEncode 8-12h de développement) et la position du fragment (jaune) utilisé comme sonde en ISH. **B.** Conservation de la séquence nucléotidique du pic *jing* entre 12 espèces de drosophiles et position du motif Col (encadré jaune) relativement au sommet du pic (magenta). **C.**Hybridation *in situ* avec la sonde intronique jing (vert). Embryon st. 11. Le progéniteur du muscle DA3 est marqué par immuno-coloration avec l'anticorps anti-Col.

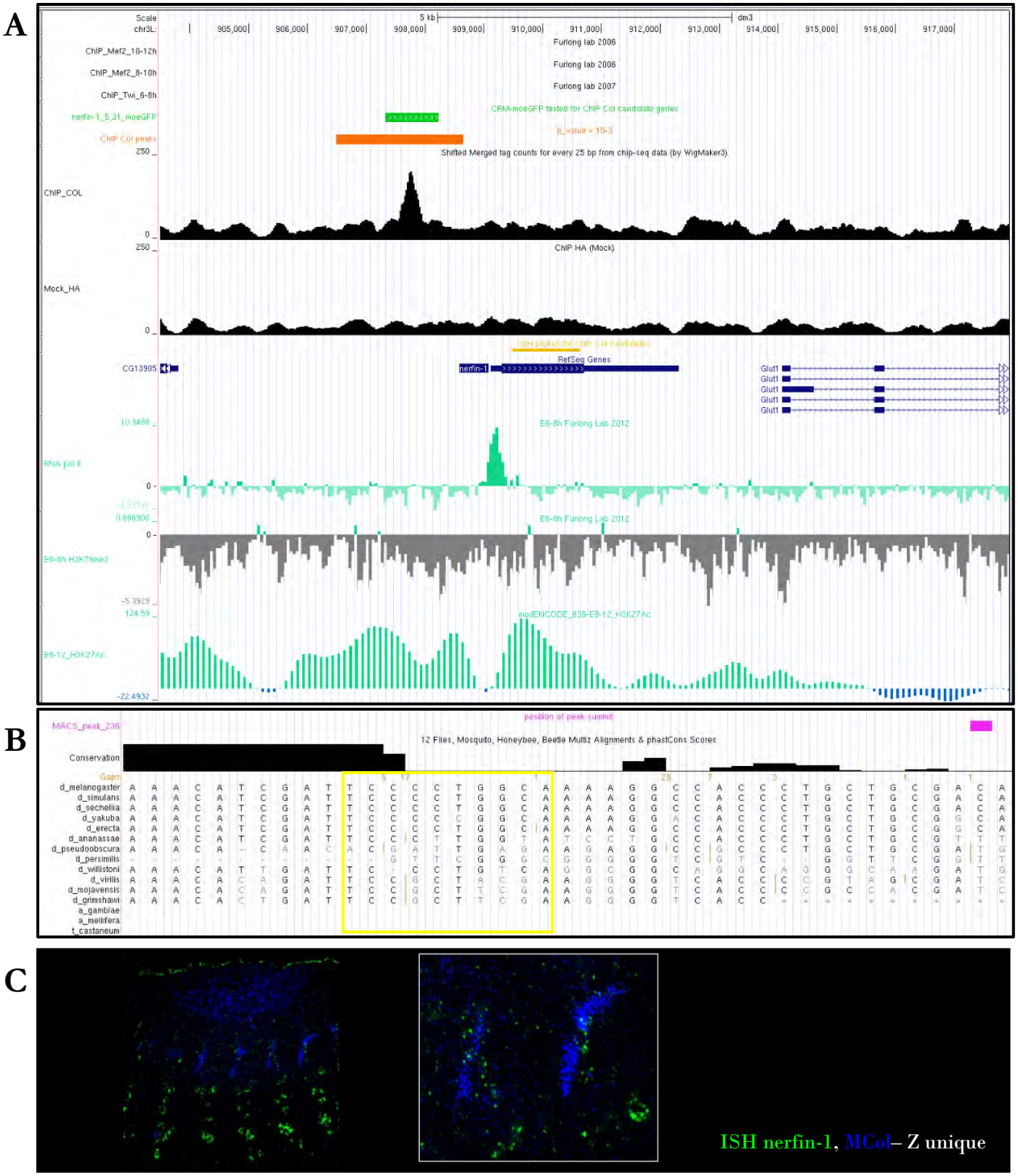


Fig. R23 – nervous finger 1 (nerfin-1)

A. Annotation de la région génomique de *nerfin-1*. La position du pic Col est notée (orange), ainsi que la position des sites de fixation de Twi (6-8h de développement) et Mef2 (8-10h et 10-12h) d'après des expériences de ChIP (E. Furlong), les profils des marques d'histones (E. Furlong : 6-8h et ModEncode 8-12h de développement) et la position du fragment (jaune) utilisé comme sonde en ISH. **B.** Conservation de la séquence nucléotidique du pic *nerfin-1* entre 12 espèces de drosophiles et position du motif Col (encadré jaune) relativement au sommet du pic (magenta). **C.**. Hybridation *in situ* avec la sonde nerfin-1 (vert). Embryon st. 14. Le muscle DA3 est marqué par immuno-coloration avec l'anticorps anti-Col.

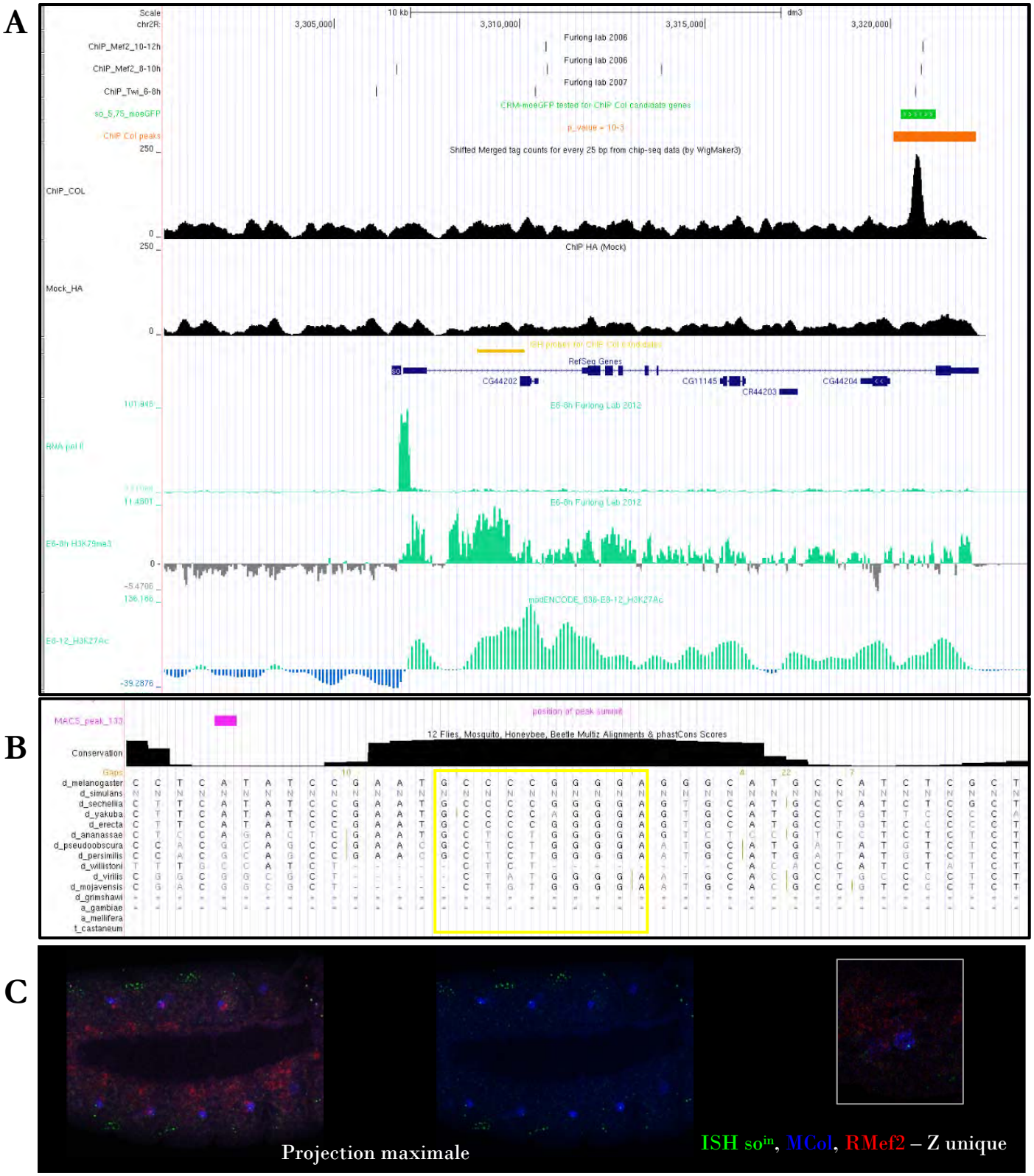


Fig. R24 – sine oculis (so)

A. Annotation de la région génomique de *so*. La position du pic Col est notée (orange), ainsi que la position des sites de fixation de Twi (6-8h de développement) et Mef2 (8-10h et 10-12h) d'après des expériences de ChIP (E. Furlong), les profils des marques d'histones (E. Furlong : 6-8h et ModEncode 8-12h de développement) et la position du fragment (jaune) utilisé comme sonde en ISH. **B.** Conservation de la séquence nucléotidique du pic *so* entre 12 espèces de drosophiles et position du motif Col (encadré jaune) relativement au sommet du pic (magenta). **C.** Hybridation *in situ* avec la sonde intronique so (vert). Embryon st. 11. Le progéniteur du muscle DA3 est marqué par immuno-coloration avec l'anticorps anti-Col et l'ensemble des noyaux musculaires avec l'anticorps anti-Mef2.

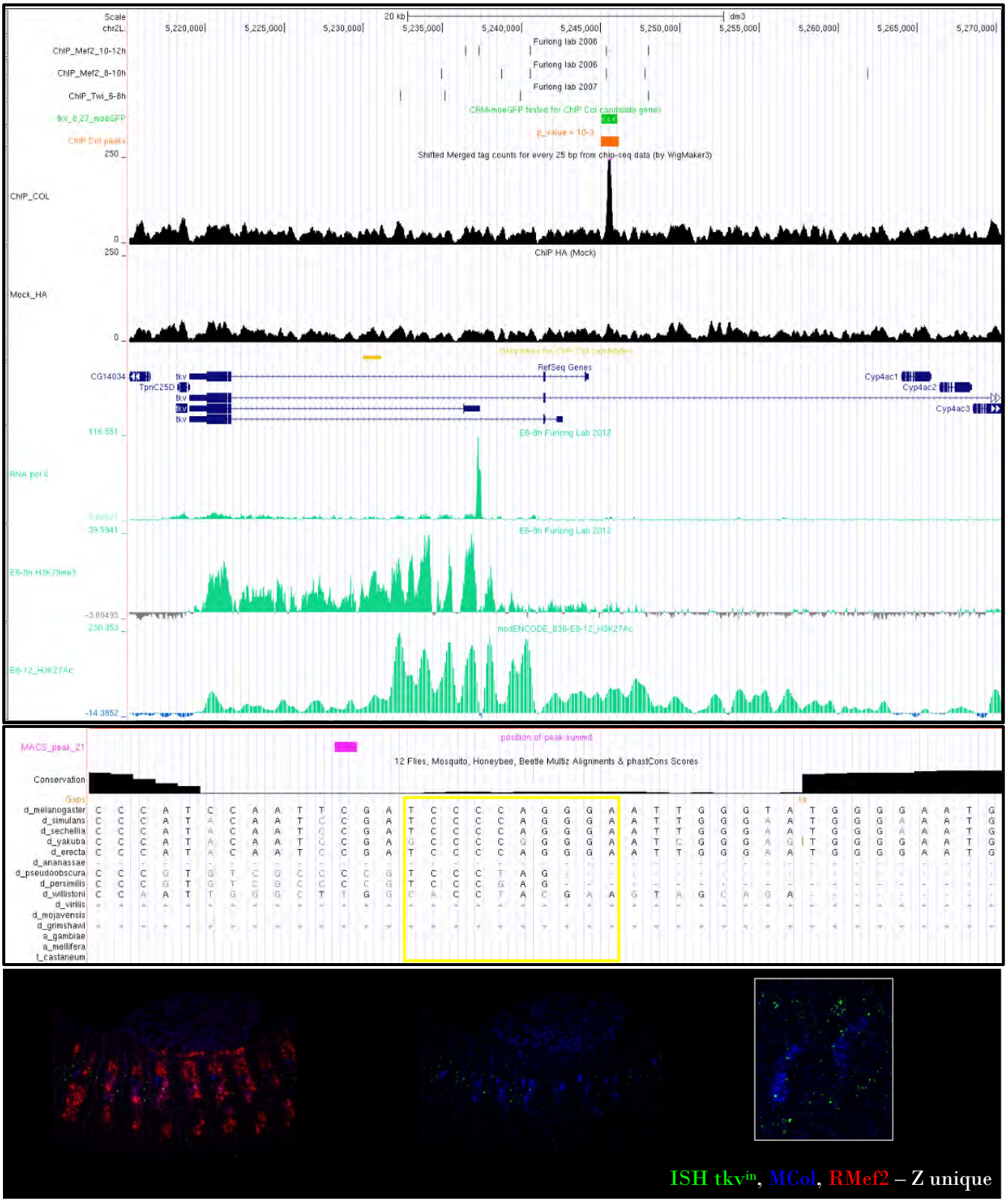


Fig. R25 – thickveins (tkv)

A. Annotation de la région génomique de *tkv*. La position du pic Col est notée (orange), ainsi que la position des sites de fixation de Twi (6-8h de développement) et Mef2 (8-10h et 10-12h) d'après des expériences de ChIP (E. Furlong), les profils des marques d'histones (E. Furlong : 6-8h et ModEncode 8-12h de développement) et la position du fragment (jaune) utilisé comme sonde en ISH. **B.** Conservation de la séquence nucléotidique du pic *tkv* entre 12 espèces de drosophiles et position du motif Col (encadré jaune) relativement au sommet du pic (magenta). **C.** Hybridation *in situ* avec la sonde intronique tkv (vert). Embryon st. 14. Le muscle DA3 est marqué par immuno-coloration avec l'anticorps anti-Col et l'ensemble des noyaux musculaires avec l'anticorps anti-Mef2.

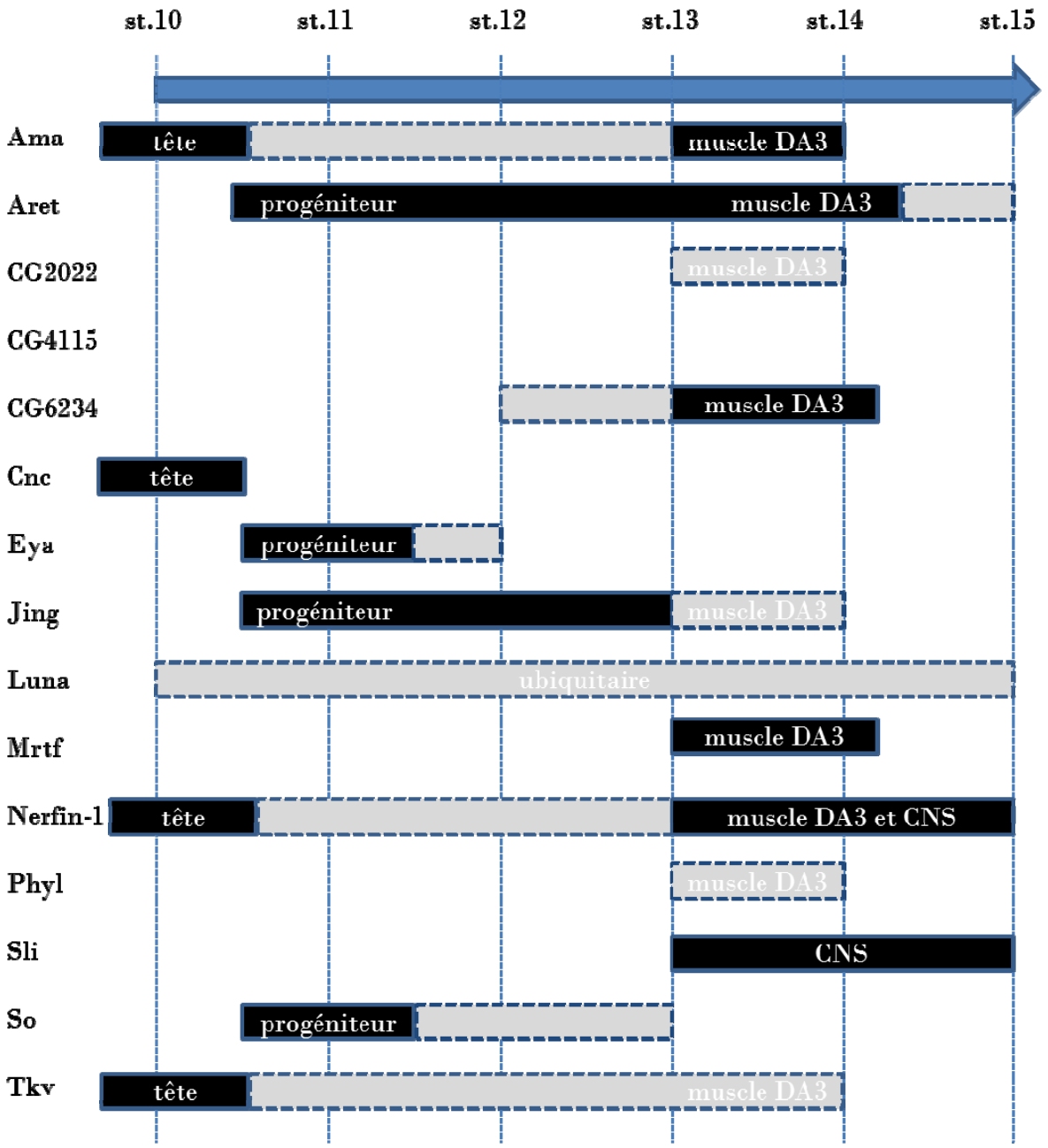


Fig. R26 –Transcription des gènes candidats dans les divers tissus exprimant Col à divers stades de l'embryogenèse.

Récapitulatif des fenêtres d'expression temporelle et spatiale des transcrits de gènes cibles potentielles de Collier étudiés, d'après leur profil d'expression en hybridation *in situ*. Les encadrés en pointillés grisés caractérisent une expression très faiblement détectée.

Rubin dans l'objectif d'étudier les gènes impliqués dans la formation du cerveau adulte de la drosophile (Pfeiffer et al., 2008), et mise à disposition de la communauté scientifique. Les lignées GMR-Gal4 contiennent des régions génomiques cis-régulatrices potentielles, sous la forme de fragments d'environ 3 kb, chevauchants sur environ 1 kb, clonés en amont de l'activateur de transcription Gal4. Des lignées GMR-Gal4 recouvrant le pic de fixation de Col étaient disponibles pour les gènes *eya* (GMR21C02 et GMR21A11), *Mrtf* (GMR56G06 et GMR60C01), *phyl* (GMR52C04, GMR74A01 et GMR51C08), *sli* (GMR32A06), et *so* (GMR15C09). Nous avons trouvé que les fragments *eya* (GMR21C02), *Mrtf* (GMR56G06), *phyl* (GMR51C08) et *so* (GMR15C09) activent le transgène UAS-lacZ dans le muscle DA3, avec plus ou moins de pénétrance. Par ailleurs, j'ai pu noter que les lignées *eya* (GMR21A11) et *phyl* (GMR52C04, GMR74A01 et GMR51C08) dirigent l'expression du rapporteur dans un sous-ensemble de neurones du SNC dont certains expriment Col. Ce résultat suggère qu'un certain nombre des sites de fixation de Col sélectionnés pourraient correspondre à une régulation par Col dans le SNC plutôt que dans le muscle. Par ailleurs, un fragment de *sli* (GMR32A06) dirige l'expression de LacZ dans d'autres muscles que le DA3, et plus particulièrement dans le DT1, muscle issu du groupe promusculaire exprimant Col (Enriquez et al., 2012), et difficile à identifier dans le cas des hybridations *in situ*. Ceci suggère que Col pourrait agir pour réguler ce gène au niveau du progéniteur du muscle DT1 en combinaison avec d'autres facteurs présents dans cette cellule, ou bien pour réprimer l'expression de *sli* dans le muscle DA3.

Bien qu'avec un léger décalage dans le temps, peut-être dû au délai engendré par le dispositif Gal4>UAS, on retrouve avec certaines lignées rapporteur l'expression notée dans le muscle DA3 par hybridation *in situ*. C'est le cas pour les gènes *eya*, *Mrtf*, et *so*. Les lignées GMR font aussi apparaître sous le contrôle des fragments sélectionnés de nouveaux territoires d'expression du lacZ non détectés lors des hybridations *in situ* ; ces territoires doivent être analysés plus en détail afin de déterminer s'il s'agit d'une expression reflétant l'expression du gène endogène ou si celle-ci apparaît suite à l'isolement du fragment de son contexte génomique. C'est le cas par exemple de l'expression observée dans certains neurones Col positif du SNC pour les lignées *eya* et *phyl*. A l'inverse, même si on retrouve une expression du rapporteur *sli* dans le SNC, je n'ai pas pu observer de colocalisation du transgène lacZ avec des neurones Col positif, la régulation de *sli* dans ces neurones pouvant dépendre d'un (ou de plusieurs !) autre CRM, d'autant plus que 4 pics de ChIP Col ont été associés à ce gène.

L'étude basée sur les GMR n'étant pas exhaustive, nous l'avons complétée par une deuxième approche plus systématique, avec des fragments d'ADN plus courts, centrés sur les pics de liaison *in vivo* de Col. Nous avons ainsi mis en œuvre le clonage d'un fragment d'ADN de 1kb

centré sur le site prédit de fixation *in vivo* de Col pour les 15 gènes candidats (*ama, aret, cg2022, cg4115, cg6234, cnc, eya, jing, luna, Mrtf, nerfin-1, phyl, sli, so et tkv*). Des lignées rapporteur moe ::GFP ont été construites, afin de déterminer les propriétés cis-régulatrices de chacun de ces fragments. <u>*Nota*</u> : les CRM sont nommés d'après le nom du gène candidat auquel ils sont associés et la hauteur du pic duquel ils ont été extraits ; ex. ama[3.25].

Une lignée témoin contenant le ^{L}CRM 2.6-0.9 de Col placé en amont du gène rapporteur moe ::GFP utilisé pour tester les CRM candidats récapitule le patron d'expression dans la tête et le muscle DA3 (Dubois et al., 2007), validant ainsi le choix de ce gène rapporteur. Sous le contrôle du ^{L}CRM 2.6-0.9, la GFP est exprimée à un niveau raisonnable. Cette construction nous permet de valider que le « squelette » du plasmide et son insertion dans le génome de la drosophile n'induit pas d'expression de la GFP en dehors de celle dirigée par le CRM cloné en amont. L'analyse des premières lignées de gènes candidats a permis de mettre en évidence la capacité des fragments testés à diriger une expression du rapporteur moe ::GFP dans l'embryon, avec des patrons d'expression bien distincts pour chacun des CRM.

Le CRM ama[3.25] donne une expression dans la tête aux stades 11-12, dans un domaine englobant le domaine de Col. D'abord plus restreint que le domaine Col, l'expression de la GFP s'élargit ensuite tandis que le domaine Col diminue. La mutation du site de fixation de Col prédit au sein de ce CRM abolit l'expression de la GFP dans la tête (cf. Annexe 4), suggérant qu'une régulation directe de Col active le gène *ama* dans ce domaine. Par contre, le CRM ama[3.25] ne contient pas l'élément cis-régulateur responsable de l'expression d'*ama* dans le muscle DA3, telle qu'observée par hybridation *in situ*.

Les CRM aret[4.76] et tkv[8.27] montrent aussi une expression de la GFP dans la tête au stade 11, dans le domaine Col positif pour aret[4.76] et un domaine plus large pour tkv[8.27]. A noter, une expression des gènes *aret* et *tkv* n'a pu être détectée dans ce tissu par hybridation *in situ*. La mutation des sites Col prédits dans les CRM aret[4.76] et tkv[8.27] ne modifie pas l'expression de la GFP dans la tête, montrant que cette expression ne dépend pas de la fixation de Col.

Les CRM aret[4.76], cg6234[7.45], eya[3.91], Mrtf[3.35], phyl[22.41], phyl[4.43], so[5.75] et tkv[8.27] dirigent l'expression de la GFP dans le muscle DA3, à partir du stade progéniteur (Mrtf[3.35], phyl[4.43] et so[5.75]) ou au cours de sa différenciation (aret[4.76], cg6234[7.45], eya[3.91], phyl[22.41] et tkv[8.27]), conformément à l'expression des transcrits observés par hybridation *in situ*. Cette expression s'étend souvent à plusieurs muscles, dorso-latéraux majoritairement (sauf Mrtf[3.35] qui semblent exprimé dans tous les muscles au stade 15), mais avec cependant des patrons différents. Ainsi, en plus du muscle DA3, le CRM aret[4.76] parait être spécifiquement actif dans le muscle LL1, le CRM eya[3.91] dans le muscle DT1, les CRM

phyl[4.43] et [22.41] dans le muscle DO4 (ou DT1), le CRM so[5.75] dans le muscle DO5, et le CRM tkv[8.27] dans le muscle DA2. Dans certains cas, l'identité de ces muscles, simplement inférée d'après leur morphologie et leur position, mérite d'être confirmée par des co-marquages avec des marqueurs spécifiques. La mutation du site Col (ou des sites dans le cas du CRM eya[3.91]) au sein de ces modules a des effets variables suivant les CRM. Dans le cas du CRM aret[4.76], on n'observe pas de modification significative du profil d'expression de la GFP dans le muscle DA3 lorsque le site de fixation prédit pour Col est muté ; de même, la mutation du CRM Mrtf[3.35] n'entraine qu'une diminution partielle de l'expression du rapporteur dans les muscles dorso-latéraux (à confirmer). Par contre, l'effet de la mutation du motif Col est plus évident pour les CRM cg6234[7.45], eya[3.91], phyl[22.41] et [4.43] et tkv[8.27], pour lesquels on observe une perte de l'expression de la GFP dans le muscle DA3 (cf. Annexes 4, 5 et 7) lorsque le site est muté. Col pourrait donc réguler de manière directe l'expression des gènes associés dans le muscle DA3. Cette perte d'expression concerne également les autres muscles dorso-latéraux dans le cas des CRM de *phyl* et *tkv*, suggérant une régulation précoce de ces CRM, correspondant à l'expression transitoire de Col dans les 3 progéniteurs dorso-latéraux (Enriquez et al., 2012). La mutation du motif Col au cœur du CRM so[5.75] n'abolit pas l'expression de la GFP, mais au contraire semble plutôt provoquer un élargissement de son domaine d'expression. Ce résultat très préliminaire reste à confirmer.

Comme les CRM correspondants de la collection GMR-gal4, les CRM eya[3.91] et sli[3.98] activent l'expression de la GFP dans le SNC de l'embryon. C'est le cas également des CRM aret[4.76] et tkv[8.27]. Le recouvrement entre les neurones Col+ et la GFP est cependant très partiel et reste à étudier plus en détail. Mis à part dans le cas du CRM sli[3.98] pour lequel la mutation du site Col entraîne une forte diminution de l'expression du gène rapporteur dans le SNC, cette mutation n'affecte pas de manière significative l'expression de la GFP.

Le CRM jing[3.91] est actif dans un progéniteur Col positif (très probablement celui du DO3/DT1) puis dans ces muscles. Lorsque le site prédit de fixation de Col est muté, l'expression de la GFP est activée de manière stochastique dans le muscle DA3 et perdue dans les autres muscles DL. Cette expression devient également observable dans la tête, dans un domaine recouvrant le domaine Col (cf. Annexe 6). Ce résultat suggère que Col pourrait agir comme répresseur, une hypothèse évoquée à ce jour dans un seul article et concernant la protéine UNC-3 de nématode.

Le CRM cnc[12.62] dirige l'expression de la GFP dans la tête, dans le domaine d'expression de Col d'abord (à laquelle s'ajoute une expression plus postérieure), puis dans un domaine distinct, postérieur et limitrophe au domaine Col. Cette expression pourrait correspondre aux

Candidat	Transcription dans les domaines *col+*	Expression des lignées Janelia (CRM-gal4>UAS-lacZ)	Expression des CRM_ChIP moe::GFP : superposition avec Col	Expression des CRM_ChIP-mut-moe::GFP	phénotype mutant	phénotype déficience
ama *ama[3,25]*	tête, muscle DA3 (st. 13-15) et glande lymphatique	-	tête (domaine plus large que le domaine Col+) *–expression probable dans un PC/FC Col+*	Perte de l'expression dans la tête *et dans le mésoderme*	Perte du muscle DA3 dans certains segments *(mutant DN uniquement)*	-
aret *aret[4,76]*	muscle DA3 : progéniteur> st.15	-	tête (domaine englobant le domaine Col+) ; muscles dorso-latéraux dont le DA3 (st. 12-14) ; *SNC (peu de neurones Col+)*	pas de modifications significatives par rapport au CRM non muté	-	Expression *de novo* de Col dans le muscle DA2 avec transformation en muscle DA3 dans certains segments
CG2022 *cg2022[8,68]*	muscle DA3 (st. 13-14) mais signal très faible	-	pas de recouvrement	pas de recouvrement	-	-
CG4115 *cg4115[3,38]*	aucun signal	-	pas de recouvrement	-	-	pas de phénotype musculaire
CG6234 *cg6234[7,45]*	muscle DA3 (st. 13-14) - neurones MD inf.	-	tête, muscle DA3 (st.13-14 - segments thoraciques)	toujours exprimé dans le mésoderme mais plus dans le muscle DA3	-	-
c'n'c *cnc[12,62]*	tête	-	tête : domaine englobant le domaine Col (recouvrement temporaire)		Perte de structures mandibulaires et labiales (Mohler et al., 1995)	-
eya *eya[3,91]*	muscle DA3 (progéniteur) - SNC	*GMR21C02 :* muscle DA3 (et DT1 ?) (st.13-15) et SNC (quelques neurones Col+) - GMR21A11 : SNC (quelques neurones Col+)	muscle DA3 (st.13-15) et d'autres muscles dorso-latéraux (DT1 ?) - SNC dont quelques neurones Col+	perte de l'expression dans le muscle DA3 mais pas dans le SNC (neurones Col + ?)	Défaut d'identité et perte de certains muscles (Liu et al., 2009)	Disparition du marquage Col dans le DA3
jing *jing[4,69]*	muscle DA3 (progéniteur) mais signal faible	-	progéniteur Col+ (DO3/DT1 ?) puis muscles dérivés - glande lymphatique	Perte de l'expression dans les muscles et expression *de novo* dans la tête et le muscle DA3	-	transformation du DA3 en DA2 avec perte du marquage Col
luna *luna[4,37]*	ubiquitaire	-	-	-	-	Défaut d'attachement des muscles
Mrtf *Mrtf[3,35]*	muscle DA3 (st.13-15)	GMR56G06 : plusieurs muscles dont le DA3 (st. 13-15)	progéniteurs Col+ puis presque tous les muscles (st.14-16) dont le DA3	Diminution probable du niveau d'expression dans les muscles DL.	-	-
nerfin-1 *nerfin-1[5,31]*	tête, muscle DA3 (faible - st. 13-14), SNC	-	tête et SNC		Transformation du DA3 en DA2 *(Trip RNAi)*	-
phyl *phyl[22,41]*	signal très faible	GMR52C04 : neurones MD sup., SNC - GMR74A01 : SNC - GMR51C08 : muscle DA3 mais peu pénétrant, SNC	muscle DA3 (st. 14-16) (et d'autres muscles dorso-latéraux)	perte d'expression dans le muscle DA3 – expression dans le SNC	Perte de certains muscles dont le DO4 et le LL1 (Artero et al., 2003)	-
phyl *phyl[4,43]*	Cf. ci-dessus	Cf. ci-dessus	muscle DA3 (st. 12-15) (et d'autres muscles dorso-latéraux)	perte d'expression dans le muscle DA3 – expression dans le SNC	Cf. ci-dessus	-
sli *sli[3,98]*	SNC	GMR32A06 : muscle DT1 probablement et SNC (neurones Col-)	*SNC (mais dans très peu de neurones Col+)*	Forte diminution de l'expression dans le SNC	Défaut d'attachement des muscles(Kramer et al., 2001)	Défaut d'attachement des muscles
so *so[5,75]*	muscle DA3 (progéniteur) - quelques cellules du SNC	GMR15C09 : muscle DA3 (st.13-15) et d'autres cellules du mésoderme	progéniteurs Col+ ; muscle DA3 et d'autres muscles dorso-latéraux (DO5 ?)	Élargissement du domaine d'expression (muscle DA2 ?, mésoderme viscéral…)	-	transformation du DO5 en DA3
tkv *tkv[8,27]*	Signal faible tête - muscle DA3 (st. 13-14)	-	tête - plusieurs muscles dorso-latéraux dont le DA3 - glande lymphatique - *quelques neurones du SNC (mais non Col+)*	Tête – glande lymphatique – perte d'expression dans le muscle DA3 à confirmer	-	-

Tableau. R27 – Récapitulatif des données d'expression de 15 gènes candidats (colonne de gauche)

De gauche à droite : hybridation *in situ*, analyse de lignées CRM-gal4 Janelia Farm, analyse de CRM définis par ChIPseq avant et après mutation du site de fixation de Col (cf. Annexe 2), phénotypes des mutants ou des déficiences associés aux gènes candidats (en gris : données de la littérature ; en noir : données non publiées du laboratoire).

données déjà acquises pour *cnc* qui, d'abord cible de Col (Crozatier et al., 1999) participe ensuite à la répression de l'activité (autorégulatrice ?) de Col dans le domaine le plus postérieur, correspondant à la partie antérieure du segment mandibulaire (Ntini and Wimmer, 2011). Nos expériences de ChIP suggèrent que Col régule de manière directe l'expression de *cnc* dans la tête. Afin de confirmer cette régulation directe, la mutation du site Col prédit dans le CRM cnc[12.62] est en cours d'analyse.

Enfin deux des CRM testés, les CRM cg2022[8.68] et cg4115[3.38] dirigent l'expression de la GFP dans des domaines ne présentant aucun recouvrement avec les domaines d'expression de Col. Le CRM cg2022[8.68] est actif dans un groupe de cellules mésodermiques dans chaque segment, distinct du domaine d'expression de Col, et la mutation du motif Col prédit au sein de ce CRM ne modifie pas ce profil d'expression de manière évidente. La présence d'un pic de fort enrichissement à cette position sur le génome pourrait refléter soit la fixation de Col au niveau d'un autre site, qui reste à identifier, soit une fixation sans lien avec une régulation transcriptionnelle (« séquestration ? »). Le CRM cg4115[3.38] s'exprime uniquement dans le mésoderme viscéral.

L'ensemble des données d'expression obtenues pour les CRM de 15 gènes cibles candidats sont récapitulées dans le tableau R27. On peut déjà conclure qu'ils correspondent à différents niveaux de régulation par Col, et dans plusieurs tissus où Col est exprimé : tête, progéniteurs des muscles dorso-latéraux, muscle DA3, SNC.

L'objectif ultime de mon projet était d'identifier des gènes cibles directes de Collier impliqués dans la réalisation de l'identité musculaire. Parmi les gènes cibles candidats validés, on note la présence de plusieurs facteurs de transcription, suggérant qu'il ne s'agit pas à proprement parler d'effecteurs de l'identité mais de « relais » et que Collier pourrait agir dans une étape amont de régulation. Parmi ces facteurs, on trouve par exemple *eya*, *so*, *Mrtf* ou possiblement, *nerfin-1*…

Un autre gène cible validé dans notre approche est le gène *cnc* qui code lui-aussi pour un FT. Il était déjà connu que Col régule son expression dans la partie antérieure du segment gnathal mandibulaire. Nos données montrent que cette régulation est directe. Un autre gène cible de Col dans le segment mandibulaire que nous avons identifié est le gène *ama*. Le panel des gènes cibles de Col s'enrichit ainsi dans plusieurs domaines.

Zoom sur quelques candidats

L'analyse des CRM pour les 15 candidats cités ci-dessus se poursuit et doit s'enrichir très prochainement d'une analyse comparée avec les mêmes CRM dont le motif de liaison prédit pour Col a été muté (cf. tableau des mutations en Annexe 2). En attendant, 5 des candidats ont retenu plus particulièrement notre attention : *cnc* et *ama* dans la tête, *eya*, *so* et *Mrtf* dans le muscle DA3. Voici d'ores et déjà quelques données bibliographiques relatives à ces gènes.

Le gène *cap'n'collar* ou ***cnc*** (CG 43286, FBgn 0262975) est un gène embryonnaire-létal qui code pour plusieurs isoformes d'un facteur de transcription a domaine riche en leucine (basic leucine zipper). Il est impliqué dans la formation du squelette céphalo-pharyngal de l'embryon de drosophile. Aucun orthologue direct n'a été décrit chez les mammifères. Le laboratoire a montré en 1999 que Col est spécifiquement requis pour l'expression de *cnc* dans la partie la plus antérieure du segment mandibulaire et la région postérieure du segment intercalaire (Crozatier et al., 1999). Plus récemment Ntini et Wimmer (Ntini and Wimmer, 2011) ont postulé une interaction entre les protéines Col et Cnc isoforme B dans cette région pour le positionnement correct de l'expression de *hh* dans le segment intercalaire, qui est sous la dépendance de Col. Cnc fait partie des facteurs de transcription étudiés par ChIPseq génomique par ModEncode (Nègre et al., 2011).

Le gène *amalgam* ou **ama** (CG2198, FBgn0000071) est un gène qui code pour une protéine d'adhésion membranaire à domaines immunoglobuline « extracellulaires » et est exprimé dans de nombreux tissus au cours du développement embryonnaire (http://insitu.fruitfly.org/cgi-bin/ex/report.pl?ftype=3&ftext=LD39923). Seul son rôle dans la guidance axonale et les interactions entre neurones a été décrit.

Le gene *eyes absent* ou ***eya*** (CG9554, FBgn0000320) est un gène embryonnaire-létal qui code pour une protéine nucléaire à activité tyrosine phosphatase (Rayapureddi et al., 2005; Tootle et al., 2003) et modifie l'activité des FTs auxquels elle se lie, en particulier le FT à homéodomaine Sine Oculis (So), un membre de la famille des protéines Six. Quatre orthologues d'*eya* ont été identifiés chez l'homme (HsapEya1-4) et la souris (MmusEya1-4). Un rôle des protéines Eya1,2 et Six1,4 dans la myogenèse a été décrite chez la souris (Grifone et al., 2007). Cependant, bien que l'expression d'eya dans le mésoderme embryonnaire de drosophile ait été bien documentée (http://insitu.fruitfly.org/cgi-bin/ex/report.pl?ftype=1&ftext=CG9554), son rôle potentiel dans ce tissu reste à étudier en détail, contrairement à son rôle dans des lignages neuronaux

spécifiques, dans l'œil adulte et les gonades. *eya* a récemment été identifié comme une cible directe de Tin dans le mésoderme (Liu et al., 2009). La même étude a montré que des embryons mutants nuls *eya* ont des défauts des muscles somatiques, en particulier de muscles dorsaux et latéraux avec une pénétrance variable. Ce phénotype est similaire à celui observé dans des embryons mutants pour le gène D-six4, un paralogue de *so* exprimé de manière similaire à *eya* dans le mésoderme embryonnaire (Clark et al., 2007; Liu et al., 2009), contrairement à *so* (mais voir plus bas). Col agit en amont d'*eya* (et en coopération avec) dans la spécification de neurones peptidergiques Tvb du SNC (Baumgardt et al., 2007) mais il reste à établir si ce contrôle est direct. Le fragment lié par Col dans nos expériences de ChIP se situe à environ 4.5 kb en amont de l'isoforme A du gène *eya* (chr2L:6540105-6541545), légèrement en amont du CRM eya_meso enhancer décrit par (Liu et al., 2009) (chr2L :6530903-6531807).

Le gene *sine oculis*, ou ***so*** (CG 11121, FBgn0003460) code pour un facteur de transcription à homéodomaine de la famille Six interagissant avec Eya dans la formation de l'œil adulte. Il possède entre 4 et 5 orthologues chez les mammifères, six1-4 et 6. Aucune expression ou rôle dans le mésoderme à l'origine des muscles du tronc n'ont encore été décrits (http://insitu.fruitfly.org/cgi-bin/ex/report.pl?ftype=3&ftext=GH15741).

Le gène ***Mrtf*** (Myocardin-related transcription factor, CG32296, FBgn0052296) code pour un FT à domaine SAP. Le nom MRTF vient de la présence d'un domaine de forte homologie avec le domaine SAP (SAF-A/B, Acinus, PIAS) des MRTF humains, des FT qui interagissent avec les SRF (serum responsive factor) ; cette interaction est conservée chez la drosophile (Han et al., 2004). La transcription de Mrtf dans le mésoderme embryonnaire de la drosophile a été décrit (http://insitu.fruitfly.org/cgi-bin/ex/report.pl?ftype=3&ftext=AT27794) mais aucune fonction dans le mésoderme n'a été décelée de l'analyse de mutant nuls générés par recombinaison homologue. Par contre la surexpression d'une forme dominante-négative dans tout le mésoderme affecte la formation des muscles dorsaux et la surexpression d'une forme « hyperactive » affecte le patron musculaire. A suivre !

Avant de poursuivre une analyse fonctionnelle de ces « nouveaux gènes », soit dans le mésoderme (*eya*, *so*, *Mrtf*) ou le segment mandibulaire (*ama*), il restait à confirmer que l'expression de ces gènes dépend de la liaison de Collier au site identifié par ChIPseq. Cette expérience est en cours et les résultats devraient être inclus dans la version définitive présentée le jour de ma soutenance.

II.3 – Contrôle combinatoire de l'identité musculaire par Collier et Nautilus

La présence de gènes cibles (potentiels) de Col dans le lignage DA3 codant pour d'autres FTs exprimés dans d'autres lignages renforce l'hypothèse d'un contrôle de l'identité musculaire par des combinaisons de FTs exprimées spécifiquement dans chacune des FC et des régulations croisées entre ces facteurs. Une situation semblable a été décrite où Collier contribue au code FT nécessaire à la spécification de neurones peptidergiques dans le SNC (Baumgardt et al., 2007) et dans la tête où Collier est requis pour la spécification des segments mandibulaire et intercalaire de l'embryon(Crozatier et al., 1999; Ntini and Wimmer, 2011). Dans les deux cas, les fonctions de Col incluent la régulation et l'interaction avec d'autres facteurs. Cependant, si l'hypothèse du contrôle combinatoire de l'identité musculaire par des facteurs de transcription identitaires (FTi) a été proposée dès 1993 (Bate and Rushton, 1993), cette hypothèse n'avait jamais été testée directement, la totalité des études sur les FTi étant centrées sur un FTi à la fois. Les phénotypes mutants de *col* et *nautilus* (Nau est l'orthologue de MyoD chez les mammifères) montrant des défauts spécifiques de formation du muscle DA3 (Crozatier and Vincent, 1999; Keller et al., 1998) joints à l'observation que Col et Nau régulent différemment l'expression de *col* dans le lignage DA3 à partir du stade FC (Dubois et al., 2007), posaient la question des rôles respectifs et combinés de ces 2 FTi dans un même lignage musculaire. En complément de ma recherche des gènes cibles de Col, j'ai donc contribué à l'étude de cette question, sous la direction de Jonathan Enriquez, alors en 4ème année de thèse. Les résultats ont été publiés dans l'article intitulé « Combinatorial coding of Drosophila muscle shape by Collier and Nautilus » (Enriquez et al., 2012) qui est résumé ci-dessous.

II.3.1 – Le groupe promusculaire « Collier » donne naissance à plusieurs progéniteurs spécifiés séquentiellement.

Dans cet article, nous avons d'abord montré que Col et Nau étaient non seulement co-exprimés dans le progéniteur à l'origine des muscles DA3 et DO5 mais également dans 2 autres progéniteurs issus du même groupe promusculaire : le progéniteur des muscles DT1 et DO3 et le progéniteur des muscles LL1 et DO4. Ces différents progéniteurs sont sélectionnés de manière séquentielle à partir du groupe promusculaire Col et chacun maintient l'expression d'une combinaison particulière de FTi qui résulte en partie d'une régulation croisée entre ces FTi. Cette observation d'une séquence temporelle précise de sélection de différents progéniteurs est la

première documentée dans le mésoderme. Les implications de cette sélection séquentielle seront par la suite approfondies par Hadi Boukhatmi (Boukhatmi et al., 2012).

II.3.2 – Collier et Nautilus : complémentarité dans la construction de l'identité musculaire

Nous avons ensuite repris l'analyse phénotypique détaillée des mutants *col* et *nau*, et examiné des combinaisons double-mutantes. Cette analyse a permis de montrer que, bien qu'exprimé seulement transitoirement jusqu'au stade progéniteur, Col était nécessaire à la spécification de l'identité de tous les muscles dorsaux-latéraux issus du groupe promusculaire Col, confirmant l'importance fonctionnelle de l'établissement du code FT à ce stade. En ce qui concerne le muscle DA3, dans lequel l'expression de Col est maintenue, l'analyse de mutants *col* amorphe et hypomorphe a montré que l'orientation finale de ce muscle résultait d'un processus d'attachement à l'épiderme en 2 étapes : 1) un attachement transitoire à 3 sites sur l'épiderme, 2 sites antérieurs (un ventral et un dorsal) et un site postérieur, 2) une étape dite de « résolution » sélectionnant les 2 sites d'attachement définitifs. En absence de Col, l'étape de résolution est défective et le muscle DA3 adopte le site antérieur dorsal plutôt que ventral, ce qui résulte en une orientation similaire au muscle dorsal DA2. Dans des mutants *col* hypomorphes, le muscle DA3 garde les 3 sites et adopte une forme triangulaire, confirmant que c'est bien l'étape de résolution qui est défective. Reste à déterminer si cette étape de résolution de l'orientation dépendant du code FTi est généralisable à d'autres lignages musculaires. L'étude de mutants *nau* a permis de mettre en évidence un double rôle de ce facteur : un contrôle « générique » de la taille des muscles, puisque toutes les fibres musculaires de l'embryon sont plus fines en absence de Nau, et un rôle de FTi dans des lignages spécifiques tels que le lignage DA3. En absence de Nau, le DA3 est transformé en son lignage jumeau « DO5 », un phénotype distinct du phénotype mutant *col*. En conclusion, l'analyse de l'expression et des fonctions de Col et Nau dans le lignage DA3 a confirmé le contrôle combinatoire de l'identité musculaire par les FTi exprimés aux stades PC et FC. L'analyse future des CRM des gènes réalisateurs de l'identité doit permettre de déterminer à quel niveau se situe la régulation transcriptionnelle par ces 2 facteurs.

Developmental Biology 363 (2012) 27–39

Contents lists available at SciVerse ScienceDirect

Developmental Biology

journal homepage: www.elsevier.com/developmentalbiology

Combinatorial coding of *Drosophila* muscle shape by Collier and Nautilus

Jonathan Enriquez [b], Mathilde de Taffin [a], Michèle Crozatier [a], Alain Vincent [a,*], Laurence Dubois [a,*]

[a] Université de Toulouse 3, Centre de Biologie du Développement, UMR 5547 CNRS and IFR 109, 118 route de Narbonne, F-31062 Toulouse cedex 09, France
[b] Department of Biochemistry and Molecular Biophysics, Columbia University, 701 W. 168th St., HHSC 1104 New York, NY 10032, USA

ARTICLE INFO

Article history:
Received for publication 24 August 2011
Revised 9 December 2011
Accepted 10 December 2011
Available online 20 December 2011

Keywords:
Collier/EBF
Nautilus/MyoD
Myogenesis
Muscle progenitors
Muscle attachment sites

ABSTRACT

The diversity of *Drosophila* muscles correlates with the expression of combinations of identity transcription factors (iTFs) in muscle progenitors. Here, we address the question of when and how a combinatorial code is translated into muscle specific properties, by studying the roles of the Collier and Nautilus iTFs that are expressed in partly overlapping subsets of muscle progenitors. We show that the three dorso-lateral (DL) progenitors which express Nautilus and Collier are specified in a fixed temporal sequence and that each expresses additionally other, distinct iTFs. Removal of Collier leads to changes in expression of some of these iTFs and mis-orientation of several DL muscles, including the dorsal acute DA3 muscle which adopts a DA2 morphology. Detailed analysis of this transformation revealed the existence of two steps in the attachment of elongating muscles to specific tendon cells: transient attachment to alternate tendon cells, followed by a resolution step selecting the final sites. The multiple cases of triangular-shaped muscles observed in *col* mutant embryos indicate that transient binding of elongating muscle to exploratory sites could be a general feature of the developing musculature. In *nau* mutants, the DA3 muscle randomly adopts the attachment sites of the DA3 or DO5 muscles that derive from the same progenitor, resulting in a DA3, DO5-like or bifid DA3-DO5 orientation. In addition, *nau* mutant embryos display thinner muscle fibres. Together, our data show that the sequence of expression and combinatorial activities of Col and Nau control the pattern and morphology of DL muscles.

Introduction

"It takes 47 different muscles to frown and only 13 to smile". This popular saying illustrates the diversity of muscles needed for simple coordinated movements. The genetic and molecular mechanisms that build up this diversity remain, however, poorly understood. The *Drosophila* larval musculature, made of a stereotyped array of around 30 different muscles in each hemi-segment, is an ideal model to study this process (Bate, 1993). Each individual muscle is composed of a single multinucleated syncytial fibre, characterised by its position and orientation with respect to the dorso-ventral (D/V) and antero-posterior (A/P) axes, size and number of nuclei, epidermal attachment sites and ultimately, innervation. These unique properties are collectively referred to as muscle identity. The *Drosophila* muscle pattern is seeded by a special class of myoblasts, called founder cells (FCs), which display the unique property of being able to undergo multiple rounds of fusion with another class of myoblasts, the Fusion Competent Myoblasts (FCMs). FCs originate from the asymmetric division of progenitors cells (PCs), selected from equivalence groups of myoblasts, called promuscular clusters, via Notch (N)-mediated lateral inhibition and short range receptor tyrosine kinase (RTK) signalling (Buff et al., 1998; Carmena et al., 1995). Promuscular clusters are themselves specified at fixed positions within the somatic mesoderm, in response to positional information issued from the ectoderm and provided by long range Wingless and Dpp signalling (Carmena et al., 1998).

Muscle diversity is first revealed by the unique patterns of "identity" transcription factors (iTFs) that accompany progenitor segregation (reviewed in Frasch, 1999; Tixier et al., 2010). The expression of a particular iTF persists in only one of the two sibling FCs derived from the division of a progenitor, such that the properties unique to each muscle could reflect the specific combination of iTFs expressed in its PC and maintained in its FC (Bate and Rushton, 1993; Baylies et al., 1998; Bourgouin et al., 1992). However, the concept of combinatorial control of muscle identity relies, so far, upon the compared expression and function of individual iTFs. Here, we address the question of when, and how the cumulative and/or combinatorial activity of several iTFs does act during the muscle specification process, using, as a paradigm, the Dorsal/Acute 3 (DA3) muscle, which originates from a PC expressing Nau and Col (Crozatier and Vincent, 1999; Dubois et al., 2007; Keller et al., 1998). Nau is the single *Drosophila* ortholog of the mammalian family of bHLH myogenic regulatory factors (MRFs) that are at the core of the myogenic regulatory network (Michelson et al., 1990; Paterson et al., 1991; Sambasivan

* Corresponding authors. Fax: +33 5 61 55 65 07.
E-mail addresses: vincent@cict.fr (A. Vincent), laurence.dubois@univ-tlse3.fr (L. Dubois).

0012-1606/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ydbio.2011.12.018

and Tajbakhsh, 2007 and Weintraub et al., 1989 for review of MRFs). Nau function has been the subject of conflicting reports. On one side, it was proposed that Nau exerts general myogenic functions, similar to vertebrate MRFs (Misquitta and Paterson, 1999; Wei et al., 2007). On the other side, muscle-specific defects observed in *nau* mutant embryos suggested that Nau acts as an iTF (Balagopalan et al., 2001; Keller et al., 1998). Col is the single *Drosophila* member of the COE (Collier/Early B-Cell Factor) family of transcription factors (Dubois and Vincent, 2001; Daburon et al., 2008). *Ci-coe* has recently been shown to be a critical determinant of atrial siphon muscle fate in the ascidian *Ciona intestinalis* (Stolfi et al., 2010) but the functions of *ebf/coe* genes in vertebrate myogenesis remain little known. EBF(s) could contribute to the transcriptional regulation of *Xenopus* muscle development, in part via a positive feedback loop between EBF and MyoD (Green and Vetter, 2011). *Drosophila* Col is expressed in a large promuscular cluster at the origin of several PCs and this expression is maintained in a single muscle, the DA3 muscle, where it is required for normal development (Crozatier and Vincent, 1999). We have previously shown that promuscular *col* activation is controlled by an "early" cis-regulatory module (CRM) that integrates positional information and Notch-mediated lateral inhibition during the process of PC selection. A separate, "late" CRM then takes over and maintains robust *col* transcription in the DA3/DO5 PC and DA3 muscle lineages. This relay both depends upon Hox information (Enriquez et al., 2010) and direct binding of Col, revealing a handover mechanism at the PC stage (Dubois et al., 2007; Enriquez and Vincent, 2010). Finally, examination of *col* transcription in *nau* mutants showed that Nau is essential for robust *col* activation in the nuclei of FCMs that fuse with the DA3 FC (Dubois et al., 2007). We further explore here the respective roles of Col and Nau in conferring the DA3 muscle specific properties.

We first show that Col and Nau are expressed together in the three PCs at the origin of all the dorso-lateral (DL) muscles, the DA3, DO3, DO4, DO5, DT1 and LL1 muscles (Beckett and Baylies, 2007; Nose et al., 1998) and that these progenitors are born sequentially. Each expresses a specific combination of iTFs that results, in part from cross-regulation already occurring at this stage. In *col* mutant embryos, all the DL muscles show specific changes in their epidermal insertion sites, showing that, even when transient, Col expression is required for DL PC identity. Detailed analysis of the DA3>DA2 muscle transformation that is observed in absence of Col revealed that the final orientation of the DA3 muscle involves two steps: a transient attachment to several epidermal sites, followed by a resolution step selecting the definitive sites. The multiple other cases of triangular-shaped DL muscles observed in *col* mutant embryos indicate that transient binding of elongating muscle to exploratory sites could be a general feature of the developing musculature. In *nau* mutant embryos the DA3 shows transformation towards its sibling, DO5 muscle, with many cases of bifid DA3/DO5 fibres, a phenotype aggravated in *nau/col* hypomorphic conditions. In addition we find that all myofibres are thinner in *nau* than wt embryos, showing that *nau* activity controls both generic and muscle-specific differentiation programmes. The early Col function in specifying DL progenitor identity, general function of Nau in ensuring proper fibre size and combined activities of Nau and Col in the DA3 muscle differentiation process provide a clear example of combinatorial transcriptional coding of muscle-specific shapes.

Materials and methods

Drosophila genetics

The following Drosophila mutant alleles and transgenic constructs were used: col^{1} (Crozatier et al., 1999), *Pcol85-Gal4, UASmCD8GFP* (*col>GFP* (Krzemien et al., 2007)), *P9Gal4, UASmCD8GFP (P9cG>GFP* (Dubois et al., 2007)), CRM276-LacZ and 4_0.9-LacZ (Enriquez et al., 2010), nau^{CK188} (Balagopalan et al., 2001), *UAS-col* (Vervoort et al., 1999), *twist-Gal4* (Baylies and Bate, 1996), $Poxm^{R361}$ (Duan et al., 2007), $Kr^{CD+}Kr^{1}$ (Romani et al., 1996), $slou^{286}$ (Knirr et al., 1999). Mutant alleles and transgenic constructs were balanced over marked chromosomes: *Cyo twist-lacZ; TM3 twist-lacZ; Cyo dfd-EYFP; TM6b dfd EYFP*. All Gal4-UAS crosses were performed at 25 °C. The strain w^{118} was used as wt reference.

Plasmid constructions and transgenic lines

A HA tag was inserted in frame at the N-terminus of the Col open reading frame, before cloning the full length *col* cDNA into the PUAS vector ($PUASCol^{HA}$). The upstream 9_0.9 *col* genomic fragment (Dubois et al., 2007) was inserted into the *attB-inslacZ* vector (Enriquez et al., 2010). The Zh8 AttP platform at position 49D on the second chromosome was used for site-specific insertion (Bischof et al., 2007).

Immunohistochemical staining and imaging

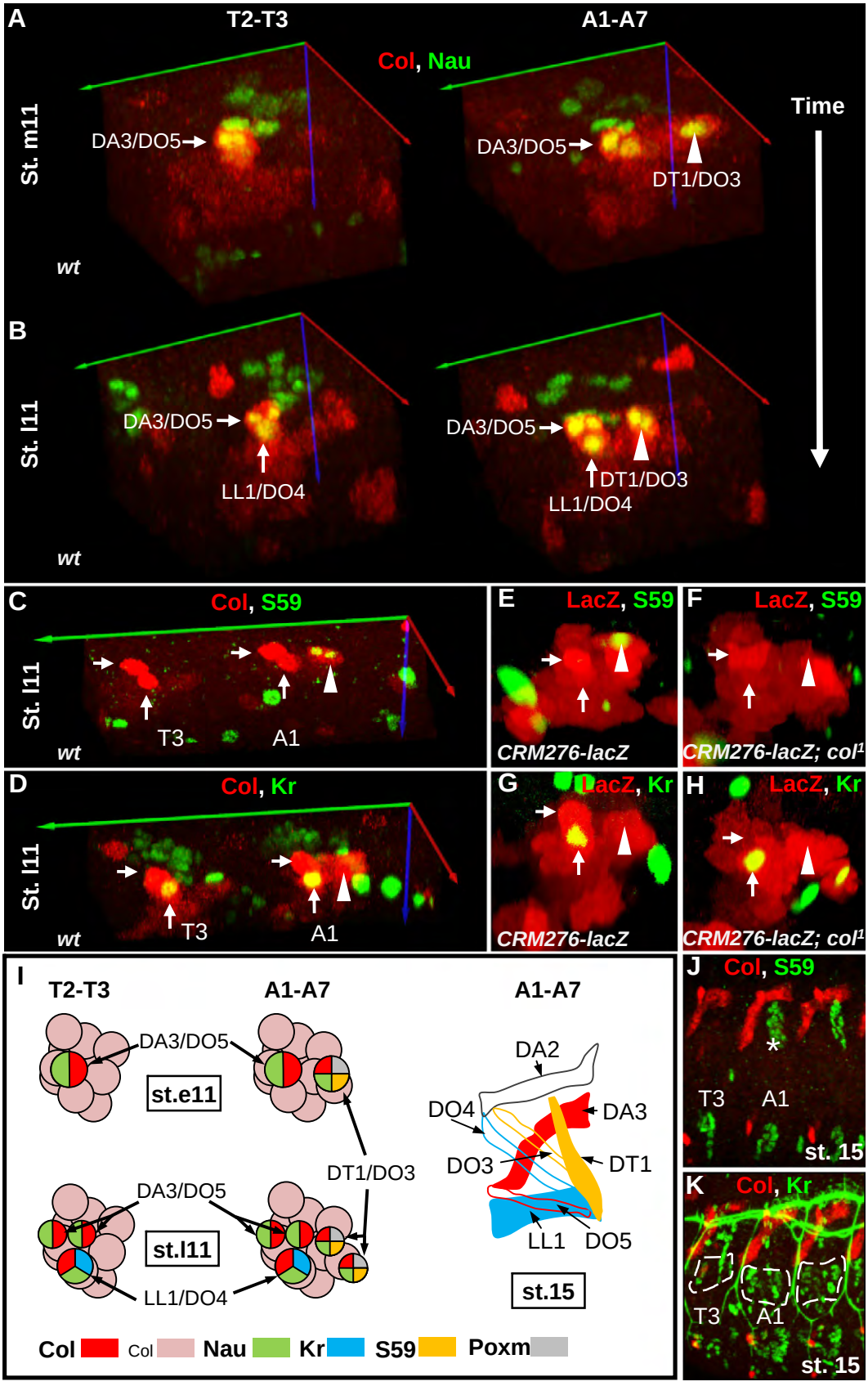
Embryos were fixed and processed for antibody staining as described (Crozatier et al., 1996). Primary antibodies were: mouse anti-Col (1/100) (Dubois et al., 2007), anti-β-galactosidase (Promega, 1/1000) and anti α-PS2 integrin, (Developmental studies Hybridoma Bank, 1/5); rabbit anti-Mef2 and anti-Nau (1/100) (provided by E. Furlong, Heidelberg, Germany and B. Paterson, Bethesda, MD, USA, respectively), anti-Kr (Gaul et al., 1987), anti S59 (Dohrmann et al., 1990); anti-Vg (provided by AJ. Simmonds, Edmonton, Canada); Secondary antibodies were: Alexa Fluor 488-conjugated goat anti-rabbit and goat anti-mouse; Alexa Fluor 555-conjugated goat anti-rabbit and goat anti-mouse; Alexa Fluor 647-conjugated goat anti-mouse (all Molecular Probes, 1/300); and biotinylated goat anti-mouse (Vector Laboratories, 1/1000). For staining with Phalloidin, embryos were manually devitellinized to avoid methanol treatment, which destabilises the cytoskeleton. 3-D reconstructions of the topology of DL progenitor and founder cells (see Fig. 1 A–H and S1,2) and muscles (see Fig. S3) were made from 200 to 500 nm thin sections acquired on a Leica SP5 confocal microscope at ×40 magnification, numerical zoom 4×, using Amira, (Visage Imaging GmbH) and Volocity (PerkinElmer) software.

SEM analysis of embryonic muscle pattern

For SEM analysis, *col* and *nau* mutant strains were balanced over the marked balancers Cyo Dfd-EYFP, and Tm6B Dfd-EYFP, respectively. Embryos were allowed to develop at 25 °C. After removal of the chorion by bleaching, live mutant embryos were selected under a Leica stereomicroscope equipped for epifluorescence. Stage-15 embryos were hand

Fig. 1. Expression of Col, Kr, Nau and S59 defines a dorso-lateral muscle progenitor code. (A,B) Col and Nau expression (yellow) between mid-stage 11 (A) and late stage 11 (B), in T (left) and A segments (right). (A) The newly born DA3/DO5 FCs and DT1/DO3 PC/FCs are indicated by a horizontal arrow and vertical arrowhead, respectively. (B) The DT1/DO3 PC has divided, and an additional progenitor (vertical white arrow) selected in both T and A segments. (C,D) Expression of Col (red) and either S59 (C) or Kr (D) (green) in late stage 11 embryos, T3 and A1 segments. The DT1 and DO3 PC/FCs express S59 while the ventral-most progenitor expresses Kr. (E,F) S59 and (G,H), Kr expression (green) in mid (E,F) and late (G,H) stage 11 wt (E,G) and col^{1} (F,H) embryos carrying the *CRM276-LacZ* transgene. LacZ staining visualises the Col expressing cluster and progenitors (I) Schematic representations. Left: temporal sequence of specification of the DA3/DO5, DT1/DO3 and LL1/DO4 PCs in T and A segments. Low and high levels of Col are indicated in pink and red, respectively, Nau in green, Kr in blue, S59 in yellow and Poxm in grey. Right: The corresponding muscle fibres at stage 15, indicating maintenance of Col, Kr and S59 expression in the DA3, LL1 and DT1 muscles, respectively (see Dohrmann et al., 1990; Ruiz-Gomez et al., 1997). (J, K) Immunostaining of stage 15 embryos for Col (red) and S59 or Kr (green). Positions of the T3 and A1 segments are indicated. (J) The DT1 muscle, indicated by a white asterisk, is specific to A segments. (K) The LL1 muscle is circled by a dotted line. In A–D, the green, red and blue arrows indicate the antero-posterior (A/P), dorso-ventral (D/V), and medio-lateral (ML) axes, respectively.

devitellinized, and fixed on their dorsal side on a coverslip coated with polylysine. They were then dissected and the fillets stretched to expose the somatic muscles. Fillets were fixed in a double aldehyde mixture (4% formaldehyde, 2.5% glutaraldehyde in 1× PBS), washed in water, and dehydrated through ethanol series. Following HMDS (hexamethyldisilazane) drying, fillets were sputtered with a gold-palladium coat



(JFC 1100 Jéol), and examined with either a Hitachi S450 microscope or TM-1000 tabletop microscope.

Results

Sequential specification of the three dorso-lateral muscle progenitors

Col is first expressed in the somatic mesoderm at embryonic stage 10, in a large cluster of myoblasts. We reported previously that this cluster gives rise to the DA3/DO5 and DT1/DO3 PCs in the T2-A7 segments and A1-A7 segments, respectively (Crozatier and Vincent, 1999; Enriquez et al., 2010). Based on reporter gene expression, we initially proposed that the second PC gave rise to the DT1 and DO4 muscles (Crozatier and Vincent, 1999), but we now favour the DT1/DO3 lineage, as proposed by (Carmena et al., 1995). The Col-expressing PCs also express Nau, a general marker of most if not all PCs and FCs (Keller et al., 1998; Michelson et al., 1990; Wei et al., 2007). The DA3/DO5 and DT1/DO3 PCs divide asynchronously, the DA3/DO5 PC dividing at mid-stage 11 (Fig. 1A) and the DT1/DO3 at late-stage 11 (Fig. 1B). When examining in detail this sequence of divisions, we noticed in late stage 11 embryos the appearance of another, large size cell, expressing both Col and Nau in the T2-A7 segments (Fig. 1B). The position of this cell, immediately ventral and internal to the DA3 and DO5 FCs, suggested that it could be an additional PC selected from the Col-expressing cluster, slightly ventral and later than the DA3/DO5 PC. Consistent with a PC identity, this cell divides into two smaller cells at the beginning of stage 12 (Fig. S1). To determine its molecular identity, we double-stained embryos for Col and either S59 or Krüppel (Kr), expressed in the DT1/DO3 and LL1/DO4 progenitors, respectively (Carmena et al., 1995; Dohrmann et al., 1990; Ruiz-Gomez et al., 1997). These double stainings confirmed the co-expression of S59 and Col in the DT1/DO3 PC (Fig. 1C) and showed that the PC that is selected late from the Col-expressing cluster expresses also Kr (Fig. 1D). It therefore corresponds to the LL1/DO4 muscle progenitor (Ruiz-Gomez et al., 1997).

Precisely timed immuno-staining experiments thus revealed that Col is expressed in all three PCs at the origin of the DL muscles (DA3, DO5, DT1, DO4, LL1, DO3, (Nose et al., 1998); Fig. 2A) and revealed that these PCs are specified in a fixed temporal sequence. At their birth time, each DL PC already expresses a specific code of iTFs in addition to Nau, either Col (DA3/DO5), Col/Kr (LL1/DO4) or Col/S59/Poxmeso (Poxm) (DT1/DO3), (Fig. 1I (Carmena et al., 1995; Cox and Baylies, 2005; Dohrmann et al., 1990; Duan et al., 2007)), suggesting that both positional and temporal clues could be involved in specification of their identity.

Epistatic relations between Col and other iTFs in dorso-lateral progenitors

The iTF code that is specific to each DL PC could result from epistatic or cross-repressive interactions between different iTFs, as observed in some other lineages (Jagla et al., 2002; Knirr et al., 1999; Lord et al., 1995). We therefore looked at S59 and Kr expression in *col*^{1} null mutant embryos. LacZ expression under the control of the early *col* CRM (CRM276) served to visualise the Col-expressing cluster and progenitors (Enriquez et al., 2010). S59 expression is lost from the DT1/DO3 PC, and therefore is downstream of Col in this lineage (Fig. 1E,F). S59 expression in the DT1 muscle precursor has recently been shown to be also dependent upon Poxm (Duan et al., 2007). We found that Poxm expression is specifically lost from the DT1/

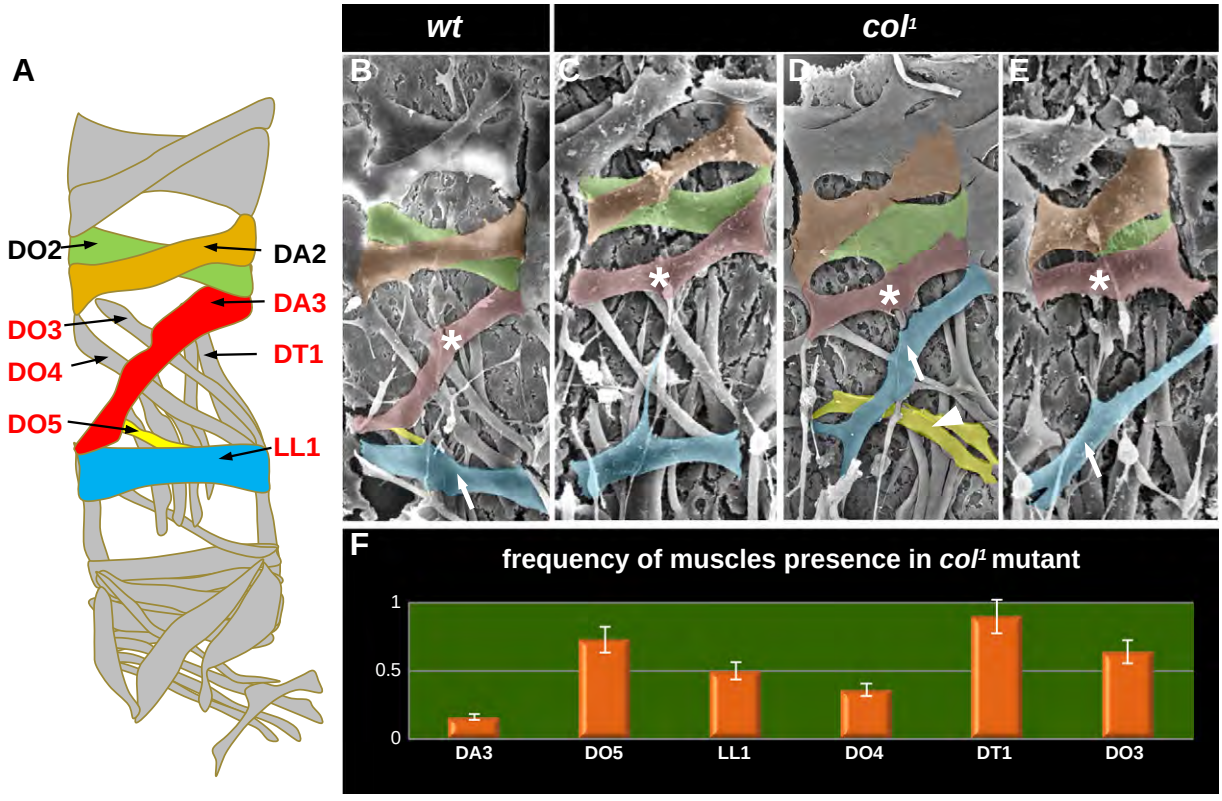


Fig. 2. Muscle phenotypes of *col* mutant embryos. (A) Schematic diagram of the muscle pattern in A2–A7 segments adapted from (Bate, 1993). (B–E) EM-scanning views of the abdominal musculature in stage 16 embryos. The internal face of one wt (B) and *col*^{1} (C–E) segment is shown. The D02, DA2, DA3, DO5 and LL1 muscles are colour-coded as in A. (C) The DA3 muscle (white asterisk) is transformed into a DA2-like muscle. (D) The LL1 muscle (white arrow) is transformed into a DA3-like muscle. In addition, the DO4 muscle (white arrowhead) adopts a DO5 morphology, while the DO3 muscle is also abnormal. Since SEM analysis cannot distinguish between the DO5 and transformed DO4 muscles, both are coloured in yellow. (E) The DA3 and LL1 muscle are transformed into DA2-like and DA3-like muscles, respectively, while the DO3, DO4 and DO5 muscles are absent. (F) Histogram representing the fraction of *col1* segments in which each indicated muscle forms normally.

DO3 progenitor in *col* mutant embryos and may therefore be an intermediate of S59 regulation by Col (Fig. S2). By contrast, Kr expression in the LL1/DO4 PC is unaffected in *col* mutant embryos and is therefore independent of Col activity (Fig. 1G,H). Inversely, Col expression is normal in either, *Kr, Pxm* or *S59* mutants (not shown). Thus Col acts upstream of Poxm and S59 in defining the DT1/DO3 progenitor identity. It regulates its own expression in the DA3/DO5 PC and, subsequently, the DA3 myofibre (Crozatier and Vincent, 1999; Enriquez et al., 2010). The restriction of Col auto-regulation to the DA3/DO5 PC suggests that this handover mechanism is dependent upon (an) other, still unknown, either activating or repressing TFs that distinguish the DA3/DO5 PC from the other DL PCs. Together, these expression data show that the molecular identity of DL muscles reflects sequences of regulations between different iTFs at the PC stage, which are specific to each PC.

Morphological transformations of the dorso-lateral muscles in col mutant embryos

While Kr and S59 remain expressed in the LL1 and DT1 muscles, respectively, Col expression is only maintained in the DA3 myofibre (Fig. 1J,K; (Dohrmann et al., 1990; Ruiz-Gomez et al., 1997)). Our analysis of the *col* mutant phenotype therefore initially focused on the DA3 muscle (Crozatier and Vincent, 1999). In order to determine which other muscles were affected, and in the absence of specific markers for the DO3, DO4 and DO5 muscles (Tixier et al., 2010), we performed morphological analyses of stage 15 embryos, using Scanning Electron Microscopy (SEM) (Fig. 2B–E) and phalloidin staining of the acto-myosin cytoskeleton of stage 17 embryos (Fig. S3). This first level of analysis revealed that all the DL muscles (Fig. 2A) were either missing or mis-specified in *col* mutant embryos at a statistically significant frequency (Fig. 2F), while the more dorsal or more ventral muscles were not affected.

The DA3 is the most often affected muscle. In 84% of *col* mutant segments (N = 48/57), it is replaced by a muscle at a more dorsal position and oriented parallel to the DA2 muscle, suggestive of a DA3 > DA2 transformation (Fig. 2C–E). The DA3 sibling, the DO5 muscle is present in 73% (N = 42/57) of segments. The LL1 and DO4 muscles are present in 51% (N = 29/57) and 36% (N = 21/57) of col^1 embryonic segments, respectively (Fig. 2F). We observed segments displaying both a DA3 > DA2 transformation and a DO5 muscles while a muscle at the position and with the same orientation than the DA3 muscle was present, suggesting a complex set of muscle transformations. Every time that both a DA3 > DA2 and a DA3-like muscles are present, the LL1 muscle is missing (16% of mutant embryos, N = 9/57; Fig. 2D, E). We conclude that the DA3-like muscle corresponds to a LL1 > DA3 transformation. In support of this interpretation, we observed muscles of intermediate morphology, i.e., displaying both the normal LL1 epidermal attachment sites and a supplementary projection towards the DA3 attachment site, a phenotype that we interpret as a partial LL1 > DA3 transformation (8%, N = 5/57; Fig. 2C and S3). Finally, two muscles oriented like the DO5 are present in some col^1 segments when the DO4 muscle is missing (Fig. 2D), indicating a re-orientation of the DO4 muscle parallel to DO5 (21% of mutant segments, N = 12/57). Similar to partial LL1 > DA3 transformations, we observed cases of DO4 muscles projecting additional, thin extensions towards the anterior attachment site of the DO5 muscle, suggestive of partial DO4 > D05 transformations (7%, N = 4/57; Fig. S3B,C). The DO3 and DT1 muscles originate from a PC that is only specified in abdominal segments (Crozatier and Vincent, 1999; Dohrmann et al., 1990; Enriquez et al., 2010). A DT1 muscle forms normally in about 90% of col^1 mutant segments (N = 46/51), while a DO3 muscle forms in 64% of segments (N = 33/51; Fig. 2F).

In summary, morphological observations show that the entire pattern of DL muscles is disrupted in *col* mutant embryos. At this level of analysis, the most obvious defect is mis-orientation of many myofibres, suggesting that in absence of Col activity, insertion site choice is often changed. While all DL muscles are affected to some extent, the most frequently observed defects are DA3 > DA2, LL1 > DA3, and DO4 > DO5 transformations, which indicate changes in progenitor identity.

Changes in muscle attachment sites upon loss or gain of Col activity

To further characterise the muscle transformations resulting from the loss of Col activity, we used specific muscle markers such as Vestigial (Vg) and Kr, which are expressed in the DA1, DA2, DA3, and LL1, and the DA1, LL1 and LT4 muscles, respectively (Fig. 3A,D; (Bate et al., 1993; Ruiz-Gomez et al., 1997)). The *col 4_0.9 lacZ* reporter gene served as a DA3 identity marker (Enriquez et al., 2010). Double Vg/LacZ staining of col^1 mutant embryos confirmed that the DA3 muscle adopts a DA2-like morphology (97% of segments at stage 15; N = 96/99), while the DA2 muscle is itself unaffected. A DA3 > DA2 transformation is already observed in hypomorphic *col* mutants ($Pcol > GFP/col^1$) where *GFP* recapitulates the *col* expression pattern (Fig. 3C, Krzemien et al., 2007). Vg staining of col^1 embryos also confirmed that the LL1 muscle adopts a DA3 morphology in at least 10% of segments (N = 8/77; Fig. 3B). The loss of Vg expression at the expected position for the LL1 or LL1 > DA3 in more than half of the segments (Fig. 3B; N = 43/77) could hide, however, other phenotypes. We therefore turned to Kr antibody staining, since Kr expression in the LL1 muscle does not depend upon Col activity. The Kr expression pattern confirmed that the LL1 muscle adopts a DA3-like morphology in a significant number of col^1 segments (Fig. 3D,E).

We then performed reciprocal experiments, i.e., examined whether expressing Col in the entire mesoderm could induce specific muscle transformations. We previously showed that Col ectopic expression in all FCs (rp298-UASCol; Dubois et al., 2007), although able to auto-activate a *col-lacZ* reporter gene, does not significantly alter the muscle pattern. To express Col earlier, including at the PC stage, we used the pan mesodermal Twist-Gal4 driver, (Greig and Akam, 1993). Vg staining of Twist > Col embryos showed that the LL1 muscle forms normally, indicating that changing the level of Col expression in the LL1 lineage does not redirect this muscle to another fate. On the contrary, the DA2 muscle adopts a DA3 or intermediate DA3/DA2 morphology in 28% (N = 35/133) and 26% (N = 37/133) of segments, respectively (Fig. 3F). The DA2 > DA3 transformation is the only specific muscle transformation that we could observe at high frequency, showing that Col ability to impose a new cell fate is strictly context-dependent, as previously observed in the central and peripheral nervous systems (Baumgardt et al., 2007; Crozatier and Vincent, 2008). The reciprocal transformations observed upon loss and gain of Col function, respectively, show that the DA2 PC is a target of reprogramming by Col and that Col activity distinguishes between the DA3 and DA2 identities (Fig. 3G).

Two steps in the selection of muscle insertion sites

The DA3 > DA2 muscle transformation in *col* mutant embryos indicated specific changes in the DA3 epidermal attachment sites. To better understand this phenotype, we first characterised the wt DA3 insertion sites, using double StripeB (SrB)/GFP staining of P9cG > GFP embryos. SrB is a specific marker for the tendon cells, which connect muscles to the epidermis (Volk and VijayRaghavan, 1994; Volohonsky et al., 2007); P9cG > GFP expression (Dubois et al., 2007) specifically labels the DA3 and, more stochastically, the DO5 muscle contours (Fig. 4A–C). Double stainings showed that, although at similar dorso/ventral (D/V) positions, the anterior insertion sites of the DA3 and DO5 muscles do not overlap. The DA3 muscle attaches to tendon cells along the segmental borders while the DO5 attaches to more internal cells. On the posterior side, the DA3 and DO5

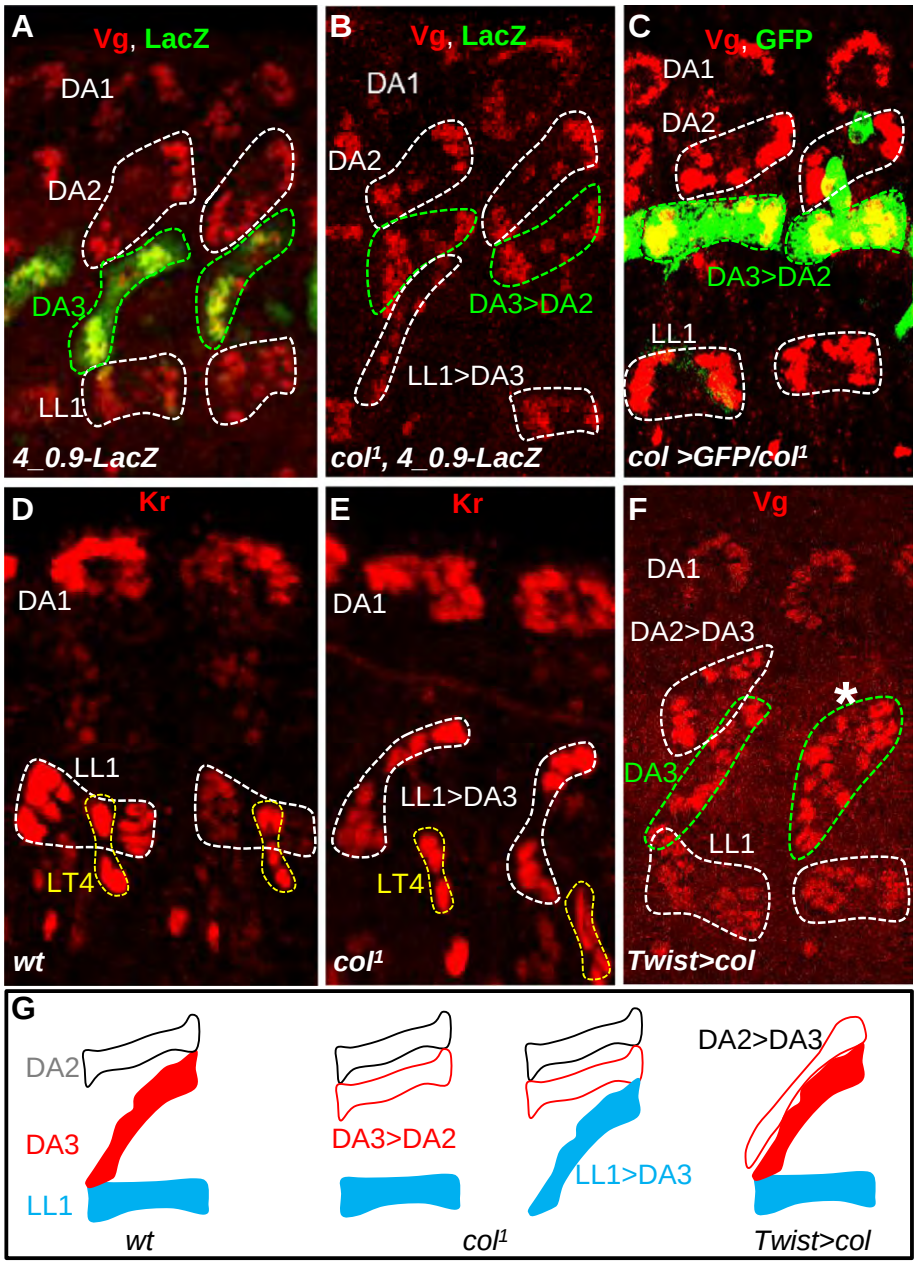


Fig. 3. Reciprocal DA3 > DA2 and DA2 > DA3 transformations in *col* loss-of-function and gain-of-function conditions. (A–F) Dorso-lateral views of stage 15 embryos stained for Vg (A–C, F) or Kr (D,E) (red), and LacZ (A–B) or GFP (C), green. Vg is expressed in the DA1, DA2 and LL1 muscles. The DA3 muscle is marked by LacZ and GFP expression, in 4_0.9LacZ and col > GFP embryos, respectively. The DA2 and LL1 muscles are circled in dotted white and the DA3 muscle in dotted green. (B,C) In *col* mutant embryos, the DA3 muscle adopts a DA2-like (DA3 > DA2) morphology. (B) One segment shows a (LL1 > DA3) transformation. (D, E) Immunostaining for Kr (red) confirms the LL1 > DA3 transformation in col^1 embryos while the LT4 muscle, circled in dotted yellow is unaffected. (F) Pan-mesodermal expression of Col (Twist > col) transforms the DA2 into a DA3-like muscle (circled green). The DA2 > DA3 transformation is either complete (asterisk, circled green) or only partial (circled white). (G) Schematic diagram of the DA3 > DA2 and LL1 > DA3 transformations in *col* loss-of-function (col^1) and reciprocal DA2 > DA3 transformation in *col* gain-of-function conditions.

attach to different tendon cells along the segmental border (Fig. 4A). Along this characterisation, we discovered that the final, acute orientation of the DA3 muscle is reached in two steps. In a first step, the DA3 muscle is attached to two distinct groups of tendon cells along the anterior segmental border and a third group along the posterior border. This three-attachment configuration gives the DA3 muscle its characteristic angled shape at stage 14 (Fig. 4B–B″). In a second step, the middle attachment site is lost, leading to a final diagonal orientation. This second step that we call the resolution step, takes place between stages 14 and 15 (Fig. 4J). In order to verify that the growing DA3 muscle forms an integrin-mediated junction with tendon cells at its intermediate attachment site, we double stained *col > GFP* embryos for GFP and the muscle specific αPS2-integrin (Bokel and Brown, 2002). This experiment indicated that αPS2-integrin accumulates at this intermediate site (stage 14; Fig. S4a), showing that a transient myotendinous junction forms. αPS2-integrin accumulation is then restricted to the DA3 final attachment sites at stage 16 (Fig. S4b), confirming that attachment site selection is a highly regulated, two-step process.

We then characterised the epidermal attachment sites of the DA3 > DA2 transformed muscle, in various combinations of *col* hypomorphic mutations, including the hypomorphic *col > GFP* allele. In heterozygous *col > GFP* embryos, a few "angled" muscles maintaining both dorsal and ventral anterior attachment sites are observed, indicating a partial DA3 > DA2 transformation (Fig. 4 C,D,G). The penetrance of this phenotype increases in homozygous *col > GFP*

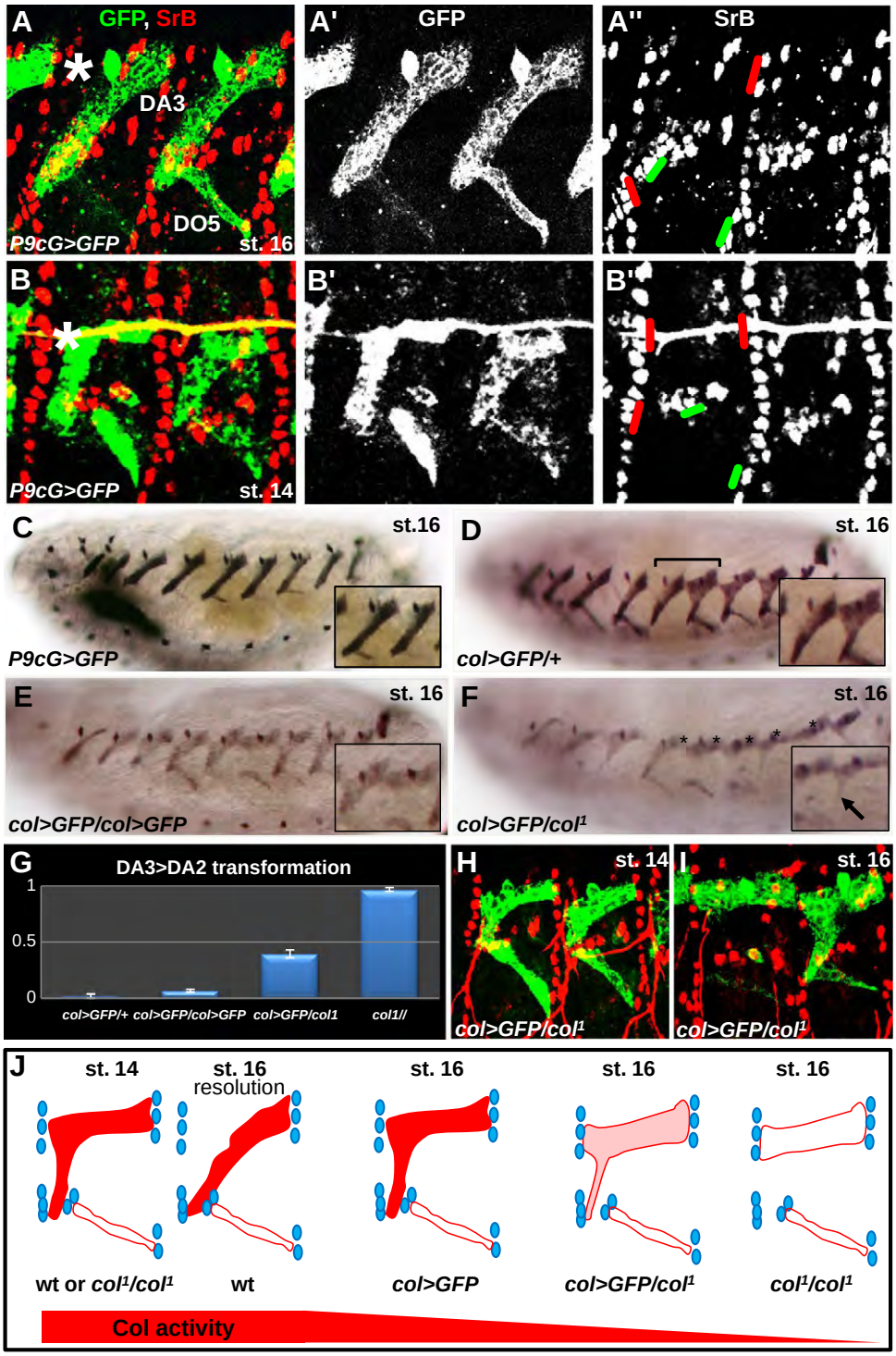
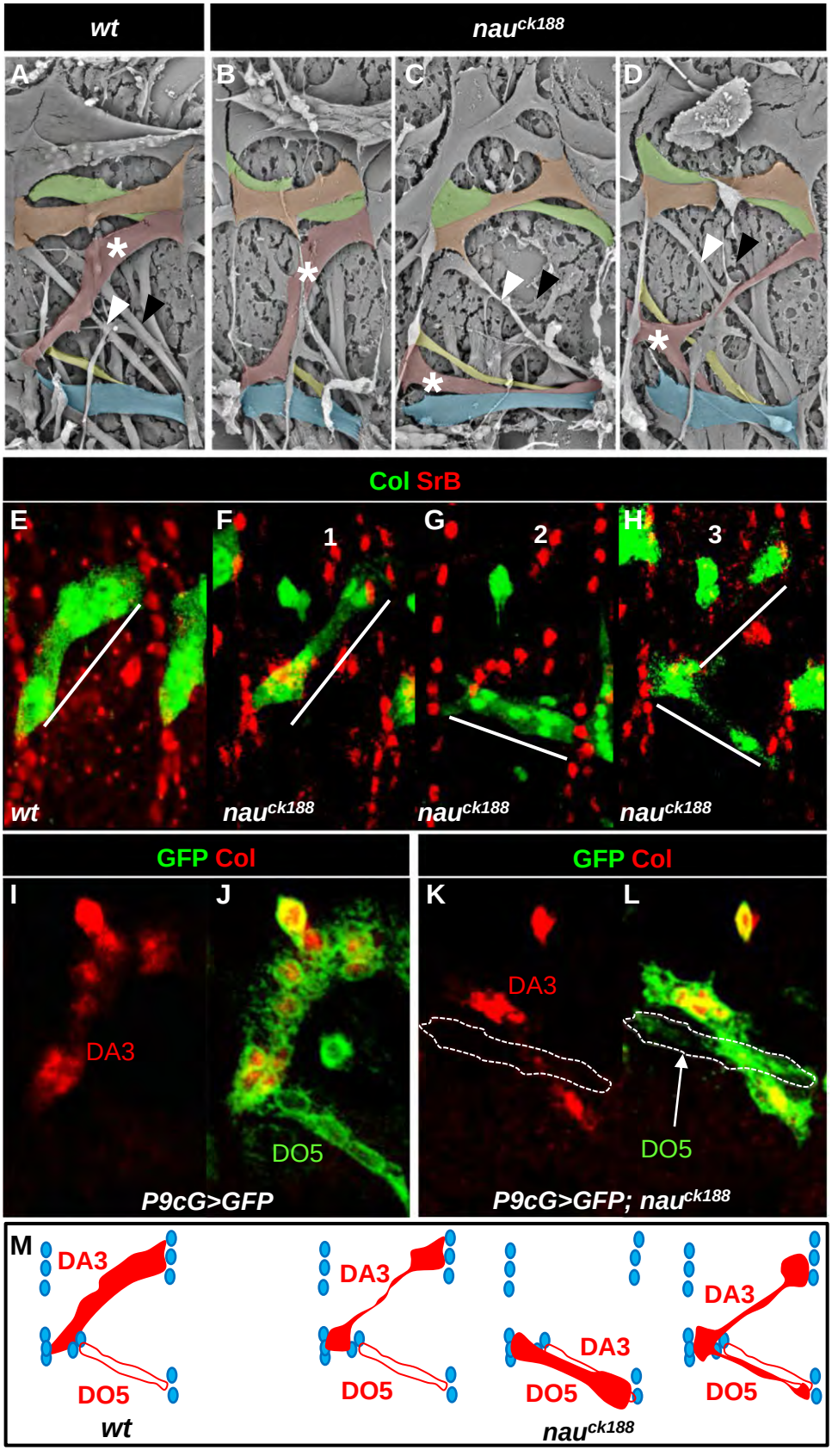


Fig. 4. Transient and final epidermal attachment sites in wt and *col* mutant embryos. (A) Stage 16, and (B), stage 14 P9cG > GFP embryos stained for GFP (DA3 and DO5 muscles; A,A′, B,B′) and SrB (A,A″, B,B″). Green and red bars in A″,B″ indicate the wt DO5 and DA3 attachment sites, respectively. The DA3 muscle shows an additional attachment site (asterisk) at stage 14. (C-F) GFP staining of the DA3 (and DO5) muscles in stage 16 P9cG > GFP embryos (C) and (D–F) various *col* > *GFP*/*col*^{1} mutant combinations, shows a progressively increasing number of DA3 > DA2 transformations; for *col*^{1} mutants, statistics are based on Vg staining, (G). Incomplete transformations maintaining three epidermal attachment sites are shown in insets in D–F. (H, I) *col* > *GFP*/*col*^{1} mutant embryos stained for SrB (red) and GFP (green). The DA3 transient attachment site observed at stage 14 (H) is maintained, leading to DA3 > DA2 re-orientation (I, left) or triangular shape muscles (I, right). (J) Diagrammatic representation of the two-steps selection of the DA3 epidermal attachment sites in wt embryos and mis-attachment phenotypes observed in different *col* mutants. The dose of *col* activity decreases from left to right (red arrow).

embryos. In addition, 7% of segments (N = 30/441) now show a complete DA3 > DA2 morphological transformation and this number increases to 40% of *col* > *GFP*/*col*^{1} segments (N = 75/189; Fig. 4E–G). In most other segments, the DA3 > DA2 muscle maintains thin projections towards the wt DA3 attachment site, suggesting again an incomplete transformation. Comparison of *col* > *GFP*/*col*^{1} embryos at stages 14 and 16 shows that the final attachment site of the DA3 > DA2 muscle corresponds to the transient DA3 attachment site that is seen in wt embryos (4B,H,I).

In summary, detailed examination of the DA3 > DA2 muscle phenotype showed that Col activity is required for attachment of the DA3 muscle to its ventral insertion site and that robustness of this

site selection requires physiological levels of Col activity. It also revealed that the final diagonal orientation of the DA3 muscle is established in two successive steps, a second, resolution step being critical (summarised in Fig. 4J). Of note, *col* mutant embryos show specific mis-attachments or triangular shapes for several muscles other than DA3 (Fig. 2 and S3), suggesting the involvement of a



resolution step for these muscles as well. Transient binding to exploratory sites during the fibre elongation process could therefore be a general feature of the developing musculature.

Nau is both required for proper muscle size and orientation of specific muscle fibres

Independent reports have described a major disorganisation of the embryonic musculature of *nau* mutant embryos (Misquitta and Paterson, 1999; Wei et al., 2007) and higher sensitivity of a subset of muscles to *nau* mutations, respectively (Balagopalan et al., 2001; Dubois et al., 2007; Keller et al., 1998). We herein re-explored the morphology of muscles in *nau* mutant embryos, using SEM analysis. It revealed a previously unnoticed phenotype which is that most fibres are much thinner in *nau* than in wt embryos (Fig. 5 and S5), confirming that, although not an essential gene, Nau plays a general myogenic regulatory function in ensuring proper fibre size. In addition, we observed many cases of either loss or mis-formation of the DA3, DO4 and DO3 muscles, while other muscles were hardly affected (Fig. 5A–D) confirming that *nau* also acts as a muscle-specific iTF (Keller et al., 1998). Puzzlingly, the severity of both phenotypes varies from embryo to embryo (Balagopalan et al., 2001; Dubois et al., 2007; Keller et al., 1998) (compare Fig. 5B–D and S5). Focusing our analysis on the DA3 muscle, we defined 4 classes of phenotypes. The DA3 muscle is either: 1) unaffected (Fig. 5A, B); 2) oriented like the DO5 muscle, (Fig. 5C); 3) attached at both the DA3 and DO5 posterior attachment sites, forming a kind of "bifid" fibre (Fig. 5D); 4) severely affected, with multiple short extensions, a phenotype that we describe as clumsy phenotype (see Fig. 6). We observed a small number of DA3>DA2 transformations (not shown), suggesting that Nau could contribute to the robustness of the handover of Col activity that is specific to the DA3/DO5 PC, consistent with Nau positively regulating *col* transcription in this lineage. To further characterise the different classes of *nau* phenotypes and calculate statistics, we used double stainings of wt and *nau* embryos for Col and Sr (Fig. 5E–H). We also counted the nuclei by staining for Mef-2 (Lilly et al., 1994; Nguyen et al., 1994) (Fig. S6). In *nau* null embryos, the DA3 muscle shows normal orientation and nuclei number (9 nuclei per DA3 muscle at stage 16; (Enriquez et al., 2010), in 51% of segments (N=115/225)). Despite having the same number of nuclei, the majority of these fibres appear much thinner than wt (Fig. 5E,F). The DA3 muscle attaches to intra-segmental tendon cells typical of a DO5 attachment and adopts the oblique orientation of the DO5 muscle in 32% (N=72/225) of segments (Fig. 5G and S6). The DA3>DO5 fibres contain only 6 nuclei on average, which is less than a wt DA3 and more than a wt DO5 (3 nuclei on average, at stage 16; Fig. S6B,E) suggesting a partial transformation. In 6% (N=14/222) of segments, the DA3 muscle displays a "bifid" DA3+DO5 morphology (Fig. 5H and S6D). In this case, the total number of nuclei is similar to wt DA3 (Fig. S6E). Finally, the DA3 muscle is absent or very poorly developed (clumsy phenotype) in 8% of segments (N=18/225), while it adopts a DA2-like orientation in 4% of segments analysed (N=9/225) (data not shown; see also Keller et al., 1998). Examination of P9cG>GFP;*nau* embryos confirmed that the DO5 muscle properly forms in the absence of Nau activity (Fig. 5I–L and S6).

In summary, our data show that, in absence of Nau activity, there is randomisation of the DA3 attachment sites between the wt DA3 and DO5 positions (Fig. 5M) as well as the formation of unstructured fibres not resembling any particular muscle, a previously observed phenotype (Keller et al., 1998). Even when correctly oriented, the *nau* mutant DA3 fibres are generally thinner than wt, despite a normal or slightly reduced number of nuclei per fibre (Fig. S6). Nau thus controls aspects of muscle fibre growth in the embryo, independently of the number of FC/FCM fusion events. *nau* phenotypes, either thinning or improper fibre orientation, vary from embryo to embryo (compare Fig. 5A–D and Fig. S5) and between segments in the same embryo (Fig. 5D), suggesting that *nau* activity may not be decisive but rather confers robustness to both generic and muscle-specific differentiation programmes.

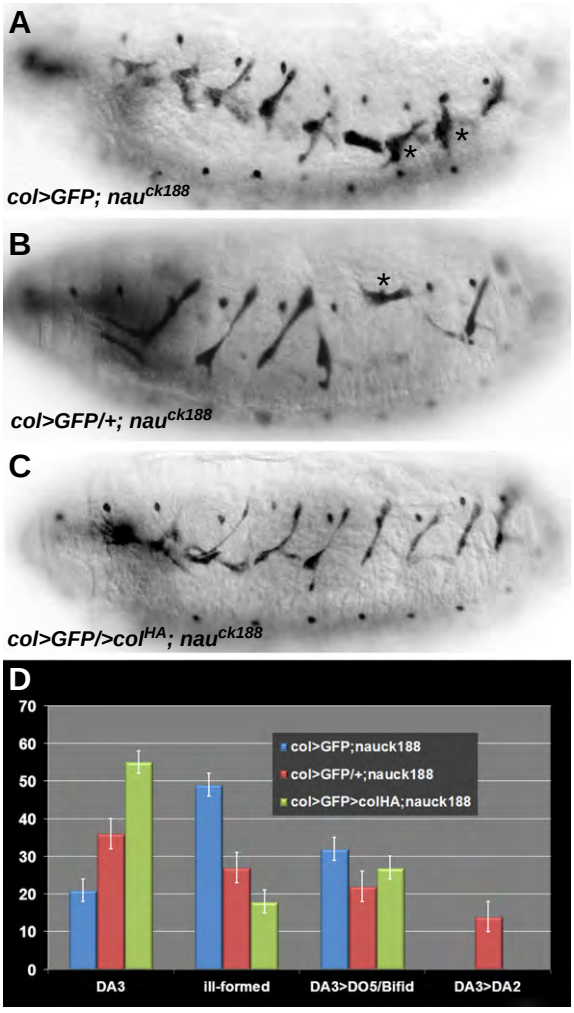


Fig. 6. Cumulative effects of *nau* and *col* mutations. (A–C) GFP staining of, (A) *col>GFP*; *nau*^{ck188}, (B) *col>GFP/+;nau*^{ck188} and (C), *col>GFP/>col*^{HA}*;nau*^{ck188} stage 16 embryos. GFP expression marks the DA3 muscle. The DA3 phenotype is aggravated in double *col>GFP;nau*^{ck188} embryos, with an increasing number of ill-formed muscles (A, black asterisks). DA3>DA2 muscles are observed in *col>GFP/+;nau*^{ck188} embryos (B, black asterisk). Both phenotypes, but not the DA3>DO5 and bifid fibres phenotypes are rescued by increased levels of Col activity (C). However rescued fibres remain thinner than in col>GFP embryos (see Fig. 4 D,E). Of note, rescue is restricted to abdominal segments. (D) Histogram indicating the number of DA3, DA3>DA2, DA3>DO5+bifid, and ill-formed muscles in *nau*^{ck188} mutant embryos expressing increasing levels of Col. Number of segments counted: N=192 *col>GFP*; *nau*^{ck188}, N=129 *col>GFP/+;nau*^{ck188}; N=123 *col>GFP/>col*^{HA}*;nau*^{ck188}.

Fig. 5. DA3 muscle phenotypes of *nau* mutant embryos. (A–D) EM-scanning views of the abdominal musculature of stage 16 wt (A) and *nau*^{ck188} (B–D) embryos. The internal face of one A segment is shown. Colour coding is as in Fig. 2. The white asterisk underlines the DA3 muscle. White and black arrowheads point to the DO4 and DO3 muscles, respectively, showing the frequent loss of these muscles in *nau*^{ck188} mutant embryos. (E–H) wt (E), and *nau*^{ck188} (F–H) embryos double-stained for Col and SrB. The orientation of wt and abnormal DA3 muscles is indicated by white lines. The DA3>DO5 muscle (G) contacts the same posterior tendon cells than wt DO5 (see Fig. 4). (I–L) Staining of *P9cG>GFP* (wt) and *P9cG>GFP;nau*^{ck188} embryos for Col (red) (IL) and GFP (green, J,L) showing a DA3>D05 transformation. The DO5 muscle is indicated by a white arrow in (L). (M) Schematic representation of the DA3 (red) and DO5 (circle red) muscle orientations and attachment sites in wt and *nau*^{ck188} mutant embryos, illustrating the reduced fibre size and stochastic orientation of the DA3 muscle in mutant embryos.

Col and Nau combinatorial control of the DA3 epidermal attachment sites

We next investigated the cumulative effect of *nau* and *col* mutations, using the hypomorphic *col>GFP* allele in order to be able to follow the morphology of the DA3 muscle. We examined trans-allelic *col>GFP/+; nau^{ck188}* and *col>GFP; nau^{ck188}* mutant combinations. Reducing the level of *col* activity in *nau* mutant embryos resulted in a strong aggravation of the DA3 muscle defects (Fig. 6A,B), with only 21% of DA3 muscles correctly oriented in *col>GFP; nau^{ck188}* compared to 51% in *nau^{ck188}* (see above) and 40% in *col>GFP/+; nauck^{188}* embryos (Fig. 6D). We also observed an increased number of clumsy muscles and DA3 muscles only attached to anterior tendon cells, phenotypes that are collectively referred to as ill-formed muscles, at the expense of DA3>DA2 muscle transformations (Fig. 6A,B,D). However, the fraction of DA3>DO5 transformations and DA3+DO5 bifid fibres, which are typical *nau* mutant phenotypes, was not significantly increased by reducing Col levels. These observations indicate a cumulative effect of *nau* and *col* mutations. Interestingly, a fraction (12%) of *col>GFP/+; nau^{ck188}* DA3 muscles adopted a DA2 morphology, the same fraction than observed in homozygous col>GFP mutant embryos (Fig. 4G), consistent with the decreased levels of *col* transcription in *nau* mutant, compared to wt embryos (Dubois et al., 2007). Nau up-regulation of *col* transcription could thus contribute to the robustness of the DA3 identity already at the progenitor stage. Restoring high levels of Col expression in *col>GFP/+; nauck^{188}* embryos, by introducing a UAS-col^{HA} transgene (*col>GFP/>col^{HA}; nauck^{188}* embryos), resulted in both full rescue of the DA3>DA2 phenotype and decreased number of ill-formed muscles, without significantly reducing the fraction of DA3>DO5 and bifid fibres (Fig. 6B–D). Together the double mutant analysis and rescue results show that combinatorial activity of *nau* and *col* is required for the correct orientation and differentiation of the growing DA3 muscle fibre (Fig. 7). Of note, even when correctly oriented, the majority of *col>GFP/>col^{HA}; nau^{ck188}* DA3 muscles remained thin (Fig. 6C), confirming that the fibre size control exerted by Nau is independent of *col* regulation.

Discussion

The larval *Drosophila* somatic musculature is made of a stereotyped set of about 30 uniquely identifiable muscles per hemisegment. Here, we show that the combinatorial activities of Col and Nau are required to establish the pattern of DL muscles and confer upon these muscles their distinctive shapes and epidermal attachment sites.

Sequential specification of the dorso-lateral muscle progenitors

Col is expressed in a promuscular cluster and the three derived PCs at the origin of DL muscles. Each of these PCs is specified at a stereotypic position and according to a precise temporal sequence, with the dorsal DA3/DO5 PC being born first and the ventral LL1/DO4 PC being born last. In addition to Col and Nau, each expresses a combination of specific iTFs, including Kr, Poxm and S59, (Fig. 1l) (Crozatier and Vincent, 1999; Dohrmann et al., 1990; Duan et al., 2007; Ruiz Gomez and Bate, 1997). Expression of a specific set of iTFs in each DL PC could thus integrate both positional and temporal cues. The textbook view is that, similar to neuroblast selection in the neuroectoderm, each muscle PC is selected via the process of lateral inhibition from an equivalence group of mesodermal cells. The parallel between neuroblast and PC selection is supported by the co-expression of the proneural gene *l(1)sc* and iTFs such as Eve or S59 in specific promuscular clusters (Carmena et al., 1995, 1998 and Baylies and Michelson,

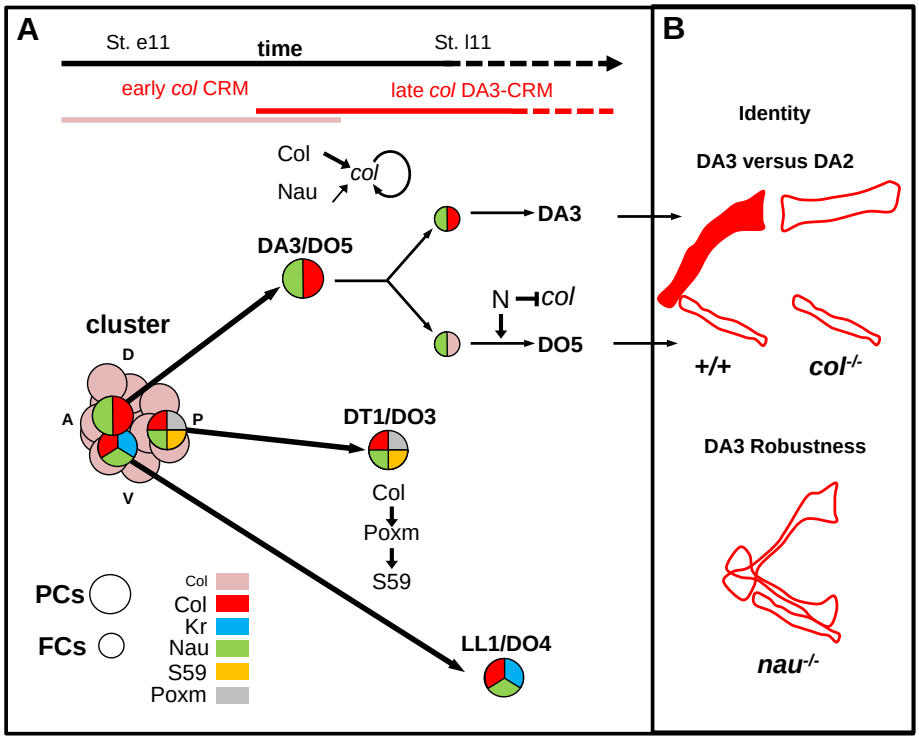


Fig. 7. The sequential and combinatorial coding of a muscle shape. (A) Schematic diagram of the DL muscle lineages. Col is expressed at low level in a promuscular cluster (pale pink) and high level in the three derived DL PCs (large colour-coded circles), which are born at fixed positions along the A–P and D–V axes. In addition to their stereotyped positioning, there exists a temporal sequence of specification of these PCs; the embryonic stages and periods of activity of the early and late *col* DA3 CRMs are indicated on a time axis (top). Each DL PC expresses different iTFs in addition to Col and Nau, such as Kr (blue), S59 (yellow) and Poxm (grey). Col activity contributes to the identity of all DL PCs, including by regulating Poxm and S59 expression in the DT1/DO3 PC and its own expression in the DA3/DO5 PC. Nau also contributes to maintain *col* expression from the PC stage. The DA3 and DO5 FCs (smaller circles) inherit both Col and Nau. B) Absence of Col activity leads to a DA3>DA2 transformation. Nau activity ensures robustness to the DA3 identity programme. This includes maintaining high level *col* transcription in this lineage (Dubois et al., 2007) and distinguishing between the DA3 and DO5 attachment sites. Thus, Col and Nau act both sequentially and combinatorially in the DA3 lineage.

2001; Frasch, 1999, for reviews). However, a deficiency of *l(1)sc* results in only minor defects of somatic muscle development (Carmena et al., 1995; Duan et al., 2007) and does not prevent the selection of the DA3/DO5 and DT1/DO3 progenitors (Crozatier and Vincent, 1999). A possibility evoked by (Duan et al., 2007), is that several PCs could be selected from large competence domains defined by expression of specific iTFs and *l(1)sc* clusters play only a limited or redundant role. Selection of the three DL PCs from a cluster of Col-expressing cells supports this view. The fact that the DA3/DO5 and LL1/DO4 PCs are born sequentially and positioned adjacent to one another (Fig. 1 and S1), suggests a reiterative selection process.

The col mutant phenotype exhibits changes in progenitor identity

The most obvious muscle pattern defects that are observed in *col* mutant embryos, are DA3>DA2 and DA3>LL1 transformations. Since the DA2, DA3 and LL1 muscles are derived from different PCs, these defects indicate changes in progenitor identity. On one side, Poxm and S59 expression in the DT1/DO3 progenitor requires Col activity. On the other side, Kr expression in the LL1/DO4 PC is independent of Col. Interestingly, Kr and S59 are expressed together in the ventral VA1/VA2 PC but, in this case, Kr regulates S59 expression (Ruiz-Gomez et al., 1997). Together, these expression data strengthen the concept of combinatorial coding of muscle identity (Bourgouin et al., 1992) at the PC stage and show that hierarchies of interactions between different iTFs are progenitor-specific.

In *Poxm* mutants, the DO3 muscle is often duplicated, likely at the expense of a DT1 muscle that is often missing (Duan et al., 2007). In *S59* mutant embryos, the DO3 and DT1 sibling muscles share ventral attachment sites and form a single syncitium in a fraction of segments (Knirr et al., 1999). Since Col acts upstream of Poxm and S59 in the DT1/DO3 progenitor, we expected the *col* mutant phenotype to overlap with the *poxm* and *S59* phenotypes. It may not be so simple, however, since the DT1 muscle is absent only in a small fraction of segments in *col* mutant embryos. Interestingly, while the DA3 muscle is transformed into a DA2-like muscle in absence of Col activity, the LL1 muscle can adopt a DA3 morphology The LL1 muscle is misoriented in *col* as well as in *Kr* mutant embryos (Dohrmann et al., 1990; Ruiz-Gomez et al., 1997). Together, these muscle reorientation phenotypes suggest that there is a range of possible attachment sites for each elongating DL muscle and that the final pattern results from a global combinatorial control. The propensity of elongating muscles to explore several attachment sites (see below), could explain why a coordinate, global regulation by combinations of iTFs is essential. The term regulatory state has been used to describe the total set of active transcription factors in a given cell at a given time (Peter and Davidson, 2011). In essence, each PC iTF code is an example of a regulatory state. The loss of one iTF reveals an alternative regulatory state and PC identity, suggesting that a given iTF is able to exert its activity only in the presence of other specific iTFs. A global analysis of this mutual dependency now requires the identification of all DL iTFs, including those expressed in the DO3, DO4 or DO5 muscles.

The myogenic functions of Nau, the Drosophila MyoD ortholog

Nau differs from other well characterised iTFs, in that it is expressed in most, if not all FCs (Wei et al., 2007), before being restricted to specific muscle precursors (Dubois et al., 2007; Keller et al., 1998; Michelson et al., 1990). SEM analysis shows that most muscles are much thinner in *nau* mutant than wt embryos. Detailed examination of the mutant DA3 muscle showed that, despite being thinner, it contained a number of nuclei close to normal. *nau* activity is thus required for embryonic muscle fibre size, but not the muscle fusion programme, per se. Whether Nau directly or indirectly regulates the synthesis and/or assembly of myofibril proteins remains to be determined. As first noted by Keller et al. (1998), DL muscles, including the DA3 muscle, are more severely affected. Taken together, these data lead us to conclude that Nau performs both general myogenic functions and specific functions in selected muscle lineages. A different threshold level of MRF activity might be needed to initiate myogenesis in different trunk and craniofacial muscles (reviewed by (Sambasivan et al., 2011)). The different Nau functions in establishing the *Drosophila* muscle pattern suggest that Nau activity is, in part conditioned by interactions with other iTFs such as Col.

Temporal, combinatorial coding and robustness of muscle shape

Co-expression of Nau and Col in the DA3/DO5 progenitor provides a good model to challenge the concept of combinatorial control of muscle identity (Dubois et al., 2007). While transformed towards a DA2 muscle in absence of Col activity, the DA3 muscle adopts the morphology of its sibling, DO5 muscle in absence of Nau. Thus, while co-expressed in the DA3/DO5 PC, Col and Nau act at different steps in the DA3 lineage. We propose the following regulatory cascade (Fig. 7): Col expression in a large cluster of myoblasts and the three derived PCs, under control of an early CRM and Hox activity (Enriquez et al., 2010), defines a domain of competence for DL muscle development. Col activity, either upstream, and/or in parallel to other iTFs, contributes to confer each DL progenitor its particular identity. The restricted ability of Col in maintaining its own expression in the DA3 FC, by direct binding to a late, DA3-specific CRM (Dubois et al., 2007), reveals a context-dependence provided by the iTF combination specific to the DA3/DO5 PC. This PC-specific handover process may explain why the DA3 muscle is the most frequently affected in *col* mutant embryos. Asymmetric division of each DL PC generates two FCs with different regulatory states (Carmena et al., 1998; Ruiz Gomez and Bate, 1997). Whereas two DA3 and two DO5 muscles form in Notch (N) loss- and gain of function conditions, respectively (Crozatier et al., 1996), Nau confers robustness to the DA3 versus DO5 differentiation programme. This Nau function involves positive regulation of *col* transcription in the DA3 syncytium nuclei (Dubois et al., 2007) and is independent of Nau function in ensuring normal fibre size.

In conclusion, our data show that the sequence of expression and combinatorial activities of Col and Nau are required to establish the pattern of DL muscles and confer upon the DA3 muscle its distinctive size and epidermal attachment sites. Identification of the gene targets of this combination is now essential to link a sequence of regulatory states to the architecture of a specific *Drosophila* muscle. Interestingly, a recent report suggests that EBF cooperates with MyoD in driving aspects of differentiation in *Xenopus* muscle cells, suggesting that there may be an ancient, evolutionarily conserved, transcriptional relationship between the COE/EBF and MyoD gene families (Green and Vetter, 2011).

Muscle targeting of specific tendon cells

Embryonic muscles connect to the chitinous exoskeleton of the developing embryo via tendon cells, which are specialised epidermal cells (Becker et al., 1997; Schnorrer and Dickson, 2004). Proper attachment of muscles requires the specific targeting of tendon cells at segmental or intra-segmental, stereotypic positions. The general view is that growing myotubes extend filopodia at their two ends, in search of attachment sites, and that muscle extension ceases when muscles have reached their targeted tendon cells. Some muscle guidance components have been described, such as the Derailed receptor tyrosine kinase for the lateral transverse muscles and the Robo and Robo2 receptors, the transmembrane protein Kon-Tiki and its associated intracellular signalling protein dGrip for ventral-longitudinal muscles (Callahan et al., 1996; Kidd et al., 1999; Schnorrer et al., 2007; Swan et al., 2004). How the precise matching

of specific muscles to specific tendon cells is achieved, however, is far from being understood (Schnorrer and Dickson, 2004; Schnorrer et al., 2007; Schweitzer et al., 2010). SEM analysis and phalloidin staining of *col* mutant embryos showed many mis-oriented muscles, suggesting targeting defects. Many fibres showed more than two attachment sites to the epidermis, however, a phenotype difficult to reconcile with a bipolar extension of muscle precursors until they connect to the epidermis. Rather, the observation that the wt DA3 muscle is transiently attached to three sites, before acquiring its fully extended bipolar morphology, indicates the existence of an exploratory step, followed by a resolution step that selects the final attachments sites. The allelic series of *col* phenotypes, which revealed many triangular shape fibres, indicates a defect in the resolution process, without ruling out that ventral elongation of the DA3 myofibre is also defective. Terminal differentiation of tendon cells is dependent upon their interaction with muscles (Schweitzer et al., 2010; Yarnitzky et al., 1997) and tendon cells could play a role in the resolution step. Triangular shape LO1 muscles were previously observed in mutants for *dgit*, which encodes a GTPase activator protein that is involved in myotube guidance (Bahri et al., 2009). Based on the *dgit* phenotype, and our own observations, we propose that the migratory path of muscles towards their targeted tendon cells can involve exploratory attachment to tendon cells along this path. Deciphering how the final, stereotyped, pattern is controlled now requires the identification of how various iTF combinations differentially regulate guidance cues.

Acknowledgements

We thank R. Cripps, E. Furlong, M. Frasch, H. Jäckle, M. Knoll, B. Patterson and T. Volk and the Bloomington Stock Center for antibodies and flies, and L. Bataillé, C. Danesin, M. Roussigné, and members of our laboratory for constructive criticisms on the manuscript. We acknowledge the help of B. Ronsin and A. Leru, TRIO Imaging platform, and J. Favier and F. Luce for fly culture.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.ydbio.2011.12.018.

References

Bahri, S.M., Choy, J.M., Manser, E., Lim, L., Yang, X., 2009. The *Drosophila* homologue of Arf-GAP GIT1, dGIT, is required for proper muscle morphogenesis and guidance during embryogenesis. Dev. Biol. 325, 15–23.
Balagopalan, L., Keller, C.A., Abmayr, S.M., 2001. Loss-of-function mutations reveal that the *Drosophila* nautilus gene is not essential for embryonic myogenesis or viability. Dev. Biol. 231, 374–382.
Bate, M., 1993. The mesoderm and its derivatives. In: Bate, M., Martinez Arias, A. (Eds.), The Development of *Drosophila melanogaster*, Vol. 2. Cold Spring Harbor Laboratory press, p. 1013.
Bate, M., Rushton, E., 1993. Myogenesis and muscle patterning in *Drosophila*. C. R. Acad. Sci. Ser. III 316, 1047–1061.
Bate, M., Rushton, E., Frasch, M., 1993. A dual requirement for neurogenic genes in *Drosophila* myogenesis. Dev. Suppl. 149–161.
Baumgardt, M., Miguel-Aliaga, I., Karlsson, D., Ekman, H., Thor, S., 2007. Specification of neuronal identities by feedforward combinatorial coding. PLoS Biol. 5, e37.
Baylies, M.K., Bate, M., 1996. Twist: a myogenic switch in *Drosophila*. Science 272, 1481–1484.
Baylies, M.K., Michelson, A.M., 2001. Invertebrate myogenesis: looking back to the future of muscle development. Curr. Opin. Genet. Dev. 11, 431–439.
Baylies, M.K., Bate, M., Ruiz Gomez, M., 1998. Myogenesis: a view from *Drosophila*. Cell 93, 921–927.
Becker, S., Pasca, G., Strumpf, D., Min, L., Volk, T., 1997. Reciprocal signaling between *Drosophila* epidermal muscle attachment cells and their corresponding muscles. Development 124, 2615–2622.
Beckett, K., Baylies, M.K., 2007. 3D analysis of founder cell and fusion competent myoblast arrangements outlines a new model of myoblast fusion. Dev. Biol. 309, 113–125.
Bischof, J., Maeda, R.K., Hediger, M., Karch, F., Basler, K., 2007. An optimized transgenesis system for *Drosophila* using germ-line-specific phiC31 integrases. Proc. Natl. Acad. Sci. U. S. A. 104, 3312–3317.
Bokel, C., Brown, N.H., 2002. Integrins in development: moving on, responding to, and sticking to the extracellular matrix. Dev. Cell 3, 311–321.
Bourgouin, C., Lundgren, S.E., Thomas, J.B., 1992. Apterous is a *Drosophila* LIM domain gene required for the development of a subset of embryonic muscles. Neuron 9, 549–561.
Buff, E., Carmena, A., Gisselbrecht, S., Jimenez, F., Michelson, A.M., 1998. Signalling by the *Drosophila* epidermal growth factor receptor is required for the specification and diversification of embryonic muscle progenitors. Development 125, 2075–2086.
Callahan, C.A., Bonkovsky, J.L., Scully, A.L., Thomas, J.B., 1996. *derailed* is required for muscle attachment site selection in *Drosophila*. Development 122, 2761–2767.
Carmena, A., Bate, M., Jimenez, F., 1995. Lethal of scute, a proneural gene, participates in the specification of muscle progenitors during *Drosophila* embryogenesis. Genes Dev. 9, 2373–2383.
Carmena, A., Gisselbrecht, S., Harrison, J., Jimenez, F., Michelson, A.M., 1998. Combinatorial signaling codes for the progressive determination of cell fates in the *Drosophila* embryonic mesoderm. Genes Dev. 12, 3910–3922.
Cox, V.T., Baylies, M.K., 2005. Specification of individual Slouch muscle progenitors in *Drosophila* requires sequential Wingless signaling. Development 132, 713–724.
Crozatier, M., Vincent, A., 1999. Requirement for the *Drosophila* COE transcription factor Collier in formation of an embryonic muscle: transcriptional response to notch signalling. Development 126, 1495–1504.
Crozatier, M., Vincent, A., 2008. Control of multidendritic neuron differentiation in *Drosophila*: the role of Collier. Dev. Biol. 315, 232–242.
Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S., Vincent, A., 1996. Collier, a novel regulator of *Drosophila* head development, is expressed in a single mitotic domain. Curr. Biol. 6, 707–718.
Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S., Vincent, A., 1999. Head versus trunk patterning in the *Drosophila* embryo; Collier requirement for formation of the intercalary segment. Development 126, 4385–4394.
Daburon, V., Mella, S., Plouhinec, J.L., Mazan, S., Crozatier, M., Vincent, A., 2008. The metazoan history of the COE transcription factors. Selection of variant HLH motif by mandatory inclusion of a duplicated exon in vertebrates. BMC Evol. Biol. 8, 131.
Dohrmann, C., Azpiazu, N., Frasch, M., 1990. A new *Drosophila* homeo box gene is expressed in mesodermal precursor cells of distinct muscles during embryogenesis. Genes Dev. 4, 2098–2111.
Duan, H., Zhang, C., Chen, J., Sink, H., Frei, E., Noll, M., 2007. A key role of Pox meso in somatic myogenesis of *Drosophila*. Development 134, 3985–3997.
Dubois, L., Vincent, A., 2001. The COE-Collier/Olf1/EBF-transcription factors: structural conservation and diversity of developmental functions. Mech. Dev. 108, 3–12.
Dubois, L., Enriquez, J., Daburon, V., Crozet, F., Lebreton, G., Crozatier, M., Vincent, A., 2007. Collier transcription in a single *Drosophila* muscle lineage: the combinatorial control of muscle identity. Development 134, 4347–4355.
Enriquez, J., Vincent, A., 2010. Segmental variations in the patterns of somatic muscles: what roles for Hox? Fly (Austin) 4.
Enriquez, J., Boukhatmi, H., Dubois, L., Philippakis, A.A., Bulyk, M.L., Michelson, A.M., Crozatier, M., Vincent, A., 2010. Multi-step control of muscle diversity by Hox proteins in the *Drosophila* embryo. Development 137, 457–466.
Frasch, M., 1999. Controls in patterning and diversification of somatic muscles during *Drosophila* embryogenesis. Curr. Opin. Genet. Dev. 9, 522–529.
Gaul, U., Seifert, E., Schuh, R., Jackle, H., 1987. Analysis of Kruppel protein distribution during early *Drosophila* development reveals posttranscriptional regulation. Cell 50, 639–647.
Green, Y.S., Vetter, M.L., 2011. EBF proteins participate in transcriptional regulation of *Xenopus* muscle development. Dev. Biol. 358, 240–250.
Greig, S., Akam, M., 1993. Homeotic genes autonomously specify one aspect of pattern in the *Drosophila* mesoderm. Nature 362, 630–632.
Jagla, T., Bidet, Y., Da Ponte, J.P., Dastugue, B., Jagla, K., 2002. Cross-repressive interactions of identity genes are essential for proper specification of cardiac and muscular fates in *Drosophila*. Development 129, 1037–1047.
Keller, C.A., Grill, M.A., Abmayr, S.M., 1998. A role for nautilus in the differentiation of muscle precursors. Dev. Biol. 202, 157–171.
Kidd, T., Bland, K.S., Goodman, C.S., 1999. Slit is the midline repellent for the robo receptor in *Drosophila*. Cell 96, 785–794.
Knirr, S., Azpiazu, N., Frasch, M., 1999. The role of the NK-homeobox gene slouch (S59) in somatic muscle patterning. Development 126, 4525–4535.
Krzemien, J., Dubois, L., Makki, R., Meister, M., Vincent, A., Crozatier, M., 2007. Control of blood cell homeostasis in *Drosophila* larvae by the posterior signalling centre. Nature 446, 325–328.
Lilly, B., Galewsky, S., Firulli, A.B., Schulz, R.A., Olson, E.N., 1994. D-MEF2: a MADS box transcription factor expressed in differentiating mesoderm and muscle cell lineages during *Drosophila* embryogenesis. Proc. Natl. Acad. Sci. U. S. A. 91, 5662–5666.
Lord, P.C., Lin, M.H., Hales, K.H., Storti, R.V., 1995. Normal expression and the effects of ectopic expression of the *Drosophila* muscle segment homeobox (msh) gene suggest a role in differentiation and patterning of embryonic muscles. Dev. Biol. 171, 627–640.
Michelson, A.M., Abmayr, S.M., Bate, M., Arias, A.M., Maniatis, T., 1990. Expression of a MyoD family member prefigures muscle pattern in *Drosophila* embryos. Genes Dev. 4, 2086–2097.
Misquitta, L., Paterson, B.M., 1999. Targeted disruption of gene function in *Drosophila* by RNA interference (RNA-i): a role for nautilus in embryonic somatic muscle formation. Proc. Natl. Acad. Sci. U. S. A. 96, 1451–1456.
Nguyen, H.T., Bodmer, R., Abmayr, S.M., McDermott, J.C., Spoerel, N.A., 1994. D-mef2: a *Drosophila* mesoderm-specific MADS box-containing gene with a biphasic expression profile during embryogenesis. Proc. Natl. Acad. Sci. U. S. A. 91, 7520–7524.

Nose, A., Isshiki, T., Takeichi, M., 1998. Regional specification of muscle progenitors in *Drosophila*: the role of the msh homeobox gene. Development 125, 215–223.
Paterson, B.M., Walldorf, U., Eldridge, J., Dubendorfer, A., Frasch, M., Gehring, W.J., 1991. The *Drosophila* homologue of vertebrate myogenic-determination genes encodes a transiently expressed nuclear protein marking primary myogenic cells. Proc. Natl. Acad. Sci. U. S. A. 88, 3782–3786.
Peter, I.S., Davidson, E.H., 2011. Evolution of gene regulatory networks controlling body plan development. Cell 144, 970–985.
Romani, S., Jimenez, F., Hoch, M., Patel, N.H., Taubert, H., Jackle, H., 1996. Kruppel, a *Drosophila* segmentation gene, participates in the specification of neurons and glial cells. Mech. Dev. 60, 95–107.
Ruiz Gomez, M., Bate, M., 1997. Segregation of myogenic lineages in *Drosophila* requires numb. Development 124, 4857–4866.
Ruiz-Gomez, M., Romani, S., Hartmann, C., Jackle, H., Bate, M., 1997. Specific muscle identities are regulated by Kruppel during *Drosophila* embryogenesis. Development 124, 3407–3414.
Sambasivan, R., Tajbakhsh, S., 2007. Skeletal muscle stem cell birth and properties. Semin. Cell Dev. Biol. 18, 870–882.
Sambasivan, R., Kuratani, S., Tajbakhsh, S., 2011. An eye on the head: the development and evolution of craniofacial muscles. Development 138, 2401–2415.
Schnorrer, F., Dickson, B.J., 2004. Muscle building; mechanisms of myotube guidance and attachment site selection. Dev. Cell 7, 9–20.
Schnorrer, F., Kalchhauser, I., Dickson, B.J., 2007. The transmembrane protein Kon-tiki couples to Dgrip to mediate myotube targeting in *Drosophila*. Dev. Cell 12, 751–766.
Schweitzer, R., Zelzer, E., Volk, T., 2010. Connecting muscles to tendons: tendons and musculoskeletal development in flies and vertebrates. Development 137, 2807–2817.
Stolfi, A., Gainous, T.B., Young, J.J., Mori, A., Levine, M., Christiaen, L., 2010. Early chordate origins of the vertebrate second heart field. Science 329, 565–568.
Swan, L.E., Wichmann, C., Prange, U., Schmid, A., Schmidt, M., Schwarz, T., Ponimaskin, E., Madeo, F., Vorbruggen, G., Sigrist, S.J., 2004. A glutamate receptor-interacting protein homolog organizes muscle guidance in *Drosophila*. Genes Dev. 18, 223–237.
Tixier, V., Bataille, L., Jagla, K., 2010. Diversification of muscle types: recent insights from *Drosophila*. Exp. Cell Res. 316, 3019–3027.
Vervoort, M., Crozatier, M., Valle, D., Vincent, A., 1999. The COE transcription factor Collier is a mediator of short-range Hedgehog-induced patterning of the *Drosophila* wing. Curr. Biol. 9, 632–639.
Volk, T., VijayRaghavan, K., 1994. A central role for epidermal segment border cells in the induction of muscle patterning in the *Drosophila* embryo. Development 120, 59–70.
Volohonsky, G., Edenfeld, G., Klambt, C., Volk, T., 2007. Muscle-dependent maturation of tendon cells is induced by post-transcriptional regulation of stripeA. Development 134, 347–356.
Wei, Q., Rong, Y., Paterson, B.M., 2007. Stereotypic founder cell patterning and embryonic muscle formation in *Drosophila* require nautilus (MyoD) gene function. Proc. Natl. Acad. Sci. U. S. A. 104, 5461–5466.
Weintraub, H., Tapscott, S.J., Davis, R.L., Thayer, M.J., Adam, M.A., Lassar, A.B., Miller, A.D., 1989. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. Proc. Natl. Acad. Sci. U. S. A. 86, 5434–5438.
Yarnitzky, T., Min, L., Volk, T., 1997. The *Drosophila* neuregulin homolog Vein mediates inductive interactions between myotubes and their epidermal attachment cells. Genes Dev. 11, 2691–2700.

III – Discussion

III.1 – Analyse des CRM contrôlant la transcription de *col* au cours de la myogenèse.

Pendant les stades embryonnaires 11 à 14, Col est exprimé dans différents tissus : le segment intercalaire, le muscle DA3 et des neurones du SNC et du SNP (Crozatier et al., 1999; Crozatier and Vincent, 1999, 2008; Demilly et al., 2011). Il régule des gènes spécifiques dans chacun de ces tissus. Le contrôle précis de l'expression de Col dans ces différents tissus est donc un élément clé de sa fonction. Le patron musculaire larvaire de la drosophile est constitué d'un ensemble de 30 muscles, chaque muscle à une position donnée le long des axes du corps présentant une morphologie/identité propre. La mise en place de ce patron stéréotypé reflète l'activation d'une combinatoire spécifique de FTSS, dans chaque progéniteur musculaire en fonction de sa position au sein du mésoderme. La première étape de spécification de l'identité musculaire est l'activation de FTSS spécifiques dans des groupes de myoblastes (groupes promusculaires) donnant naissance aux progéniteurs des muscles, en fonction de la position de ces groupes au sein du mésoderme. L'étude pionnière de l'expression du FTSS Even-skipped (Eve) dans les groupes promusculaires les plus dorsaux a permis d'identifier un CRM intégrant à la fois l'information de position en provenance de l'ectoderme (signalisations Wnt et Dpp) et une information tissulaire mésodermique (Halfon et al., 2000). Ce CRM, appelé MHE (**M**uscle and **H**eart **E**nhancer), possède des sites de fixation pour les facteurs Mad, dTCF et Pnt, les effecteurs nucléaires des voies de signalisation Dpp, Wg et Ras/MAPK, et pour les facteurs mésodermiques Twi et Tin. Un algorithme de recherche (CodeFinder) a été développé par l'équipe de Martha Bulyk sur la base de cette caractérisation afin d'identifier *in silico* d'autres fragments génomiques regroupant des sites de fixation pour 4 au moins de ces 5 facteurs comme autant de CRM mésodermiques dorsaux potentiels. C'est à partir de cette recherche que le ^{E}CRM de *col*, actif dans un groupe promusculaire dorso-latéral, donc ventral au groupe promusculaire Eve, a été identifié (Boukhatmi, 2013; Enriquez et al., 2010). L'interprétation de ces résultats est que la même information de position est donc lue différemment par différents CRM « promusculaires ». L'identification récente d'un CRM contrôlant l'activation du gène *tup* à une position intermédiaire entre Eve et Col (Boukhatmi et al., 2012) renforce cette hypothèse mais les mécanismes moléculaires sous-jacents restent cependant à définir expérimentalement. Des études comparées des 3 CRM promusculaires de *eve*, *col* et *tup* devraient permettre de progresser sur cette question.

En raison de la construction de l'algorithme, les prédictions CodeFinder ont tendance à sélectionner des fragments génomiques plus larges que la taille moyenne des CRMs. Une identification plus précise du ^{E}CRM de *col* restait une condition préliminaire nécessaire à l'étude plus approfondie de ce CRM dans le groupe promusculaire, en particulier par modifications dans un contexte génomique « normal » (voir ci-dessous). La construction de nouveaux transgènes m'a permis d'identifier un fragment de 1,4kb pour le ^{E}CRM qui est uniquement actif dans le groupe promusculaire Col (le CRM initial était aussi actif dans la glande lymphatique). Il contient 1 site de liaison conservé pour les facteurs Pnt et Mad, 2 sites pour le facteur Tin et 3 sites pour le facteur Twi. Une analyse par mutation dirigée des CRM promusculaires de *eve*, *col* et *tup* devrait permettre de comprendre le mécanisme de positionnement des groupes promusculaires. L'hypothèse de travail est que le niveau de signalisation Dpp et Wg (et Ras/MAPK ?) conditionne l'activité de chaque CRM. Comment ? Plusieurs scénarios sont possibles : lecture directe, différentielle de ce niveau en fonction du nombre, de la répartition et des affinités relatives des sites de fixation pour Mad et dTCF dans chaque CRM ; intégration de la durée de l'interaction entre le CRM et ces FT ; compétition avec des régulateurs négatifs eux-mêmes en aval de Mad et Dpp. En réalisant des mutations uniques ou groupées des sites de fixation prédits pour les facteurs de transcription Mad, dTCF (et Pnt) sur chacun de ces CRM, notamment pour les sites conservés au cours de l'évolution, on pourra mesurer l'impact de chacune des voies dans la spécification identitaire de chaque groupe promusculaire et distinguer entre ces scénarios. L'utilisation de pFlyFos pour l'étude du ^{E}CRM de *col* permettra ensuite de confirmer l'hypothèse retenue en contexte génomique reconstitué.

L'existence de sites de fixation de Twi et Tin sur le ^{E}CRM et sur le ^{L}CRM prédits *in silico*, a été confortée *in vivo* par des expériences de ChIP-chip (Sandmann et al., 2007). Concernant Tin, Hadi Boukhatmi a cependant montré qu'il n'était pas nécessaire à l'activation promusculaire de *col* , le nombre de myoblastes exprimant *col* étant même augmenté dans un mutant *tin* (Boukhatmi et al., 2012) ; Il pourrait donc être impliqué dans une régulation négative, probablement indirecte. La liaison de Twi sur le ^{E}CRM est en accord avec le rôle précoce de Twi dans la spécification du mésoderme somatique à l'origine des muscles larvaires. La fonction de la liaison de Twi au ^{L}CRM restait par contre à déterminer. La mutation du site de fixation de Twi dans un gène rapporteur comportant uniquement le ^{L}CRM n'a pas révélé de rôle majeur de ce site dans la régulation tardive de Col. Quel pourrait être alors son rôle ? Il est intéressant de noter que la fixation de Twi sur le ^{L}CRM précède la fenêtre temporelle d'activité de ce CRM mais corrèle avec la fenêtre temporelle d'activité du ^{E}CRM. Une hypothèse envisagée est que la fixation concomitante de Twi sur les $^{E-}$ et ^{L}CRM favoriserait l'ouverture et la communication entre ces deux CRM distants de

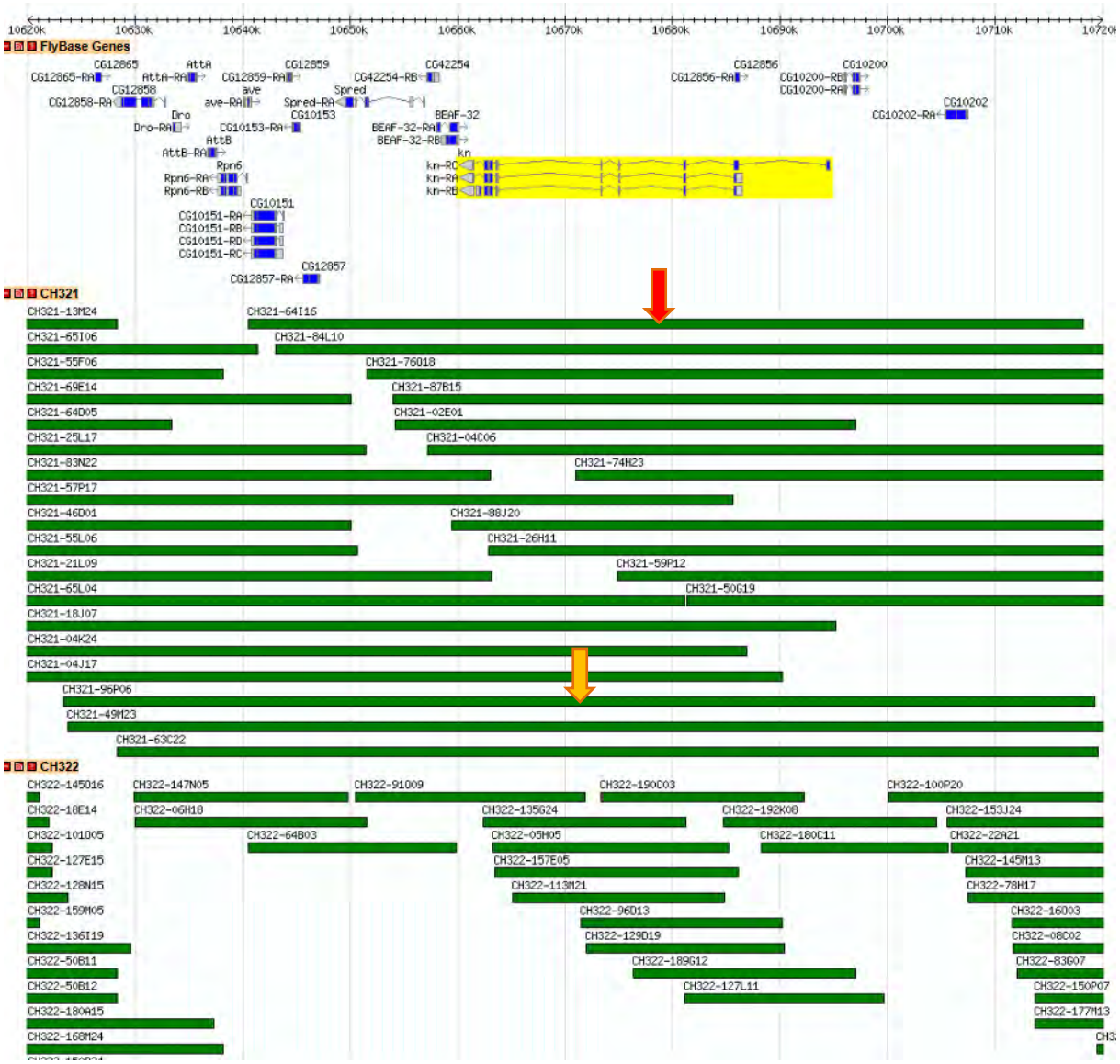


Fig. D1 – Les différents BAC p[acman] recouvrant le gène *collier*

La région génomique de *col* (ou *knot* – *kn*) est insérée dans plusieurs BAC de type p[acman] (Venken et al., 2009). Les p[acman] de type CH321 (série supérieure) contiennent environ 80 kb de séquences génomiques, ceux de type CH322 (série inférieure) environ 20 kb. La stratégie de ChAPseq a été initiée avec le p[acman] CH312-64I16 (flèche rouge) et l'équipe de K. White a étiqueté le p[acman] CH321-96P06 (flèche jaune). Tous deux contiennent l'ensemble du gène *col* ainsi que plusieurs kilobases de régions génomiques en amont et en aval.

plusieurs kilobases au stade cellule progéniteur, lors du processus de relais entre ces 2 CRM (Boukhatmi et al., 2012; Enriquez et al., 2010). Cette hypothèse devra être testée si une interaction en cis des $^{E-}$ et ^{L}CRM est établie (cf. II.1.3). Une autre hypothèse est celle d'un recrutement collectif des FTSS mésodermiques tels que Tin et Twi au sein d'un complexe de pré-transcription comportant les ^{E}CRM et ^{L}CRM (Junion et al., 2012).

La recherche systématique d'éléments régulateurs dans la région génomique de *col* entreprise par J.L. Frendo à l'aide des lignées GMR et Vienna tiles (cf. II.2.2.f) a permis d'exclure l'existence de CRM musculaires embryonnaires autres que le ^{E}CRM et ^{L}CRM déjà caractérisés. En parallèle, l'analyse des éléments cis-régulateurs du gène identitaire de *tup* a montré l'existence d'un CRM précoce et d'un CRM tardif soumis à autorégulation, une situation très semblable à la régulation de *col* (Boukhatmi et al., 2012). Le mécanisme de relais observé entre $^{E-}$ et ^{L}CRM pourrait donc être un mécanisme général de mémoire de l'identité de la cellule progéniteur lors de la différenciation musculaire. Il est maintenant impératif de tester cette hypothèse dans un contexte génomique et de dépasser la simple utilisation de gènes rapporteurs. J'ai donc amorcé une stratégie d'utilisation de pFlyFos, même si cette stratégie peut aussi révéler des surprises (voir ci-dessous).

III.2 – « ChAPseq » : une stratégie qui reste à privilégier

Mon objectif initial était d'identifier des cibles directes de Col spécifiquement impliquées dans la réalisation de l'identité du muscle DA3. D'où le choix d'une stratégie de ChAPseq basée sur l'expression ciblée de la protéine Col^{BLRP}. Un avantage de cette stratégie était la possibilité d'étudier en parallèle le rôle des CRM de *col* « en contexte ». Les résultats obtenus avec une approche conventionnelle de ChIPseq confirment que la mise au point de la stratégie de ChAPseq doit être poursuivie (cf. II.2.1 et III.3). En première instance, ma stratégie de ChAPseq était basée sur l'utilisation d'un BAC de type P[acman] (Venken et al., 2009), contenant des fragments d'ADN génomique plus longs que les pFlyFos (80 kb en moyenne contre 30 kb pour les pFlyFos), et incluant ainsi toutes les régions régulatrices de *col*. Le P[acman] choisi, CH321-64I16 (cf. Fig. D1), s'est malheureusement avéré non fonctionnel en tests de sauvetage phénotypique. J'ai donc recommencé avec un pFlyFos, le pFlyFos #022589 que j'ai étiqueté. Le pFlyFosCol* permet un sauvetage complet de la létalité associée à la mutation *col^{1}*. Malheureusement, en dépit du sauvetage phénotypique, les premiers immuno-marquages que j'ai réalisés avec le pFlyFos Col*-HA.BLRP ne m'ont pas permis de détecter le peptide HA inséré en

partie carboxy-terminale de la protéine Col. Plusieurs hypothèses peuvent être avancées : soit l'étiquette HA n'est pas présente, soit elle n'est pas accessible à l'anticorps et ne peut donc pas être détectée. Un séquençage de l'ADN du pFlyFos *col*-HA.BLRP* intégré dans le génome doit me permettre d'exclure qu'une mutation soit intervenue. Si la position de l'étiquette en C-terminal de la protéine la rend inaccessible à l'anticorps il faudra positionner cette étiquette en N-terminal, avec le risque de recombinaisons Flp/FRT aléatoires que cela comporte (cf. Matériel et méthodes). Une autre hypothèse est l'absence d'expression de l'isoforme Col étiquetée. Il existe en effet deux isoformes majoritaires du transcrit *col* qui sont générées par épissage alternatif (Crozatier et al., 1996), et qui codent pour deux isoformes protéiques différentes, répertoriées B et D dans Flybase (Marygold et al., 2013). Le codon-stop de l'isoforme B de 575 acides aminés est situé dans un intron épissé dans les transcrits de l'isoforme D de 557 acides aminés. Les deux isoformes B et D sont fonctionnelles *in vitro* (Ntini and Wimmer, 2011). J'ai choisi d'étiqueter la forme B puisque l'ADNc correspondant est celui inclus dans le transgène P5colcDNA utilisé depuis 1999 pour le sauvetage de la létalité embryonnaire. La non-détection de l'étiquette HA pourrait cependant refléter l'expression exclusive de l'isoforme D qui est codée par les transcrits *col* majoritaires dans l'embryon (http://flybase.org/cgi-bin/gbrowse/dmel/?Search=1;name=FBgn0001319). Dans la continuité des études à grande échelle menées par le consortium ModEncode, une nouvelle collection de lignées de drosophiles contenant des P[acman] étiquetés a été générée par R. Spokony (laboratoire de K. White). Une lignée porte le P[acman] CH321-96P06 (cf. Fig.D1) dans lequel le gène *col* a été étiqueté en C-terminal avec une étiquette GFP-BLRP sur l'isoforme D. Je suis actuellement en train de tester si une copie de P[acman] CH321-96P06 permet de sauver la létalité embryonnaire du mutant *col1* et de vérifier le patron d'expression de la GFP dans les embryons de cette lignée. Celle-ci pourrait permettre de contourner les premières difficultés rencontrées.

De manière intéressante, le pFlyFos *col** permet aussi un sauvetage du phénotype de perte de région centrale de l'aile alors que le CRM précédemment identifié comme actif dans le disque d'aile (Hersh and Carroll, 2005) n'est pas inclus dans les séquences amont de *col* du pFlyFos *col**. Ce résultat suggère l'existence d'un autre CRM actif dans l'aile et partiellement redondant avec le CRM amont. Il serait intéressant de déterminer s'il s'agit d'un « shadow enhancer » (Hong et al., 2008), conférant de la robustesse et/ou de la précision spatiale et temporelle à l'expression de Col dans le disque d'aile. L'analyse de l'expression larvaire des lignées GMR précédemment utilisées pour la recherche systématique des CRM de *col* dans l'embryon devrait rapidement apporter une réponse à la question de redondance des CRM actifs dans l'aile.

III.3 – ChIPseq : identifier de nouvelles cibles directes de Collier

Suite aux difficultés imprévues rencontrées lors de la mise en œuvre d'une stratégie ChAPseq, le cœur de mon travail de thèse a consisté à rechercher les cibles directes de Col au cours de la myogenèse embryonnaire par une méthode de ChIPseq classique à partir d'embryons entiers. L'objectif était d'identifier des effecteurs de l'identité du muscle DA3, c'est-à-dire des gènes qui, régulés par Col au cours de la formation du muscle DA3, lui confèrent sa morphologie particulière. Afin de se situer pendant la phase de réalisation du programme identitaire, nous avons choisi de travailler avec des embryons aux stades 13-14 (10-14h de développement), lorsque la cellule fondatrice fusionne avec des myoblastes naïfs pour former la fibre musculaire. Le programme identitaire est activé dans les noyaux des myoblastes intégrés dans la fibre musculaire en croissance, via la propagation de l'expression de Col qui joue un rôle de chef d'orchestre (Dubois et al., 2007). Le challenge technique était d'identifier des cibles de Collier alors qu'il n'est exprimé que dans 2 à 3 noyaux par muscle DA3 à ce stade. Notre stratégie initiale tenait compte de la disponibilité des résultats de ChIPseq Col obtenus par le consortium ModEncode à partir d'embryons entre 0 et 12h de développement (Nègre et al., 2011). Nous avions prévu de comparer les deux jeux de données, obtenues avec deux types d'anticorps anti-Col différents, et de se focaliser sur les gènes communs. Cependant cette comparaison n'a pas identifié de recouvrement significatif (seulement 7 régions génomiques communes aux deux jeux de données sont détectées, qui présentent à la fois un recouvrement partiel– cf. Annexe 3). Une recherche *de novo* de motifs n'a pas permis de retrouver un enrichissement significatif du motif EBF (TCCCnnGGGA ; (Treiber et al., 2010b)) dans le jeu de donnés ModEncode (cf. analyse en Annexe 3). Nous avons donc testé les anticorps utilisés par le consortium ModEncode par immunocytochimie sur embryons fixés et constaté l'absence de détection de la protéine Col. L'ensemble nous a donc conduits à ne pas pouvoir considérer les données ModEncode. L'identification *de novo* du motif de liaison d'EBF et sa présence sur 90% des séquences immuno-précipitées avec les anticorps monoclonaux du laboratoire indiquent au contraire la spécificité de nos expériences de ChIP Col. Nous avons donc décidé de poursuivre l'analyse des 373 gènes candidats issus de ces expériences. La recherche *de novo* de motifs enrichis sur les fragments immuno-précipités permet de définir le motif consensus de fixation de Col *in vivo* ; la conservation de ce motif– TCCCnnGGGA - entre Col et EBF confirme les données *in vitro* et reflète la conservation de structure du DBD des protéines COE au cours de l'évolution (Daburon et al., 2008). Une étude plus détaillée montre que 174 des 373 motifs considérés pour établir ce consensus en possèdent le cœur CCCnnGGG. Ce cœur est-il pour autant nécessaire

afin que le site soit fonctionnel ? En utilisant des gènes rapporteurs, (Dubois et al., 2007) ont montré que, bien que différent de la matrice COE définie *in vivo*, le site TGTCTGGGGA de fixation de Col sur le ^{L}CRM est fonctionnel, puisque sa mutation (TGTCTGCCCA) entraine une perte d'expression du gène rapporteur. C'est également le cas du motif fixé par Col sur le promoteur de *hh* (CCCCAATGGC) (Ntini and Wimmer, 2011). Des variations par rapport au motif consensus semblent donc tolérées. Parmi les 15 gènes candidats dont les CRM ont été plus particulièrement étudiés (cf.II.2.2.f), 6 contiennent des motifs « dégénérés » par rapport au motif consensus (c'est-à-dire ayant au moins une mutation dans le triplet de C ou de G qui constituent le cœur du motif). La comparaison du patron d'expression observé avec les CRM sauvages ou mutés au niveau du motif Col nous offrira un panel plus large pour conclure quant à la relation motif-fonction. On peut d'ores et déjà noter que parmi les 6 CRM pour lesquels la mutation du site Col entraîne une modification d'expression du gène rapporteur dans le domaine Col, 5 possèdent un motif étendu, ou restreint. Collier se lie-t-il *in vivo* avec davantage d'affinité au motif consensus « étendu » de haute affinité défini *in vitro* pour EBF, le palindrome -ATTCCCnnGGGAAT- (Travis et al., 1993) ? C'est ce que suggère la recherche *de novo* de motifs enrichis à partir des 100 pics les plus élevés du ChIP-Col : les bases A et T en position 5' et 3' de la séquence cœur du motif CCCnnGGG qui s'en dégage deviennent prépondérantes, se rapprochant davantage de la matrice du motif consensus étendu (cf. Fig. R13-A). De plus, le remplacement du site endogène de fixation de Col sur le ^{L}CRM par le motif -ATTCCCnnGGGAAT- confère au gène rapporteur un plus haut niveau d'expression dans le muscle DA3 (L. Dubois, données non publiées). La comparaison de l'enrichissement relatif des fragments immunoprécipités, mesuré par qPCR, montrant un enrichissement environ deux fois supérieur pour le motif consensus étendu comparé au motif sauvage appuie cette conclusion. Une expression ectopique est par ailleurs observée dans le muscle DA2 lorsque le site sauvage du ^{L}CRM est remplacé par le site consensus étendu, un muscle dans lequel *col* est normalement réprimé par Tup (Boukhatmi et al., 2012). La nature du site participe donc à une répression efficace de sa cible dans ce lignage. L'ensemble de ces données montre que la séquence du site lié par Col *in vivo* est un paramètre important de la régulation par ce FT. Tous les motifs de type consensus étendus sont-ils pour autant fixés par Col *in vivo* ? La recherche du motif –ATTCCCnnGGGAAT- sur l'ensemble du génome de la drosophile montre qu'il est présent 8 fois. Sur les 8, un seul (le CRM associé au *cg12484*) est retrouvé parmi les 413 fragments immuno-précipités, indiquant que le contexte dans lequel se situe un site de fixation pour Collier est plus déterminant encore, en accord avec le caractère tissu-spécifique des cibles de Col actuellement connues. Ainsi la surexpression de Col dans tout le mésoderme ne permet la réactivation du gène

endogène ou du gène rapporteur ^{L}CRM-lacZ que dans 2 muscles, le DA2 et le LL1 et non pas dans tous les muscles comme on aurait pu s'y attendre (Crozatier and Vincent, données non publiées ; (Dubois et al., 2007). La fixation de Col, ou tout du moins l'activation d'un CRM « Col-dépendant » n'est donc possible que dans un contexte cellulaire donné : dans les noyaux où Col est exprimé, la présence de sites de fixation de haute affinité n'est pas suffisante pour recruter Col, et dans les noyaux où Col n'est pas exprimé, l'ajout de ce facteur ne permet pas forcément d'activer ses cibles révélées dans d'autres contextes. Cette observation est vérifiée dans plusieurs tissus : dans le cas d'une surexpression de Col dans tous les neurones du SNC, seuls quelques neurones ont la capacité de réactiver les gènes *eya* et *ap*, cibles de Col dans le lignage où Col est normalement exprimé (Baumgardt et al., 2007). De même, *ppk*, cible de Col dans les neurones multidendritiques de classe IV, n'est réactivé que dans un nombre restreint de neurones du PNS après surexpression de Col dans tous les neurones avec un pilote Elav>gal4 (Crozatier and Vincent, 2008). En outre un transgène rapporteur contenant 3 copies en tandem du motif –ATTCCCnnGGGAAT- n'est exprimé dans aucune cellule exprimant Col (D. Valle et J. Oyallon, résultats non publiés), montrant que des sites de haute affinité ne sont pas suffisants pour recruter Col à toutes épreuves. La fixation de ce facteur sur son site dépend donc probablement de la grammaire des CRM et de leur état chromatinien en fonction du type cellulaire et donc de la présence éventuelle de co-facteurs de Col. Et si l'affinité de Col pour son site est variable suivant le motif fixé, j'entends par là la stabilité de la liaison Col-ADN, la variation du motif contribue peut-être à une régulation plus fine de la transcription des gènes cibles de Col, à sa dynamique, avec des gènes transcrits de manière plus stable lorsque la fixation de Col est stabilisée, et de manière plus dynamique pour des CRM possédant un motif Col plus éloigné du motif consensus. Afin de vérifier cette hypothèse, une étude quantitative du niveau de transcription en fonction de la nature du site occupé par Col pour un CRM donné reste à faire. La modification du motif Col sur le ^{L}CRM en son motif consensus palindromique au sein du pFlyFos a été prévue en ce sens.

La recherche *de novo* de motifs sur les séquences des fragments liés par Col *in vivo* n'a pas mis en évidence d'autres motifs enrichis que le site de Col qui nous auraient permis de définir une « grammaire » des éléments cis-régulateurs musculaires de Col. Si cette grammaire existe, il se peut qu'elle soit masquée dans notre collection de fragments à cause de l'expression de Col dans plusieurs tissus aux stades choisis pour notre étude, dont le SNC et le muscle, auxquels il faut ajouter la tête en raison de la présence dans notre collection de quelques embryons aux stades 11-12. En effet le mélange de CRM dépendant de l'activité de Col dans ces 3 tissus peut diluer le code propre à chacun des tissus. Notre tentative de répartition tissulaire de ces CRM d'après la littérature sur leur gène cible prédit, en les associant soit au système nerveux soit au muscle, n'a

pas plus permis d'identifier des motifs spécifiques à chaque groupe (cf. II.2.2.d). L'association erronée des CRM à un gène ou bien une fonction différente de ces gènes sous le contrôle de Col pourrait en être la cause. L'objectif est donc de reprendre cette recherche en intégrant, dans un premier temps, les résultats obtenus pour les CRM candidats validés *in vivo*, et dans un deuxième temps les CRM issus du ChAPseq qui, par définition, devraient être tissu-spécifiques. Pour l'instant, la majorité des CRM testés sont des CRM « composites », exprimés dans plusieurs tissus, comme c'est le cas pour des versions initiales des ^{E}CRM et ^{L}CRM. De même que pour les CRM de *col*, il pourra donc être nécessaire de restreindre d'abord ces modules en isolant les éléments requis dans chacun des tissus avant d'entreprendre la recherche d'une grammaire spécifique. Une deuxième hypothèse est qu'il n'y ait pas de grammaire claire des CRM de Col suivant le tissu dans lequel il est exprimé. Notre modèle de CRM à grammaire définie est en fait fortement basé sur ce que nous connaissons du ^{L}CRM, et du rôle de la combinatoire « DA3 » (Col, Nau, Hox), mais ce modèle n'est peut-être pas généralisable à l'ensemble des CRM en aval de Col. Par ailleurs, un phénomène de recrutement collectif des facteurs de transcription pourrait rendre une grammaire difficilement lisible. Une observation intéressante concernant les CRM musculaires des gènes candidats déjà analysés mérite d'être notée. Ils sont actifs dans plusieurs muscles issus du groupe promusculaire exprimant Col. En plus du muscle DA3, le CRM aret[4.76] parait être actif spécifiquement dans le muscle LL1, le CRM eya[3.91] dans le muscle DT1, le CRM so[5.75] dans le muscle DO5 et le CRM tkv[8.27] dans le muscle DA2 (cf. Fig. I2). Cette identification, simplement inférée d'après la morphologie et la position de ces muscles, reste cependant à valider par l'intermédiaire de marqueurs spécifiques. Mais de cette expression différentielle ressurgit l'hypothèse de l'intégration de combinatoires de facteurs de transcription identitaires au niveau des CRM des gènes réalisateurs. Dans ce contexte, il serait intéressant de vérifier si par exemple Krüppel, FTi du LL1, n'agirait pas sur le CRM aret[4.76] en combinaison avec Col ou si S59, FTi du DT1, n'agirait pas sur le CRM eya[3.91]… (Enriquez et al., 2012) et de chercher leur site de fixation éventuel au sein de ces CRM. De la même manière, la recherche de motifs Nau sur les CRM permettant une expression dans le DA3 pourrait nous permettre de généraliser ou non le modèle du ^{L}CRM aux autres cibles de Col dans ce muscle.

Quelques gènes cibles putatifs de Col étaient déjà connus dans la littérature. On trouve *cnc* (Crozatier et al., 1999) et *hh* (Ntini and Wimmer, 2011) dans la tête, *col* lui-même dans le muscle DA3, *eya*, *ap*, *dimm, nplp1* et *dop*R dans le SNC (Baumgardt et al., 2007) et *ppk* dans les neurones multidendritiques de classe IV (Crozatier and Vincent, 2008). La régulation directe de ces gènes par Col n'a néanmoins été montrée que pour *hh* ; ce gène ne fait pas partie des candidats identifiés dans nos expériences de ChIP. Comme nous l'avons dit plus haut, le motif de fixation

de Col dans le CRM de *hh* est un motif éloigné du consensus, sa fixation par Col n'est donc peut-être que transitoire et n'a pas pu être détectée dans notre cas, d'autant plus que la fenêtre temporelle d'expression de *hh* dans la tête correspond plutôt à des embryons aux stades 10-11, sous-représentés dans notre collection. Alors qu'il est exprimé dans la même fenêtre spatio-temporelle, le gène *cnc* fait par contre partie des candidats identifiés dans notre ChIP, avec le 3^{e} pic le plus haut (enrichissement = 12.62). Contrairement à *hh*, le motif de fixation de Col prédit dans le CRM cnc[12.62] est un motif consensus étendu (ATTCCCCAGGGACC), ce qui pourrait expliquer l'enrichissement différentiel observé. L'identification de ce CRM reproduisant l'expression de *cnc* et contenant un site de fixation Col qui parait fonctionnel établit que la régulation de *cnc* par Col doit être directe (validation en cours). La fixation de Col sur son propre CRM au sein du progéniteur du DA3 reste à définir plus précisément. En effet, les séquences obtenues au niveau du promoteur de Col ont dû être retravaillées manuellement afin de distinguer la présence de pics à cette position car la présence résiduelle de l'ADNc ayant servi à produire la protéine recombinante Col utilisée lors de l'élution des fragments a entraîné une surreprésentation de ces séquences au niveau du gène *col*. La détection des pics se faisant notamment par comparaison avec l'environnement proche, plus rien ne paraissait significatif aux abords des exons de *col* présentant un enrichissement extraordinaire… Après traitement, on distingue bien 4 pics aux abords du promoteur de *col*, dont l'un, le plus significatif, correspond parfaitement au site d'autorégulation précédemment défini. Il faut tout de même souligner que l'alignement des séquences au fur et à mesure du séquençage de nos échantillons n'a pas pris en compte la présence du transgène ^{L}CRM-Col^{Cons}-lacZ dans nos fragments et ces séquences partiellement redondantes avec le ^{L}CRM de *col* ne peuvent donc pas en être distinguées… Cette « boîte noire » devra donc être tout particulièrement étudiée avec la mise au point du ChAPseq. Parmi les quelques gènes cibles connus de Col dans le SNC et présents parmi les candidats issus du ChIPseq, on trouve *eya* et *ap*, qui codent pour des facteurs de transcription participant à la définition de l'identité de sous-types de neurones, et potentiellement *nplp1*, gène d'identité « terminale » de ces neurones (l'attribution du pic étant plutôt en faveur du gène zip, plus proche...) (Baumgardt et al., 2007). Le gène *eya* participe également à la formation des muscles durant l'embryogenèse (Liu et al., 2009) ; Col régule directement cette expression d'*eya* , par l'intermédiaire d'un CRM situé environ 10 kb en amont de celui précédemment identifié pour la régulation par Tin. Les gènes *dimm* et *dop*R ne font pas partie des candidats cibles directs de Col identifiés par ChIPseq. L'hypothèse la plus probable est que Col régule ces gènes de manière indirecte. Enfin *ppk*, marqueur de différenciation finale des neurones multidendritiques de classe IV, n'est pas non plus présent parmi les cibles identifiées, mais cela pourrait être dû à son

expression plus tardive (on ne peut observer de transcrits *ppk* qu'à partir du stade 15). Des cibles directes de EBF-1 (Treiber et al., 2010b), EBF-2 (Rajakumari et al., 2013) et de Unc-3 (Kratsios et al., 2012), les orthologues de Col chez les vertébrés et C.*elegans* , respectivement, ont également été identifiées. L'exploitation de ces données dans le cadre de la recherche de gènes cibles de Col durant la myogenèse (ou à défaut dans le SNC ou la tête !) était cependant difficile, ces recherches ayant été réalisées dans des tissus autres que les muscles (lymphocytes B pour EBF-1, tissu adipeux brun pour EBF-2) ou motoneurones en phase de différentiation terminale dans C. *elegans*). Une des conclusions tirées par les auteurs de leur étude d'Unc-3 est que les protéines COE pourraient être des régulateurs de l'identité des motoneurones cholinergiques, une fonction conservée au cours de l'évolution entre les protostomes et les deutérostomes. Cette conclusion semble en contradiction avec une étude de (Demilly et al., 2011), montrant que Collier n'est pas exprimé dans les motoneurones mais dans des sous-types d'interneurones dans l'embryon de drosophile, dont seul un sous-ensemble est potentiellement cholinergique. Néanmoins, les gènes *ace*, *cha* et *vacht* ayant été identifiés comme gènes cibles directes de Unc-3 dans les motoneurones cholinergiques de C. *elegans* et faisant également partie des 413 gènes cibles de Col initialement considérés, j'ai tiré parti de l'existence de lignées GMR (*ace* : GMR55F07 – *cha*/*VaChT* : GMR59A07) pour tester l'activité des fragments contenant le pic Col et leur recouvrement avec l'expression de Col dans les neurones. Ces deux lignées permettent effectivement l'expression du gène rapporteur dans le SNC de l'embryon mais le recouvrement de cette expression avec les neurones Col positifs ne concerne qu'un ou deux neurones. Il faut noter que le pic associé au gène *ace* ne possède pas de site Col d'après la prédiction MEME. Je n'ai donc pas étudié plus avant la possibilité que ces gènes soient des cibles directes de Col mais cette analyse devrait être reprise dans le cadre d'une étude approfondie de la conservation de fonctions de Col/Unc-3.

Une analyse globale des fonctions prédites des gènes candidats issus du ChIP Col a révélé un grand nombre de facteurs de transcription (cf. Fig. R12 : 81 protéines se fixant à l'ADN, 12 participant à une régulation positive des gènes…). Au contraire d'Unc-3 dans les motoneurones cholinergiques de C. *elegans*, Col n'agirait donc pas comme sélecteur terminal du processus d'acquisition de l'identité musculaire mais plutôt comme l'initiateur d'un réseau de régulation, avec un niveau de régulation supplémentaire entre Col et les gènes réalisateurs terminaux de cette identité musculaire. La mise en place de ce niveau de régulation pourrait avoir lieu au stade progéniteur musculaire puisque c'est à ce stade que l'on observe par exemple la co-expression de Col et des gènes *eya*, *so*, *jing*... La proportion importante de facteurs de transcription dans les cibles directes de Col pourrait indiquer un processus de « feedforward regulation » des gènes effecteurs, comme mis en évidence dans le SNC (Baumgardt et al., 2007).

Oaz est le seul interacteur direct connu pour les EBF. Or il se trouve que plusieurs pics issus du ChIP Col, dont celui de plus fort enrichissement (22.41 !) se trouvent dans la région intergénique en amont des gènes phyl et Oaz ! L'attribution de ces pics à l'un ou l'autre des gènes n'a pas encore pu être tranchée. S'il s'avérait que Col régulait directement Oaz, cela ajouterait encore un niveau de régulation au réseau d'interactions entre ces deux facteurs.

Aucun des candidats issus du ChIPseq analysés à ce jour n'est exprimé spécifiquement dans le muscle DA3. L'acquisition d'une identité propre à chaque muscle passerait donc davantage par la régulation différentielle de mêmes cibles dans des muscles différents que par l'activation de gènes spécifiques à chaque muscle. C'est la conclusion déjà proposée par L. Bataillé (Bataillé et al., 2010) qui montre que la régulation différentielle du processus de fusion dans différents muscles passe par la régulation des niveaux d'expression des gènes *mp20*, *pax* et *mspo* dans chacun de ces muscles, sous le contrôle des facteurs de transcription identitaires.

L'analyse des cibles directes d'EBF-1 dans les lymphocytes B (Treiber et al., 2010b) et d'EBF-2 dans le tissu adipeux brun (Rajakumari et al., 2013) suggèrent enfin un nouveau rôle possible des facteurs COE. En effet, EBF-1 et 2 semblent se fixer précocement à leurs cibles et contribuer ainsi au recrutement d'autres facteurs de transcription nécessaires à la transcription de ces cibles (dont Pax5 dans le cas d'EBF-1 et Pparγ dans le cas d'EBF-2). Ce rôle éventuel de « priming » pour la transcription reste encore à explorer pour Col.

Conclusions et perspectives

L'analyse globale de nos données de ChIPseq a permis d'identifier des gènes cibles de Collier dans plusieurs tissus différents. Elle a permis de dégager un motif nucléotidique consensus lié par Collier *in vivo*. Ce motif est un élément essentiel de l'activité régulatrice de Col. Pour autant, les variations autour du consensus et la faible proportion de ces motifs liée par Col *in vivo* montrent que cette liaison est soumise à des mécanismes de régulation autres que la stricte reconnaissance de l'ADN. L'analyse de plusieurs CRM confirme que la liaison de Col est dépendante du type cellulaire et donc de l'état chromatinien des gènes et/ou d'autres facteurs de transcription présents dans ce type cellulaire. Cependant, aucune cis-grammaire spécifique pouvant rendre compte de cette liaison contextuelle n'a pu être mise en évidence ; la diversité des tissus dans lesquels Col est exprimé au cours de l'embryogenèse rend probablement illusoire la recherche d'une grammaire spécifique en l'état, sauf à devoir caractériser fonctionnellement un très grand nombre de CRM différents. Une solution d'avenir que j'ai initiée est une stratégie d'identification expérimentale des cibles de Col qui soit spécifique d'un tissu et d'un stade de développement

précis (ChAPseq). La proportion importante de facteurs de transcription parmi les cibles de Col suggère des mécanismes de contrôle combinatoire, y compris de « *feedforward* » pour la mise en place de l'identité musculaire. Si tel est le cas, la recherche de CRM dépendant de Col bénéficierait d'une analyse de ChAPseq (lignage-spécifique) pour ces autres facteurs. Enfin, comprendre le contrôle transcriptionnel de l'identité musculaire nécessite de progresser sur le rôle des effecteurs de cette identité. Les analyses de ChAPseq seront donc jointes à un programme de recherche « intégré » comprenant des analyses de transcriptomes de muscles spécifiques et un crible génétique « phénotypique » à grande échelle. Etablir des connexions précises entre régulation transcriptionnelle et acquisition d'une morphologie propre à chaque muscle reste un enjeu d'actualité.

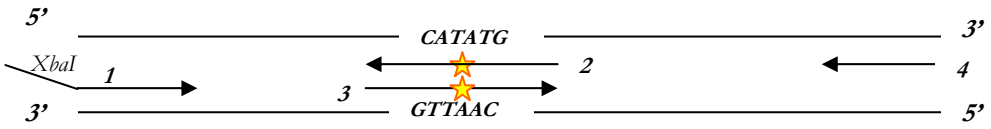
IV - Matériel et Méthodes

Découpage du ^{E}CRM

Le CRM276 (2393 pb) a été découpé en 2 fragments de 1010 pb (CRM276-A) et de 1383 pb (CRM276-B) (Fig. R1-A) au niveau du site de restriction pour l'enzyme *BglII*.

Mutation du site Twi sur le ^{L}CRM

La mutation du site Twi 5'- CATATG -3' en site inactif 5'- GTTAAC☆ -3' dans le fragment 2.6-0.9 du ^{L}CRM (fragment compris entre -0.9 et -2.6 kb en amont du TSS de *col*) (cf. Fig.R2) a été réalisée par la méthode des 4 oligonucléotides avec les oligonucléotides suivants :



1 = 5'- GCTCTAGAAGCGAGCTGGAAAC (ajout du site XbaI)

2 = 5'- CAACTCCGTC**GTTAAC**☆CAAGACATTC – 3'

3 = 5'- GAATGTCTTG**GTTAAC**☆GACGGAGTTG – 3'

4 = 5'- CAATCGGAAGGACTCTCATCGCC – 3'

Construction de gènes rapporteurs pour l'étude de sites prédits de fixation des protéines Hox sur le ^{L}CRM

Les mutations/ délétions *Hox5-Hox8* de sites de liaison des protéines Hox sur le ^{L}CRM, prédits par bio-informatique, ont été introduites au sein du fragment compris entre -0.9 et -4 kb en amont du SIT de *col* (cf. Fig.R2), placé en amont du rapporteur lacZ. Les modifications apportées sont les suivantes :

- *Hox5* = délétion du site Hox4 couplée à une mutation du site Hox3
- *Hox6* = délétion des sites Hox4 et AbdB couplées avec une mutation du site Hox3

- *Hox7* = délétion du site Hox4
- *Hox6* = délétion des sites Hox4 et AbdB

La mutation du site Hox3 est la suivante 5'- TAATTA –3' > 5'- GGGGTA -3'.

Stratégie de recombineering du FlyFos Col

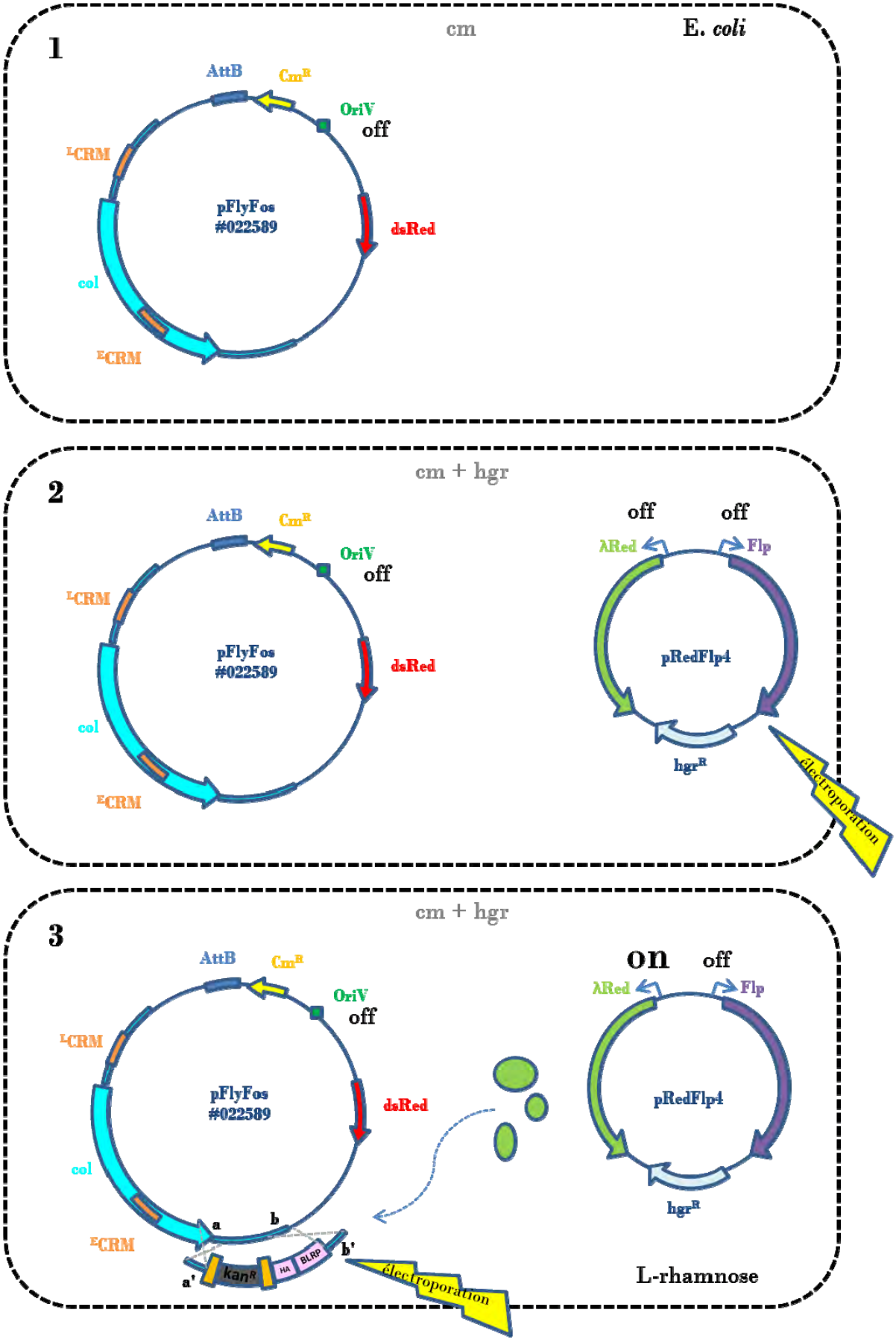
L'utilisation d'un pFlyFos (Ejsmont et al., 2009) contenant la région transcrite du gène *col* et 10kb de séquences amont (pFlyFos # 022589) permet d'étudier des CRM musculaires dans un contexte génomique reconstitué. Les fosmides FlyFos permettent d'utiliser le recombineering, c'est-à-dire différentes méthodes de recombinaisons homologues (système Red du phage λ) ou site-spécifiques (Flp/FRT) utilisées séquentiellement dans la bactérie *E.coli* puis pour le système – AttP-AttB pour l'intégration site-spécifique dans le génome de la drosophile. Les étapes de recombinaison ont été adaptées du protocole proposé pour le FlyFos (Ejsmont et al., 2009), avec une préparation des bactéries compétentes légèrement différente et adaptée à des échantillons uniques plutôt qu'à des recombinaisons en « batch » (plaques de 96 puits) (Fig. MM1).

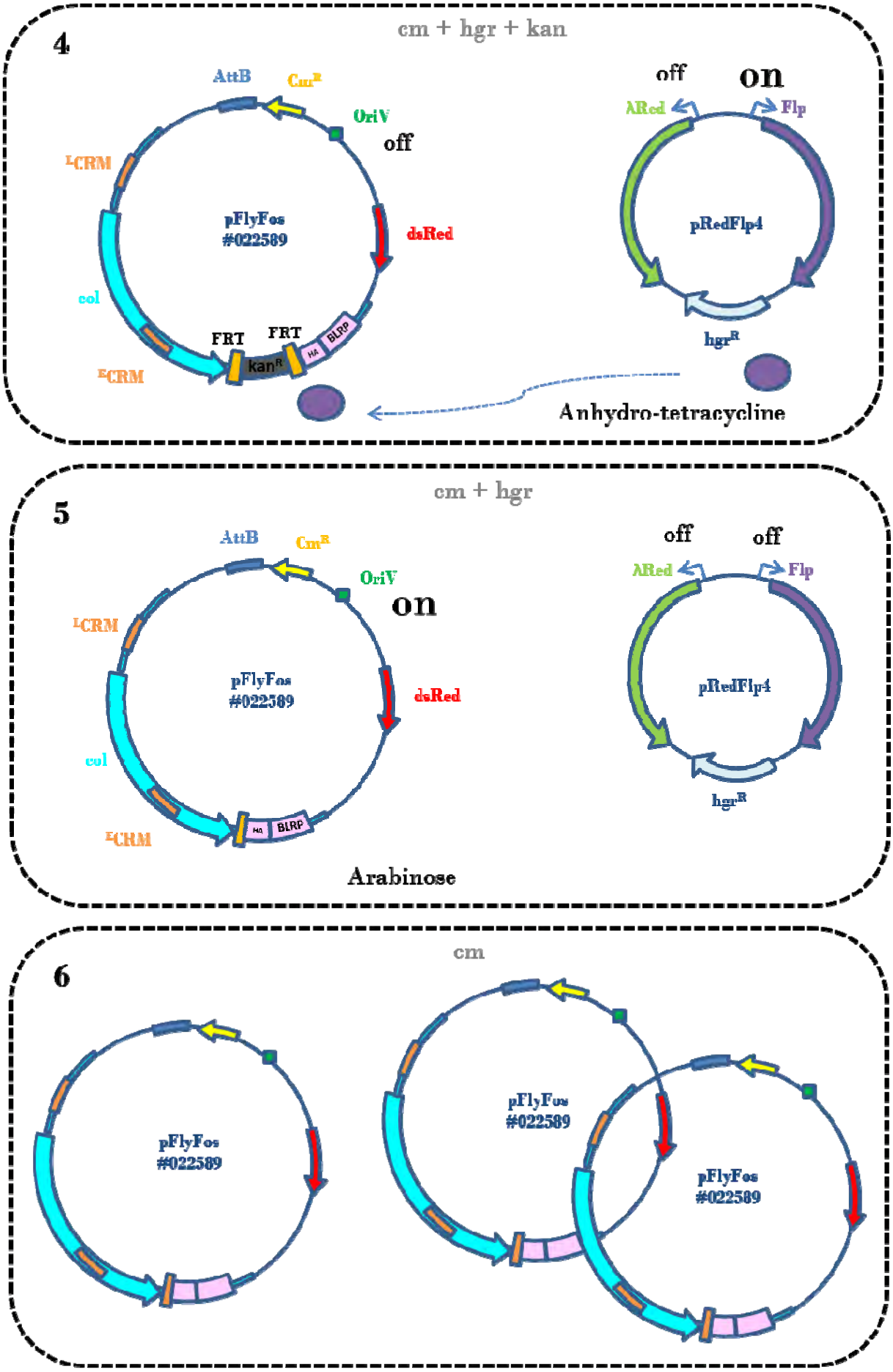
Nous avons choisi d'étiqueter la protéine ColB avec une étiquette Tev-HA-BLRP en position C-terminale. Le peptide HA doit permettre de visualiser spécifiquement la protéine Col transgénique et la différencier de la protéine Col endogène. Le peptide BLRP (**B**iotin **L**igase **R**ecognition **P**eptide) est reconnu par l'enzyme BirA comme site de biotinylation *in vivo* chez la drosophile. Enfin le motif Tev est un motif permettant le clivage spécifique de l'étiquette pour une élution spécifique de la protéine étiquetée. L'étiquette Tev-HA-BLRP a été introduite en C-terminal, juste en amont du codon Stop de l'isoforme Col B, afin d'éviter les erreurs lors des recombinaisons successives au niveau des CRM de *col* dues à la présence d'un site FRT, vestige de la recombinaison antérieure. Les régions d'homologie utilisée pour la recombinaison homologue sont les suivantes :

a' : 5' – CCACATCCACATCAGCCGTGGCACAATCCGGCCGTGTCAGCAGCCACGGCGGCGGCCGTT – 3'

b' : 5' – GTCCTCATCCTCATCCTCCGTTCAGCTCCATCTGGCTGGCGTTGGGAGTCCGGGAAATGC – 3'

L'étiquetage du gène *col* dans pFlyFos # 022589 a été effectuées dans *E. coli* transformée avec le plasmide pRedFlp4 qui porte à la fois l'opéron Red du phage λ et la flipase Fl (cf. Fig.MM1). Après validation par séquençage, l'ADN des fosmides intacts et modifiés est purifié et injecté dans des embryons de drosophile issues de lignées compétentes pour la transgénèse médiée par ΦC31 AttP/AttB (Bischof et al., 2007). Les mouches transgéniques sont sélectionnées par la présence du gène dsRed.


cm
E. coli
1
AttB
CmR
OriV
off
LCRM
pFlyFos
#022589
dsRed
col
ECRM
cm + hgr
2
AttB
CmR
OriV
off
LCRM
pFlyFos
#022589
dsRed
col
ECRM
off
off
λRed
Flp
pRedFlp4
hgrR
électroporation
cm + hgr
3
AttB
CmR
OriV
off
LCRM
pFlyFos
#022589
dsRed
col
ECRM
a
b
b'
a'
kanR
HA
BLRP
électroporation
on
off
λRed
Flp
pRedFlp4
hgrR
L-rhamnose


4
cm + hgr + kan
AttB
Cm^R
OriV
off
LCRM
pFlyFos
#022589
dsRed
col
FRT
FRT
HA
BLRP
kan^R
ECRM
off
on
λRed
Flp
pRedFlp4
hgr^R
Anhydro-tetracycline
5
cm + hgr
AttB
Cm^R
OriV
on
LCRM
pFlyFos
#022589
dsRed
col
HA
BLRP
ECRM
off
off
λRed
Flp
pRedFlp4
hgr^R
Arabinose
6
cm
pFlyFos
#022589
pFlyFos
#022589
pFlyFos
#022589

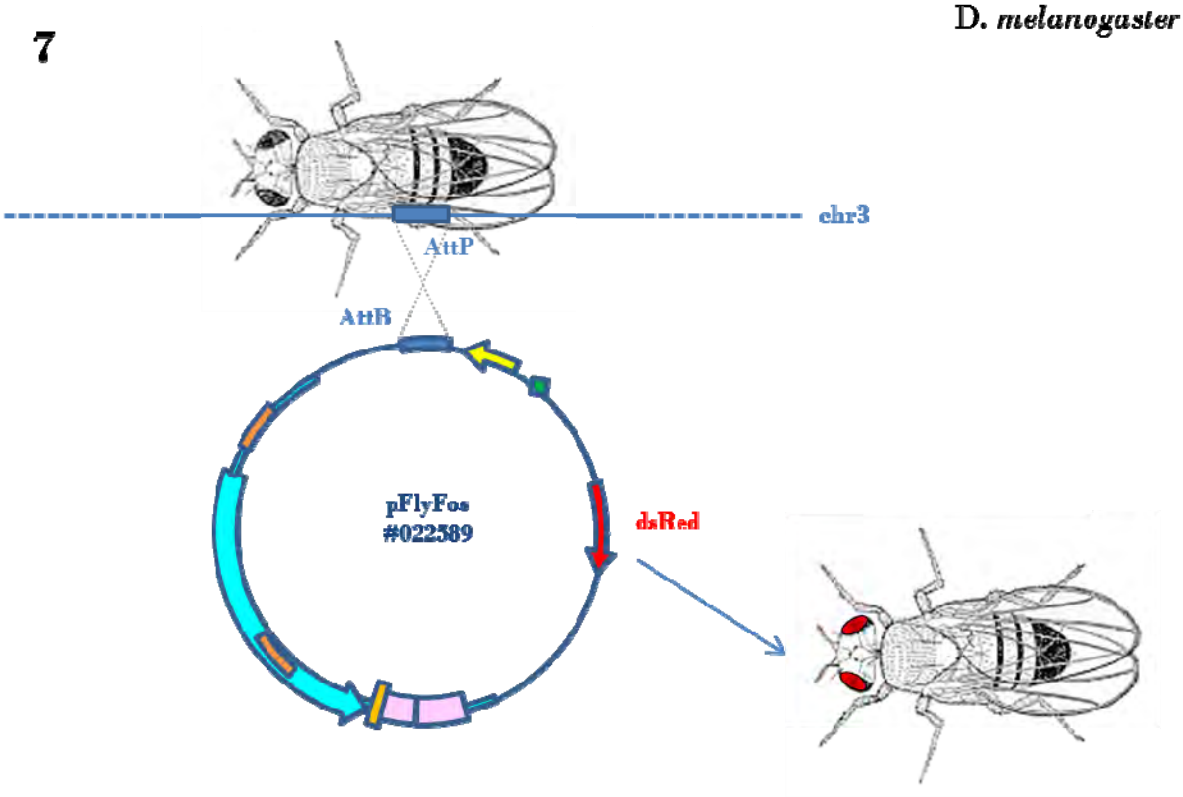


Fig. MM1 – Les étapes du recombineering

Le recombineering est la succession des étapes de recombinaison homologue ou site-spécifique permettant de modifier un fosmide dans la bactérie (E. *coli*) puis de l'intégrer dans le génome de la drosophile. Sur cet exemple, la recombinaison a pour objectif d'intégrer une étiquette HA-BLRP (rose) en C-terminal de la protéine Col dans le pFlyFos *col*. 7 étapes sont nécessaires. **1**. Transformation d'une souche E.*coli* avec le fosmide pFlyFos *col* portant le gène de résistance au chloramphénicol. (Esjmont *et al*. 2009). **2**. Électroporation du plasmide pRedFlp4 (Esjmont *et al*. 2009), comportant l'opéron Red du phage **λ,** le gène de la flippase Flp et le gène de résistance à l'hygromycine suivie de l'induction au L-rhamnose du système Red du phage **λ** permettant la recombinaison homologue via l'action des 3 protéines Exo, Bet et Gam, puis **3**. Électroporation de la cassette à intégrer, dont les extrémités a' et b' sont homologues à la région ciblée pour la recombinaison (a-b) ; cette cassette contient un gène de résistance à la kanamycine afin de sélectionner les évènements de recombinaison **4**. Activation de la Flp par ajout d'anhydro-tétracycline pour permettre la délétion du gène de résistance à la kanamycine flanqué de 2 sites FRT et la mise en phase de lecture de l'étiquette et de la protéine Col. **5**. Activation de l'origine de réplication OriV à l'arabinose pour obtenir **6**. un haut niveau de copies de pFlyFos *col** avant purification et **7**. Injection dans une lignée de drosophile possédant une plateforme de recombinaison AttP. L'intégration du fosmide dans le génome de la drosophile peut être sélectionnée grâce à la présence du gène dsRed.

Nota : (i) Si d'autres modifications doivent être réalisées par recombinaison homologue, le plasmide predFlp4 doit être maintenu dans la bactérie à l'étape 5.

(ii) Les différents fragments d'ADN ne sont pas représentés à l'échelle.

Caractérisation des anticorps monoclonaux anti-Col

Afin de caractériser les épitopes reconnus par les différents anticorps monoclonaux anti-Col, différentes protéines Col ont été générées à partir d'une matrice d'ADNc codant pour la protéine ColB complète étiquetée avec un motif HA en position N-terminale. Une digestion de la matrice avec les enzymes de restriction EcoRV, BseRI, BglII, SphI, NsiI, ClaI a permis de générer diverses formes de la protéine Col plus ou moins raccourcies en C terminal, en utilisant le kit de synthèse protéique TnT germe de blé (TnT® Coupled Wheat Germ Extract System – Promega). Chaque anticorps a été testé par Western-blot sur ces différents formes protéiques.

Immuno-précipitation de la chromatine (ChIP)

L'immuno-précipitation de la chromatine avec les Ac Col a été réalisée suivant le protocole décrit par (Sandmann et al., 2006a), avec quelques modifications.

Le protocole de Sandmann *et al.* a été optimisé pour des anticorps polyclonaux. Pour se rapprocher de ces conditions, nous avons choisi d'utiliser un mélange d'anticorps reconnaissant deux épitopes différents et de diminuer la force des lavages (tampon RIPA plus doux) puisque la spécificité de liaison des anticorps monoclonaux est supérieure à celle des polyclonaux. Pour garder les mêmes conditions pour l'expérience contrôle, nous avons choisi un anticorps anti-HA monoclonal (HA.11 Clone 16B12 – Covance).

Le temps de fixation ADN/protéine (« crosslinking ») nécessaire pour une expérience de ChIP dépend de la protéine considérée. Nous avons testé 2 temps de fixation (15 minutes et 25 minutes) et vérifié l'enrichissement final en fragments d'ADN reconnus par Col par qPCR avec les deux sites de fixation connus de Col (sur le CRM tardif de *col* lui-même, endogène ou transgénique). Un temps de fixation de 25 minutes est apparu optimal.

La taille des fragments immuno-précipités doit se situer entre 200 et 400 pb avant séquençage. Afin d'obtenir une fragmentation homogène de la chromatine en fragments de cette taille plusieurs tests de paramétrages du sonicateur ont été nécessaires pour tenir compte du volume des échantillons, du temps de fixation des embryons… Finalement les échantillons ont été soumis à 25 cycles d'ultrasons [25s ON – 45s OFF] en High Level Energy (Bioruptor Diagenode), dans un bain maintenu à 3°C.

Une des difficultés majeures de cette expérience venait du faible nombre de noyaux par embryons exprimant Collier (environ 200, soit moins de 0.1% des noyaux totaux…). Afin

d'augmenter le rendement d'immuno-précipitation et d'enrichir au maximum les fragments précipités en fragments fixés par Col, nous avons opté pour une élution compétitive par ajout de protéine Collier recombinante au moment de l'élution des fragments. L'élution compétitive nous a permis d'améliorer considérablement l'enrichissement des fragments immuno-précipités en fragments fixés par Col (l'enrichissement passe de 2 à 3 en prenant le locus rp49 comme référence). Cependant, cette étape a également introduit une contamination de nos fragments d'ADN immuno-précipités par les restes d'ADN matrice de la protéine Col recombinante. Il faut donc prévoir plusieurs étapes de purification de la protéine recombinante afin d'éliminer toutes traces du plasmide.

V - Bibliographie

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

Bally-Cuif, L., Dubois, L., Vincent, A., 1998. Molecular cloning of Zcoe2, the zebrafish homolog of Xenopus Xcoe2 and mouse EBF-2, and its expression during primary neurogenesis. Mech Dev 77, 85-90.

Bataillé, L., Delon, I., Da Ponte, J.P., Brown, N.H., Jagla, K., 2010. Downstream of identity genes: muscle-type-specific regulation of the fusion process. Dev Cell 19, 317-328.

Bate, M., 1990. The embryonic development of larval muscles in Drosophila. Development 110, 791-804.

Bate, M., Rushton, E., 1993. Myogenesis and muscle patterning in Drosophila. C R Acad Sci III 316, 1047-1061.

Baumgardt, M., Miguel-Aliaga, I., Karlsson, D., Ekman, H., Thor, S., 2007. Specification of neuronal identities by feedforward combinatorial coding. PLoS Biol 5, e37.

Baylies, M.K., Bate, M., 1996. twist: a myogenic switch in Drosophila. Science 272, 1481-1484.

Baylies, M.K., Bate, M., Ruiz Gomez, M., 1998. Myogenesis: a view from Drosophila. Cell 93, 921-927.

Beckett, D., Kovaleva, E., Schatz, P.J., 1999. A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. Protein Sci 8, 921-929.

Bejerano, G., Siepel, A.C., Kent, W.J., Haussler, D., 2005. Computational screening of conserved genomic DNA in search of functional noncoding elements. Nat Methods 2, 535-545.

Bellen, H.J., O'Kane, C.J., Wilson, C., Grossniklaus, U., Pearson, R.K., Gehring, W.J., 1989. P-element-mediated enhancer detection: a versatile method to study development in Drosophila. Genes Dev 3, 1288-1300.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.B., 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A 99, 757-762.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., Celniker, S.E., 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol 5, R61.

Bettencourt, R., Ip, Y.T., 2004. Learning the codes of fly immunity. Mol Cell 13, 1-2.

Biggin, M.D., Tjian, R., 1989. Transcription factors and the control of Drosophila development. Trends Genet 5, 377-383.

Bischof, J., Maeda, R.K., Hediger, M., Karch, F., Basler, K., 2007. An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. Proc Natl Acad Sci U S A 104, 3312-3317.

Black, B.L., Olson, E.N., 1998. Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. Annu Rev Cell Dev Biol 14, 167-196.

Boettiger, A.N., Levine, M., 2009. Synchronous and stochastic patterns of gene activation in the Drosophila embryo. Science 325, 471-473.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., Furlong, E.E., 2012. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet 44, 148-156.

Boukhatmi, H., 2013. Mise en place de l'identité des muscles au cours de la spécification des myoblastes chez la Drosophile., thèse de Biologie moléculaire, cellulaire et du développement. Université Toulouse III - Paul Sabatier, Centre de Biologie du Développement - Toulouse, pp. 1-194.

Boukhatmi, H., Frendo, J.L., Enriquez, J., Crozatier, M., Dubois, L., Vincent, A., 2012. Tup/Islet1 integrates time and position to specify muscle identity in Drosophila. Development 139, 3572-3582.

Bourgouin, C., Lundgren, S.E., Thomas, J.B., 1992. Apterous is a Drosophila LIM domain gene required for the development of a subset of embryonic muscles. Neuron 9, 549-561.

Bryson-Richardson, R.J., Currie, P.D., 2008. The genetics of vertebrate myogenesis. Nat Rev Genet 9, 632-646.

Buckingham, M., 2006. Myogenic progenitor cells and skeletal myogenesis in vertebrates. Curr Opin Genet Dev 16, 525-532.

Buff, E., Carmena, A., Gisselbrecht, S., Jiménez, F., Michelson, A.M., 1998. Signalling by the Drosophila epidermal growth factor receptor is required for the specification and diversification of embryonic muscle progenitors. Development 125, 2075-2086.

Carmena, A., Bate, M., Jiménez, F., 1995. Lethal of scute, a proneural gene, participates in the specification of muscle progenitors during Drosophila embryogenesis. Genes Dev 9, 2373-2383.

Carmena, A., Gisselbrecht, S., Harrison, J., Jiménez, F., Michelson, A.M., 1998. Combinatorial signaling codes for the progressive determination of cell fates in the Drosophila embryonic mesoderm. Genes Dev 12, 3910-3922.

Clark, I.B., Jarman, A.P., Finnegan, D.J., 2007. Live imaging of Drosophila gonad formation reveals roles for Six4 in regulating germline and somatic cell migration. BMC Dev Biol 7, 52.

Corradi, A., Croci, L., Broccoli, V., Zecchini, S., Previtali, S., Wurst, W., Amadio, S., Maggi, R., Quattrini, A., Consalez, G.G., 2003. Hypogonadotropic hypogonadism and peripheral neuropathy in Ebf2-null mice. Development 130, 401-410.

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

Croci, L., Chung, S.H., Masserdotti, G., Gianola, S., Bizzoca, A., Gennarini, G., Corradi, A., Rossi, F., Hawkes, R., Consalez, G.G., 2006. A key role for the HLH transcription factor EBF2COE2,O/E-3 in Purkinje neuron migration and cerebellar cortical topography. Development 133, 2719-2729.

Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S., Vincent, A., 1996. Collier, a novel regulator of Drosophila head development, is expressed in a single mitotic domain. Curr Biol 6, 707-718.

Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S., Vincent, A., 1999. Head versus trunk patterning in the Drosophila embryo; collier requirement for formation of the intercalary segment. Development 126, 4385-4394.

Crozatier, M., Vincent, A., 1999. Requirement for the Drosophila COE transcription factor Collier in formation of an embryonic muscle: transcriptional response to notch signalling. Development 126, 1495-1504.

Crozatier, M., Vincent, A., 2008. Control of multidendritic neuron differentiation in Drosophila: the role of Collier. Dev Biol 315, 232-242.

Daburon, V., Mella, S., Plouhinec, J.L., Mazan, S., Crozatier, M., Vincent, A., 2008. The metazoan history of the COE transcription factors. Selection of a variant HLH motif by mandatory inclusion of a duplicated exon in vertebrates. BMC Evol Biol 8, 131.

Demilly, A., Simionato, E., Ohayon, D., Kerner, P., Garcès, A., Vervoort, M., 2011. Coe genes are expressed in differentiating neurons in the central nervous system of protostomes. PLoS One 6, e21213.

Dubois, L., 1999. Role d'une nouvelle famille de facteurs de transcription, les proteines coe, dans le processus de differenciation neuronale chez le xenope et le poisson zebre, Developmental biology. Université Paul Sabatier - Toulouse III, p. 159 p.

Dubois, L., Bally-Cuif, L., Crozatier, M., Moreau, J., Paquereau, L., Vincent, A., 1998. XCoe2, a transcription factor of the Col/Olf-1/EBF family involved in the specification of primary neurons in Xenopus. Curr Biol 8, 199-209.

Dubois, L., Enriquez, J., Daburon, V., Crozet, F., Lebreton, G., Crozatier, M., Vincent, A., 2007. Collier transcription in a single Drosophila muscle lineage: the combinatorial control of muscle identity. Development 134, 4347-4355.

Dubois, L., Vincent, A., 2001. The COE--Collier/Olf1/EBF--transcription factors: structural conservation and diversity of developmental functions. Mech Dev 108, 3-12.

Ejsmont, R.K., Sarov, M., Winkler, S., Lipinski, K.A., Tomancak, P., 2009. A toolkit for high-throughput, cross-species gene engineering in Drosophila. Nat Methods 6, 435-437.

Enriquez, J., Boukhatmi, H., Dubois, L., Philippakis, A.A., Bulyk, M.L., Michelson, A.M., Crozatier, M., Vincent, A., 2010. Multi-step control of muscle diversity by Hox proteins in the Drosophila embryo. Development 137, 457-466.

Enriquez, J., de Taffin, M., Crozatier, M., Vincent, A., Dubois, L., 2012. Combinatorial coding of Drosophila muscle shape by Collier and Nautilus. Dev Biol 363, 27-39.

Epstein, H.F., Waterston, R.H., Brenner, S., 1974. A mutant affecting the heavy chain of myosin in Caenorhabditis elegans. J Mol Biol 90, 291-300.

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

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., van Steensel, B., 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell 143, 212-224.

Fischer, J.A., Giniger, E., Maniatis, T., Ptashne, M., 1988. GAL4 activates transcription in Drosophila. Nature 332, 853-856.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., Jacq, B., Arpin, M., Bellaiche, Y., Bellusci, S., Benaroch, P., Bornens, M., Chanet, R., Chavrier, P., Delattre, O., Doye, V., Fehon, R., Faye, G., Galli, T., Girault, J.A., Goud, B., de Gunzburg, J., Johannes, L., Junier, M.P., Mirouse, V., Mukherjee, A., Papadopoulo, D., Perez, F., Plessis, A., Rossé, C., Saule, S., Stoppa-Lyonnet, D., Vincent, A., White, M., Legrain, P., Wojcik, J., Camonis, J., Daviet, L., 2005. Protein interaction mapping: a Drosophila case study. Genome Res 15, 376-384.

Fraga, D., Meulia, T. and Fenster, S., 2008. Real-Time PCR. Current Protocols - Essential Laboratory Techniques. Unit 10.3, 10.13.11–10.13.34.

Gaertner, B., Johnston, J., Chen, K., Wallaschek, N., Paulson, A., Garruss, A.S., Gaudenz, K., De Kumar, B., Krumlauf, R., Zeitlinger, J., 2012. Poised RNA polymerase II changes over developmental time and prepares genes for future expression. Cell Rep 2, 1670-1683.

Garel, S., Garcia-Dominguez, M., Charnay, P., 2000. Control of the migratory pathway of facial branchiomotor neurones. Development 127, 5297-5307.

Garel, S., Marín, F., Grosschedl, R., Charnay, P., 1999. Ebf1 controls early cell differentiation in the embryonic striatum. Development 126, 5285-5294.

Garel, S., Marín, F., Mattéi, M.G., Vesque, C., Vincent, A., Charnay, P., 1997. Family of Ebf/Olf-1-related genes potentially involved in neuronal differentiation and regional specification in the central nervous system. Dev Dyn 210, 191-205.

Gehring, W.J., 1985. Homeotic genes, the homeobox, and the spatial organization of the embryo. Harvey Lect 81, 153-172.

Green, Y.S., Vetter, M.L., 2011. EBF proteins participate in transcriptional regulation of Xenopus muscle development. Dev Biol 358, 240-250.

Grifone, R., Demignon, J., Giordani, J., Niro, C., Souil, E., Bertin, F., Laclef, C., Xu, P.X., Maire, P., 2007. Eya1 and Eya2 proteins are required for hypaxial somitic myogenesis in the mouse embryo. Dev Biol 302, 602-616.

Hagman, J., Belanger, C., Travis, A., Turck, C.W., Grosschedl, R., 1993. Cloning and functional characterization of early B-cell factor, a regulator of lymphocyte-specific gene expression. Genes Dev 7, 760-773.

Hagman, J., Gutch, M.J., Lin, H., Grosschedl, R., 1995. EBF contains a novel zinc coordination motif and multiple dimerization and transcriptional activation domains. EMBO J 14, 2907-2916.

Hagman, J., Lukin, K., 2005. Early B-cell factor 'pioneers' the way for B-cell development. Trends Immunol 26, 455-461.

Hagman, J., Travis, A., Grosschedl, R., 1991. A novel lineage-specific nuclear factor regulates mb-1 gene transcription at the early stages of B cell differentiation. EMBO J 10, 3409-3417.

Halfon, M.S., Carmena, A., Gisselbrecht, S., Sackerson, C.M., Jiménez, F., Baylies, M.K., Michelson, A.M., 2000. Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. Cell 103, 63-74.

Halfon, M.S., Zhu, Q., Brennan, E.R., Zhou, Y., 2011. Erroneous attribution of relevant transcription factor binding sites despite successful prediction of cis-regulatory modules. BMC Genomics 12, 578.

Han, Z., Li, X., Wu, J., Olson, E.N., 2004. A myocardin-related transcription factor regulates activity of serum response factor in Drosophila. Proc Natl Acad Sci U S A 101, 12567-12572.

Hartmann, H., Guthöhrlein, E.W., Siebert, M., Luehr, S., Söding, J., 2013. P-value-based regulatory motif discovery using positional weight matrices. Genome Res 23, 181-194.

Hattori, Y., Sugimura, K., Uemura, T., 2007. Selective expression of Knot/Collier, a transcriptional regulator of the EBF/Olf-1 family, endows the Drosophila sensory system with neuronal class-specific elaborated dendritic patterns. Genes Cells 12, 1011-1022.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E., Ren, B., 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39, 311-318.

Hens, K., Feuz, J.D., Isakova, A., Iagovitina, A., Massouras, A., Bryois, J., Callaerts, P., Celniker, S.E., Deplancke, B., 2011. Automated protein-DNA interaction screening of Drosophila regulatory elements. Nat Methods 8, 1065-1070.

Herrmann, C., Van de Sande, B., Potier, D., Aerts, S., 2012. i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. Nucleic Acids Res 40, e114.

Hersh, B.M., Carroll, S.B., 2005. Direct regulation of knot gene expression by Ultrabithorax and the evolution of cis-regulatory elements in Drosophila. Development 132, 1567-1577.

Hong, J.W., Hendrix, D.A., Levine, M.S., 2008. Shadow enhancers as a source of evolutionary novelty. Science 321, 1314.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44-57.

Ingham, P.W., 1988. The molecular genetics of embryonic pattern formation in Drosophila. Nature 335, 25-34.

Jackson, D.J., Meyer, N.P., Seaver, E., Pang, K., McDougall, C., Moy, V.N., Gordon, K., Degnan, B.M., Martindale, M.Q., Burke, R.D., Peterson, K.J., 2010. Developmental expression of COE across the Metazoa supports a conserved role in neuronal cell-type specification and mesodermal development. Dev Genes Evol 220, 221-234.

Jagla, T., Bellard, F., Lutz, Y., Dretzen, G., Bellard, M., Jagla, K., 1998. ladybird determines cell fate decisions during diversification of Drosophila somatic muscles. Development 125, 3699-3708.

Jimenez, M.A., Akerblad, P., Sigvardsson, M., Rosen, E.D., 2007. Critical role for Ebf1 and Ebf2 in the adipogenic transcriptional cascade. Mol Cell Biol 27, 743-757.

Jinushi-Nakao, S., Arvind, R., Amikura, R., Kinameri, E., Liu, A.W., Moore, A.W., 2007. Knot/Collier and cut control different aspects of dendrite cytoskeleton and synergize to define final arbor shape. Neuron 56, 963-978.

John, S., Sabo, P.J., Canfield, T.K., Lee, K., Vong, S., Weaver, M., Wang, H., Vierstra, J., Reynolds, A.P., Thurman, R.E., Stamatoyannopoulos, J.A., 2013. Genome-Scale Mapping of DNase I Hypersensitivity. Curr Protoc Mol Biol Chapter 27, Unit21.27.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.H., Birney, E., Furlong, E.E., 2012. A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell 148, 473-486.

Karlić, R., Chung, H.R., Lasserre, J., Vlahovicek, K., Vingron, M., 2010. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A 107, 2926-2931.

Keller, C.A., Grill, M.A., Abmayr, S.M., 1998. A role for nautilus in the differentiation of muscle precursors. Dev Biol 202, 157-171.

Kieslinger, M., Folberth, S., Dobreva, G., Dorn, T., Croci, L., Erben, R., Consalez, G.G., Grosschedl, R., 2005. EBF2 regulates osteoblast-dependent differentiation of osteoclasts. Dev Cell 9, 757-767.

Kieslinger, M., Hiechinger, S., Dobreva, G., Consalez, G.G., Grosschedl, R., 2010. Early B cell factor 2 regulates hematopoietic stem cell homeostasis in a cell-nonautonomous manner. Cell Stem Cell 7, 496-507.

Knirr, S., Frasch, M., 2001. Molecular integration of inductive and mesoderm-intrinsic inputs governs even-skipped enhancer activity in a subset of pericardial and dorsal muscle progenitors. Dev Biol 238, 13-26.

Kratsios, P., Stolfi, A., Levine, M., Hobert, O., 2012. Coordinated regulation of cholinergic motor neuron traits through a conserved terminal selector gene. Nat Neurosci 15, 205-214.

Krzemień, J., Dubois, L., Makki, R., Meister, M., Vincent, A., Crozatier, M., 2007. Control of blood cell homeostasis in Drosophila larvae by the posterior signalling centre. Nature 446, 325-328.

Kudrycki, K., Stein-Izsak, C., Behn, C., Grillo, M., Akeson, R., Margolis, F.L., 1993. Olf-1-binding site: characterization of an olfactory neuron-specific promoter motif. Mol Cell Biol 13, 3002-3014.

Lagergren, A., Månsson, R., Zetterblad, J., Smith, E., Basta, B., Bryder, D., Akerblad, P., Sigvardsson, M., 2007. The Cxcl12, periostin, and Ccl9 genes are direct targets for early B-cell factor in OP-9 stroma cells. J Biol Chem 282, 14454-14462.

Levy, A., Noll, M., 1981. Chromatin fine structure of active and repressed genes. Nature 289, 198-203.

Liberg, D., Sigvardsson, M., Akerblad, P., 2002. The EBF/Olf/Collier family of transcription factors: regulators of differentiation in cells originating from all three embryonal germ layers. Mol Cell Biol 22, 8389-8397.

Lin, H., Grosschedl, R., 1995. Failure of B-cell differentiation in mice lacking the transcription factor EBF. Nature 376, 263-267.

Lin, M.H., Bour, B.A., Abmayr, S.M., Storti, R.V., 1997. Ectopic expression of MEF2 in the epidermis induces epidermal expression of muscle genes and abnormal muscle development in Drosophila. Dev Biol 182, 240-255.

Liu, Y.H., Jakobsen, J.S., Valentin, G., Amarantos, I., Gilmour, D.T., Furlong, E.E., 2009. A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. Dev Cell 16, 280-291.

Maier, H., Ostraat, R., Gao, H., Fields, S., Shinton, S.A., Medina, K.L., Ikawa, T., Murre, C., Singh, H., Hardy, R.R., Hagman, J., 2004. Early B cell factor cooperates with Runx1 and mediates epigenetic changes associated with mb-1 transcription. Nat Immunol 5, 1069-1077.

Marygold, S.J., Leyland, P.C., Seal, R.L., Goodman, J.L., Thurmond, J., Strelets, V.B., Wilson, R.J., consortium, F., 2013. FlyBase: improvements to the bibliography. Nucleic Acids Res 41, D751-757.

Massari, M.E., Murre, C., 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. Mol Cell Biol 20, 429-440.

Mella, S., 2004. Histoire évolutive des facteurs de transcription de la famille COE : expression comparée au cours de la morphogenèse de l'aile de drosophile et des membres de tétrapodes et évolution de leur structure génomique chez les bilatériens, Developmental biology. Université Paul Sabatier - Toulouse III, p. 193 p.

Mella, S., Soula, C., Morello, D., Crozatier, M., Vincent, A., 2004. Expression patterns of the coe/ebf transcription factor genes during chicken and mouse limb development. Gene Expr Patterns 4, 537-542.

Miyata, T., Suga, H., 2001. Divergence pattern of animal gene families and relationship with the Cambrian explosion. Bioessays 23, 1018-1027.

Müller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., Simon, J.A., 2002. Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. Cell 111, 197-208.

Nelson, A.C., Wardle, F.C., 2013. Conserved non-coding elements and cis regulation: actions speak louder than words. Development 140, 1385-1395.

Nguyen, H.T., Bodmer, R., Abmayr, S.M., McDermott, J.C., Spoerel, N.A., 1994. D-mef2: a Drosophila mesoderm-specific MADS box-containing gene with a biphasic expression profile during embryogenesis. Proc Natl Acad Sci U S A 91, 7520-7524.

Ntini, E., Wimmer, E.A., 2011. Second order regulator Collier directly controls intercalary-specific segment polarity gene expression. Dev Biol 360, 403-414.

Nègre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R.F., Chen, J., Hwang, L., Cheng, C., Auburn, R.P., Davis, M.B., Domanus, M., Shah, P.K., Morrison, C.A., Zieba, J., Suchy, S., Senderowicz, L., Victorsen, A., Bild, N.A., Grundstad, A.J., Hanley, D., MacAlpine, D.M., Mannervik, M., Venken, K., Bellen, H., White, R., Gerstein, M., Russell, S., Grossman, R.L., Ren, B., Posakony, J.W., Kellis, M., White, K.P., 2011. A cis-regulatory map of the Drosophila genome. Nature 471, 527-531.

O'Riordan, M., Grosschedl, R., 1999. Coordinate regulation of B cell differentiation by the transcription factors EBF and E2A. Immunity 11, 21-31.

Pang, K., Matus, D.Q., Martindale, M.Q., 2004. The ancestral role of COE genes may have been in chemoreception: evidence from the development of the sea anemone, Nematostella vectensis (Phylum Cnidaria; Class Anthozoa). Dev Genes Evol 214, 134-138.

Pankratz, M., Jäckle, H., 1993. Blastoderm segmentation. In *The development of* Drosophila melanogaster, Cold Spring Harbor, New York. ed.

Papatsenko, D., Goltsev, Y., Levine, M., 2009. Organization of developmental enhancers in the Drosophila embryo. Nucleic Acids Res 37, 5665-5677.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B.L., Couronne, O., Eisen, M.B., Visel, A., Rubin, E.M., 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444, 499-502.

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

Pfeiffer, B.D., Jenett, A., Hammonds, A.S., Ngo, T.T., Misra, S., Murphy, C., Scully, A., Carlson, J.W., Wan, K.H., Laverty, T.R., Mungall, C., Svirskas, R., Kadonaga, J.T., Doe, C.Q., Eisen, M.B., Celniker, S.E., Rubin, G.M., 2008. Tools for neuroanatomy and neurogenetics in Drosophila. Proc Natl Acad Sci U S A 105, 9715-9720.

Philippakis, A.A., Busser, B.W., Gisselbrecht, S.S., He, F.S., Estrada, B., Michelson, A.M., Bulyk, M.L., 2006. Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells. PLoS Comput Biol 2, e53.

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

Pozzoli, O., Bosetti, A., Croci, L., Consalez, G.G., Vetter, M.L., 2001. Xebf3 is a regulator of neuronal differentiation during primary neurogenesis in Xenopus. Dev Biol 233, 495-512.

Prasad, B.C., Ye, B., Zackhary, R., Schrader, K., Seydoux, G., Reed, R.R., 1998. unc-3, a gene required for axonal guidance in Caenorhabditis elegans, encodes a member of the O/E family of transcription factors. Development 125, 1561-1568.

Rajakumari, S., Wu, J., Ishibashi, J., Lim, H.W., Giang, A.H., Won, K.J., Reed, R.R., Seale, P., 2013. EBF2 determines and maintains brown adipocyte identity. Cell Metab 17, 562-574.

Ranganayakulu, G., Zhao, B., Dokidis, A., Molkentin, J.D., Olson, E.N., Schulz, R.A., 1995. A series of mutations in the D-MEF2 transcription factor reveal multiple functions in larval and adult myogenesis in Drosophila. Dev Biol 171, 169-181.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C.O., Forrest, A.R., Gough, J., Grimmond, S., Han, J.H., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C.R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R.D., Tegnér, J., Lenhard, B., Teichmann, S.A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D.A., Ideker, T., Hayashizaki, Y., 2010. An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140, 744-752.

Rayapureddi, J.P., Kattamuri, C., Chan, F.H., Hegde, R.S., 2005. Characterization of a plant, tyrosine-specific phosphatase of the aspartyl class. Biochemistry 44, 751-758.

Richard, A.F., Demignon, J., Sakakibara, I., Pujol, J., Favier, M., Strochlic, L., Le Grand, F., Sgarioto, N., Guernec, A., Schmitt, A., Cagnard, N., Huang, R., Legay, C., Guillet-Deniau, I., Maire, P., 2011. Genesis of muscle fiber-type diversity during mouse embryogenesis relies on Six1 and Six4 gene expression. Dev Biol 359, 303-320.

Ries, D., Meisterernst, M., 2011. Control of gene transcription by Mediator in chromatin. Semin Cell Dev Biol 22, 735-740.

Ritter, D.I., Li, Q., Kostka, D., Pollard, K.S., Guo, S., Chuang, J.H., 2010. The importance of being cis: evolution of orthologous fish and mammalian enhancer activity. Mol Biol Evol 27, 2322-2332.

Roby, Y.A., Bushey, M.A., Cheng, L.E., Kulaga, H.M., Lee, S.J., Reed, R.R., 2012. Zfp423/OAZ mutation reveals the importance of Olf/EBF transcription activity in olfactory neuronal maturation. J Neurosci 32, 13679-13688a.

Rodriguez, P., Braun, H., Kolodziej, K.E., de Boer, E., Campbell, J., Bonte, E., Grosveld, F., Philipsen, S., Strouboulis, J., 2006. Isolation of transcription factor complexes by in vivo biotinylation tagging and direct binding to streptavidin beads. Methods Mol Biol 338, 305-323.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F., Washietl, S., Arshinoff, B.I., Ay, F., Meyer, P.E., Robine, N., Washington, N.L., Di Stefano, L., Berezikov, E., Brown, C.D., Candeias, R., Carlson, J.W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M.Y., Will, S., Alekseyenko, A.A., Artieri, C., Booth, B.W., Brooks, A.N., Dai, Q., Davis, C.A., Duff, M.O., Feng, X., Gorchakov, A.A., Gu, T., Henikoff, J.G., Kapranov, P., Li, R., MacAlpine, H.K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S.K., Riddle, N.C., Sakai, A., Samsonova, A., Sandler, J.E.,

Schwartz, Y.B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K.H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S.E., Brent, M.R., Cherbas, L., Elgin, S.C., Gingeras, T.R., Grossman, R., Hoskins, R.A., Kaufman, T.C., Kent, W., Kuroda, M.I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J.W., Ren, B., Russell, S., Cherbas, P., Graveley, B.R., Lewis, S., Micklem, G., Oliver, B., Park, P.J., Celniker, S.E., Henikoff, S., Karpen, G.H., Lai, E.C., MacAlpine, D.M., Stein, L.D., White, K.P., Kellis, M., Consortium, m., 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787-1797.

Ruiz Gómez, M., Bate, M., 1997. Segregation of myogenic lineages in Drosophila requires numb. Development 124, 4857-4866.

Ruiz-Gómez, M., 1998. Muscle patterning and specification in Drosophila. Int J Dev Biol 42, 283-290.

Salmon-Divon, M., Dvinge, H., Tammoja, K., Bertone, P., 2010. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics 11, 415.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., Furlong, E.E., 2007. A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev 21, 436-449.

Sandmann, T., Jakobsen, J.S., Furlong, E.E., 2006a. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. Nat Protoc 1, 2839-2855.

Sandmann, T., Jensen, L.J., Jakobsen, J.S., Karzynski, M.M., Eichenlaub, M.P., Bork, P., Furlong, E.E., 2006b. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. Dev Cell 10, 797-807.

Schaub, C., Nagaso, H., Jin, H., Frasch, M., 2012. Org-1, the Drosophila ortholog of Tbx1, is a direct activator of known identity genes during muscle specification. Development 139, 1001-1012.

Schuh, R., Aicher, W., Gaul, U., Côté, S., Preiss, A., Maier, D., Seifert, E., Nauber, U., Schröder, C., Kemler, R., 1986. A conserved family of nuclear proteins containing structural elements of the finger protein encoded by Krüppel, a Drosophila segmentation gene. Cell 47, 1025-1032.

Senger, K., Armstrong, G.W., Rowell, W.J., Kwan, J.M., Markstein, M., Levine, M., 2004. Immunity regulatory DNAs share common organizational features in Drosophila. Mol Cell 13, 19-32.

Seyres, D., Röder, L., Perrin, L., 2012. Genes and networks regulating cardiac development and function in flies: genetic and functional genomic approaches. Brief Funct Genomics 11, 366-374.

Sigvardsson, M., Clark, D.R., Fitzsimmons, D., Doyle, M., Akerblad, P., Breslin, T., Bilke, S., Li, R., Yeamans, C., Zhang, G., Hagman, J., 2002. Early B-cell factor, E2A, and Pax-5 cooperate to activate the early B cell-specific mb-1 promoter. Mol Cell Biol 22, 8539-8551.

Sigvardsson, M., O'Riordan, M., Grosschedl, R., 1997. EBF and E47 collaborate to induce expression of the endogenous immunoglobulin surrogate light chain genes. Immunity 7, 25-36.

Simionato, E., Ledent, V., Richards, G., Thomas-Chollier, M., Kerner, P., Coornaert, D., Degnan, B.M., Vervoort, M., 2007. Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. BMC Evol Biol 7, 33.

Smith, E.M., Gisler, R., Sigvardsson, M., 2002. Cloning and characterization of a promoter flanking the early B cell factor (EBF) gene indicates roles for E-proteins and autoregulation in the control of EBF expression. J Immunol 169, 261-270.

Sorge, S., Ha, N., Polychronidou, M., Friedrich, J., Bezdan, D., Kaspar, P., Schaefer, M.H., Ossowski, S., Henz, S.R., Mundorf, J., Rätzer, J., Papagiannouli, F., Lohmann, I., 2012. The cis-regulatory code of Hox function in Drosophila. EMBO J 31, 3323-3333.

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Kellis, M., Furlong, E.E., Birney, E., 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol 13, R49.

Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., Ruby, J.G., Brennecke, J., Hodges, E., Hinrichs, A.S., Caspi, A., Paten, B., Park, S.W., Han, M.V., Maeder, M.L., Polansky, B.J., Robson, B.E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D.G., Eastman, D.A., Rice, M., Weir, M., Hahn, M.W., Park, Y., Dewey, C.N., Pachter, L., Kent, W.J., Haussler, D., Lai, E.C., Bartel, D.P., Hannon, G.J., Kaufman, T.C., Eisen, M.B., Clark, A.G., Smith, D., Celniker, S.E., Gelbart, W.M., Kellis, M., 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450, 219-232.

Stolfi, A., Gainous, T.B., Young, J.J., Mori, A., Levine, M., Christiaen, L., 2010. Early chordate origins of the vertebrate second heart field. Science 329, 565-568.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., van Helden, J., 2012. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40, e31.

Tixier, V., Bataillé, L., Jagla, K., 2010. Diversification of muscle types: recent insights from Drosophila. Exp Cell Res 316, 3019-3027.

Tootle, T.L., Silver, S.J., Davies, E.L., Newman, V., Latek, R.R., Mills, I.A., Selengut, J.D., Parlikar, B.E., Rebay, I., 2003. The transcription factor Eyes absent is a protein tyrosine phosphatase. Nature 426, 299-302.

Travis, A., Hagman, J., Hwang, L., Grosschedl, R., 1993. Purification of early-B-cell factor and characterization of its DNA-binding specificity. Mol Cell Biol 13, 3392-3400.

Treiber, N., Treiber, T., Zocher, G., Grosschedl, R., 2010a. Structure of an Ebf1:DNA complex reveals unusual DNA recognition and structural homology with Rel proteins. Genes Dev 24, 2270-2275.

Treiber, T., Mandel, E.M., Pott, S., Györy, I., Firner, S., Liu, E.T., Grosschedl, R., 2010b. Early B cell factor 1 regulates B cell gene networks by activation, repression, and transcription-independent poising of chromatin. Immunity 32, 714-725.

Tsai, R.Y., Reed, R.R., 1997. Cloning and functional characterization of Roaz, a zinc finger protein that interacts with O/E-1 to regulate gene expression: implications for olfactory neuronal development. J Neurosci 17, 4159-4169.

Tsai, R.Y., Reed, R.R., 1998. Identification of DNA recognition sequences and protein interaction domains of the multiple-Zn-finger protein Roaz. Mol Cell Biol 18, 6447-6456.

van Bemmel, J.G., Filion, G.J., Rosado, A., Talhout, W., de Haas, M., van Welsem, T., van Leeuwen, F., van Steensel, B., 2013. A network model of the molecular organization of chromatin in Drosophila. Mol Cell 49, 759-771.

Venken, K.J., Carlson, J.W., Schulze, K.L., Pan, H., He, Y., Spokony, R., Wan, K.H., Koriabine, M., de Jong, P.J., White, K.P., Bellen, H.J., Hoskins, R.A., 2009. Versatile P[acman] BAC libraries for transgenesis studies in Drosophila melanogaster. Nat Methods 6, 431-434.

Vervoort, M., Crozatier, M., Valle, D., Vincent, A., 1999. The COE transcription factor Collier is a mediator of short-range Hedgehog-induced patterning of the Drosophila wing. Curr Biol 9, 632-639.

Visel, A., Rubin, E.M., Pennacchio, L.A., 2009. Genomic views of distant-acting enhancers. Nature 461, 199-205.

Wang, M.M., Reed, R.R., 1993. Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. Nature 364, 121-126.

Wang, S.S., Betz, A.G., Reed, R.R., 2002. Cloning of a novel Olf-1/EBF-like gene, O/E-4, by degenerate oligo-based direct selection. Mol Cell Neurosci 20, 404-414.

Wang, S.S., Lewcock, J.W., Feinstein, P., Mombaerts, P., Reed, R.R., 2004. Genetic disruptions of O/E2 and O/E3 genes reveal involvement in olfactory receptor neuron projection. Development 131, 1377-1388.

Wang, S.S., Tsai, R.Y., Reed, R.R., 1997. The characterization of the Olf-1/EBF-like HLH transcription factor family: implications in olfactory gene regulation and neuronal development. J Neurosci 17, 4149-4158.

Zandi, S., Mansson, R., Tsapogas, P., Zetterblad, J., Bryder, D., Sigvardsson, M., 2008. EBF1 is essential for B-lineage priming and establishment of a transcription factor network in common lymphoid progenitors. J Immunol 181, 3364-3372.

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

BIBLIOGRAPHIE des FIGURES

Artero, R., Furlong, E.E., Beckett, K., Scott, M.P., Baylies, M., 2003. Notch and Ras signaling pathway effector genes expressed in fusion competent and founder cells during Drosophila myogenesis. Development 130, 6257-6272.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., Furlong, E.E., 2012. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet 44, 148-156.

Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S., Vincent, A., 1999. Head versus trunk patterning in the Drosophila embryo; collier requirement for formation of the intercalary segment. Development 126, 4385-4394.

Dubois, L., Enriquez, J., Daburon, V., Crozet, F., Lebreton, G., Crozatier, M., Vincent, A., 2007. Collier transcription in a single Drosophila muscle lineage: the combinatorial control of muscle identity. Development 134, 4347-4355.

Enriquez, J., Boukhatmi, H., Dubois, L., Philippakis, A.A., Bulyk, M.L., Michelson, A.M., Crozatier, M., Vincent, A., 2010. Multi-step control of muscle diversity by Hox proteins in the Drosophila embryo. Development 137, 457-466.

Fraga, D., Meulia, T. and Fenster, S., 2008. Real-Time PCR. Current Protocols - Essential Laboratory Techniques. Unit 10.3, 10.13.11–10.13.34.

Kramer, S.G., Kidd, T., Simpson, J.H., Goodman, C.S., 2001. Switching repulsion to attraction: changing responses to slit during transition in mesoderm migration. Science 292, 737-740.

Liu, Y.H., Jakobsen, J.S., Valentin, G., Amarantos, I., Gilmour, D.T., Furlong, E.E., 2009. A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. Dev Cell 16, 280-291.

Mohler, J., Mahaffey, J.W., Deutsch, E., Vani, K., 1995. Control of Drosophila head segment identity by the bZIP homeotic gene cnc. Development 121, 237-247.

Ntini, E., Wimmer, E.A., 2011. Second order regulator Collier directly controls intercalary-specific segment polarity gene expression. Dev Biol 360, 403-414.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F., Washietl, S., Arshinoff, B.I., Ay, F., Meyer, P.E., Robine, N., Washington, N.L., Di Stefano, L., Berezikov, E., Brown, C.D., Candeias, R., Carlson, J.W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M.Y., Will, S., Alekseyenko, A.A., Artieri, C., Booth, B.W., Brooks, A.N., Dai, Q., Davis, C.A., Duff, M.O., Feng, X., Gorchakov, A.A., Gu, T., Henikoff, J.G., Kapranov, P., Li, R., MacAlpine, H.K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S.K., Riddle, N.C., Sakai, A., Samsonova, A., Sandler, J.E., Schwartz, Y.B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K.H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S.E., Brent, M.R., Cherbas, L., Elgin, S.C., Gingeras, T.R., Grossman, R., Hoskins, R.A., Kaufman, T.C., Kent, W., Kuroda, M.I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J.W., Ren, B., Russell, S., Cherbas, P., Graveley, B.R., Lewis, S., Micklem, G., Oliver, B., Park, P.J., Celniker, S.E., Henikoff, S., Karpen, G.H., Lai, E.C., MacAlpine, D.M., Stein, L.D., White,

K.P., Kellis, M., Consortium, m., 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787-1797.

Travis, A., Hagman, J., Hwang, L., Grosschedl, R., 1993. Purification of early-B-cell factor and characterization of its DNA-binding specificity. Mol Cell Biol 13, 3392-3400.

Treiber, T., Mandel, E.M., Pott, S., Györy, I., Firner, S., Liu, E.T., Grosschedl, R., 2010. Early B cell factor 1 regulates B cell gene networks by activation, repression, and transcription-independent poising of chromatin. Immunity 32, 714-725.

Venken, K.J., Carlson, J.W., Schulze, K.L., Pan, H., He, Y., Spokony, R., Wan, K.H., Koriabine, M., de Jong, P.J., White, K.P., Bellen, H.J., Hoskins, R.A., 2009. Versatile P[acman] BAC libraries for transgenesis studies in Drosophila melanogaster. Nat Methods 6, 431-434.

VI – Annexes

Annexe 1 : Tableau récapitulatif des données d'expression des 30 candidats (hybridations *in situ* et analyse de leurs CRMs via les lignées J. Farm)

Annexe 2 : Tableau récapitulatif des CRM candidats étudiés

Annexe 3 : Comparaison entre les résultats des ChIP Col ModEncode (ref.618) et de notre laboratoire

Annexe 4 : Profil d'expression des CRM ama[3.25] et cg6234[7.45], avant et après mutation du site prédit de fixation de Col

Annexe 5 : Profil d'expression des CRM eya[3.91] et phyl[22.41], avant et après mutation du site prédit de fixation de Col

Annexe 6 : Profil d'expression du CRM jing[4.69] avant et après mutation du site prédit de fixation de Col

Annexe 7 : Profil d'expression du CRM tkv[8.27] avant et après mutation du site prédit de fixation de Col

Candidat	hauteur du pic (enrichissement)	Profil hybridations *in situ* (domaines Col positifs)	CRM-gal4>UAS-lacZ (lignées J. Farm)
ama	3,25	Tête - muscle DA3 (st.13-15) -glande lymphatique	-
aret	4,76 - 4,09	Muscle DA3 (progéniteur>st.15)	-
cg12484	3,4	Muscle DA3 (st13-14) mais patron confus	**GMR92A02 :** quelques neurones du SNC
cg2022	8,68 - 5,68 - 2,86	Muscle DA3 (st. 13-14) mais signal très faible	-
cg34371	4,91 - 3,03	Muscle DA3 (st. 15) mais signal faible	-
cg4115	3,38	Aucun signal	-
cg4161	7,11 - 3,42	Muscle DA3 (st13-15)	-
cg6234	7,45	Muscle DA3 (st. 13-14) – neurones MD inférieurs	-
cnc	12,62	tête	-
Dys	3,73	Muscle DA3 (st. 13-15)	-
eya	3,91	Muscle DA3 (progéniteur) - SNC	**GMR21C02 :** muscle DA3 (et DT1 ?) (st.13-15) et SNC (quelques neurones Col+) - **GMR21A11 : SNC** (quelques neurones Col+)
jing	8,57 - 4,69 - 4,56 - 3,33 - 2,52	Muscle DA3 (progéniteur) mais signal très faible	-
kuz	2,92	Aucun signal	-
luna	4,37 - 2,65	ubiquitaire	-
mbl	6,41 - 3,1	Muscle DA3 (st. 13-15) et neurones MD	-
Mrtf	3,35	Muscle DA3 (st. 13-15)	**GMR56G06** : plusieurs muscles dont le DA3 (st. 13-15)
nerfin-1	5,31	Tête - muscle DA3 (faible, st.13-14) - SNC	-
numb	4,62	Aucun signal	-
phyl	22,41 - 10,54 - 4,43 - 3,34	Signal très faible	**GMR52C04** : neurones MD sup., SNC - **GMR74A01** : SNC - **GMR51C08** : muscle DA3 mais peu pénétrant, SNC
pum	6,27 - 5,92 - 4,83 - 4,72	Neurones MD et quelques cellules du SNC	-
px	9,71 - 2,97 - 2,57	Aucun signal	-
salr	4,87	Tête – muscle DA3 (st.14) mais signal très faible	-
sens-2	3,71 (-6,31)	Neurones MD	-
sli	11,3 - 5,04 - 4,1 - 3,98	SNC	**GMR32A06 :** muscle DT1 probablement et SNC (neurones Col-)
smr	13,6	Muscle DA3 st. 13-15 mais patron confus	-
so	5,75	Muscle DA3 (progéniteur) – quelques cellules du SNC	**GMR15C09 :** muscle DA3 (st.13-15) et d'autres cellules du mésoderme
ten-m	2,79	Signal très faible	-
tkv	8,27	Signal très faible : tête – muscle DA3 (st.13-14)	-
tl	8,36	Signal très faiblr	-
unc-5	6,21 - 4,93	SNC mais patron confus	**GMR93E11, GMR93F01, GMR93F02, GMR93E10** : quelques neurones Col+ du SNC. **GMR93F01** : faible expression dans le muscle DA3

Annexe 1. Tableau récapitulatif des données d'expression des 30 candidats (hybridations *in situ* et analyse de leurs CRMs via les lignées J. Farm)

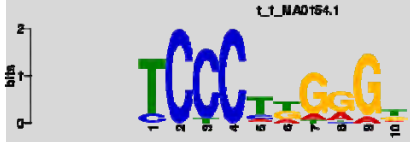
Candidat	coordonnées CRM	taille CRM (pb)	Site Col	Site Col muté
ama[3,25]	chr3R:2587010-2587909	899	GTCCT**CCC**AT**GGG**AACGGCCT	GTCC**TCC**AT**CCC**AACGGCCT
aret[4,76]	chr2L:12251861-12252820	959	GATGC**CT**CCG**GGG**AGCTGTGA	GATGC**CT**CCG**CCC**AGCTGTGA
cg2022[2,86]	chr3R:815209-816233	1024	TTGGC**CCTTT**G**GG**ACAGAGAA	TTGGC**CCTTT**C**CC**ACAGAGAA
cg4115[3,38]	Chr3R:8166523-8167422	899	TGATT**CCC**AG**GGG**AAAAAGGT	TGATT**CCC**AG**CCC**AAAAAGGT
cg6234[7,45]	chr3R:8474739-8475647	908	GCGTT**CCC**AA**GGG**ATTTGGAA	GCGTT**GGG**AA**GGG**ATTTGGAA
cnc[12,62]	chr3R:19037100-19038139	1039	CATTTGGT**CCCT**G**GAG**AATTG	CATTTGGT**GGG**TG**GAG**AATTG
eya[3,91]	Chr2L:6540104-6541545	1441	**CCCTGGGGT**…(379pb)…**TCCCTGGGG**	**CCCTG**CCC**T**…(379pb)…**TCCCTG**CCC
jing[4,69]	chr2R:2482099-2483002	903	AGTG**CCCG**CT**GGG**AATTTTCT	AGTG**CCCG**CT**CCC**AATTTTCT
[L]CRM Col (Dubois *et al.* 2007)	chr2R:10688751-10687377	1374	AT**GTC**TG**GGG**AC	AT**GTC**TG**CCC**AC
Mrtf[3,35]	chr3L:2744714-2745691	977	TTTCC**CCCCT**G**GG**AATTCATG	TTTCC**CCCCT**C**CC**AATTCATG
nerfin-1[5,31]	Chr3L:907323-908228	905	TT**CCCCT**G**GC**AA..(22pb)..ATC**GCCTTGGGA**	TT**GGG**CT**GGC**AA..(22pb)..ATC**GCCTT**C**CC**A
oaz (/phyl)[22,41]	chr2R:10324868-10325829	961	CAACT**CCCT**G**GGG**AGTTTTCA	CAACT**CCCT**G**CCC**AGTTTTCA
oaz (/phyl)[4,43]	chr2R:10339692-10340791	1099	AAAAT**CCCAG**G**GG**AACCCTTT	AAAAT**CCCAG**C**CC**AACCCTTT
sli[3,98]	chr2R:11807800-11808717	917	AGTGT**CCCCC**G**AG**ATTAGAGC	AGTGT**GGG**CC**GAG**ATTAGAGC
so[5,75]	chr2R:3320280 -3321219	939	GAATG**CCCCG**G**GG**AGTGCATG	GAATG**CCCCG**C**CC**AGTGCATG
tkv[8,27]	chr2L:5245055-5246039	984	TCGAT**CCCCA**G**GG**AATTGGGT	TCGAT**GGG**CA**GGG**AATTGGGT

Annexe 2. Tableau récapitulatif des CRM candidats étudiés

Les coordonnées des CRM (2e colonne) correspondent à la *release* R5 du génome de D. *melanogaster*. Les sites Col tels que positionnés par la recherche *de novo* de MEME sont en caractères gras dans la 4e colonne, et la mutation effectuée sur ces motifs est en rouge dans la 5e colonne.

Recherche de motifs *de novo* - analyse avec la suite MEME sur les fragments issus des ChIP Col

A

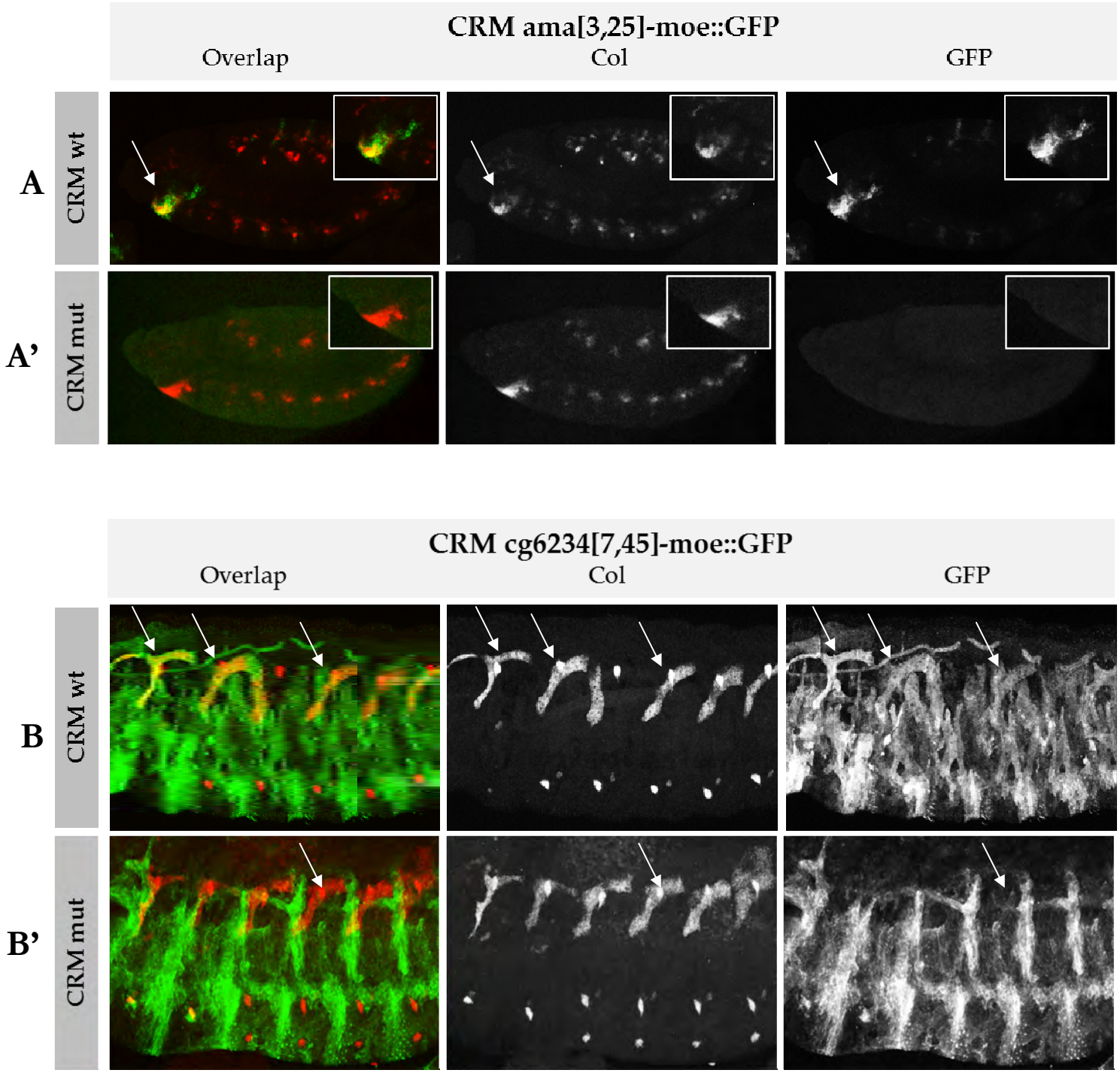
ModEncode (ref. 618)	Ce rapport
Echantillon analysé	
785 fragments – 826 pb en moy.	413 fragments – 1000 pb
Meilleur motif enrichi (MEME)	
	
177 motifs (23%) – e-value = 1.3 e-211	324 motifs (78%) – e-value = 2.6 e-107
Meilleure matrice associée (TOMTOM)	
	
Klu – p-value = 2.8 e-04	EBF – p-value = 1.89 e-06

B

Intersection entre les données ChIP Col issues de ModEncode (ref. 618) et de notre laboratoire		
Coordonnées des pics ModEncode	**Coordonnées des pics du laboratoire**	**Gène candidat associé**
chr2R : 18398133 - 18399277	chr2R : 18396772 - 18398478	px
chr3L : 5353146 - 5354039	chr3L : 5353229 - 5354317	cg4769
chr3R : 939086 - 939749	chr3R : 938900 - 940991	cg43131
chr3R : 22623271 - 22623935	chr3R : 22623076 - 22624179	Tl
chrX : 12629699 - 12630878	chrX : 12629543 - 12630848	smr
chrX : 17692804 - 17693541	chrX : 17692628 - 17694369	odsH
chrX : 8631545 - 8632253	chrX : 8630183 - 8631732	l(1)G0020

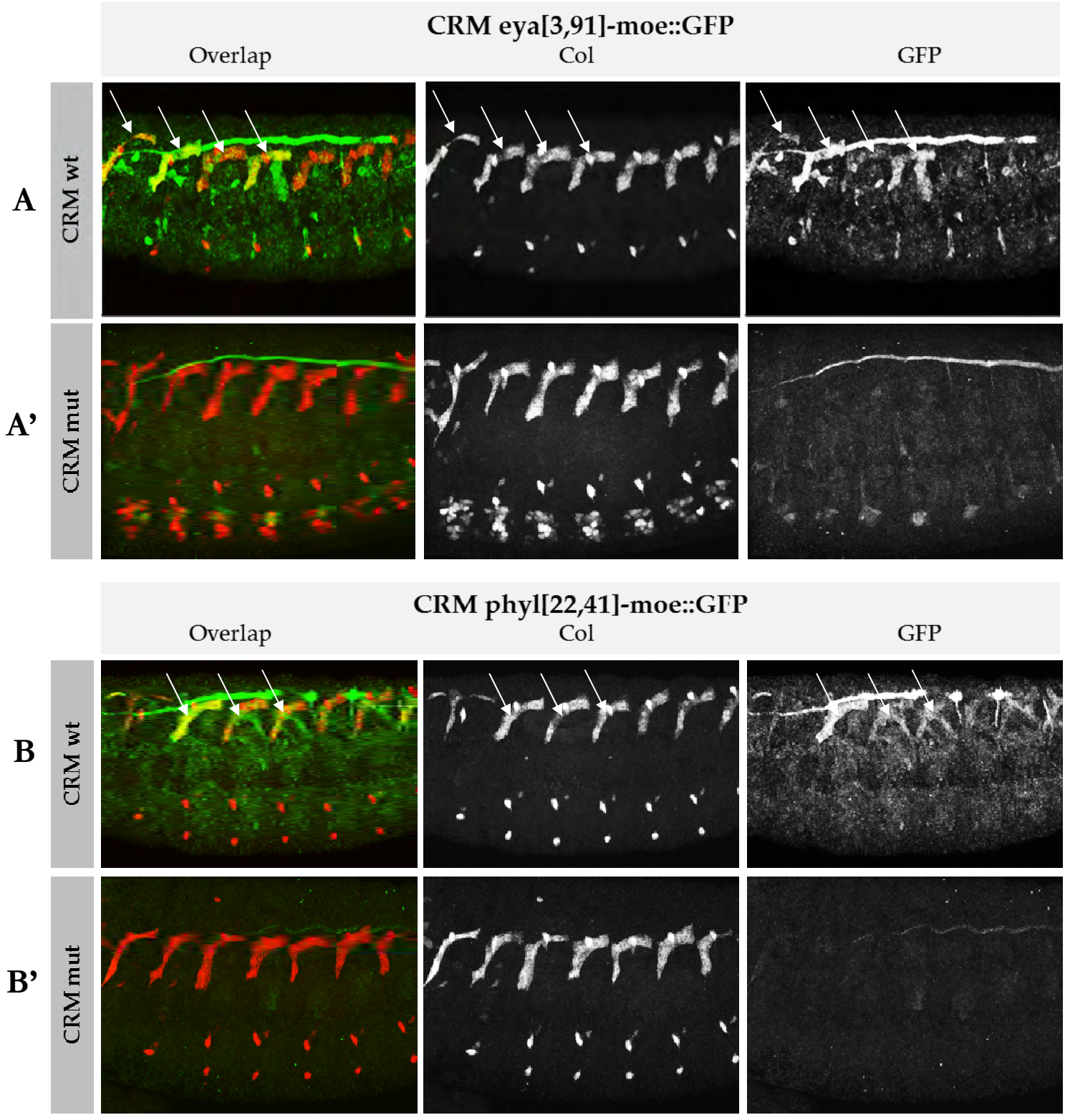
Annexe 3. Comparaison entre les résultats des ChIP Col ModEncode (ref.618) et de notre laboratoire

A : Recherche *de novo* avec la suite MEME de motifs enrichis sur les fragments issus des 2 expériences de ChIP Col (nos données, et ModEncode (Roy et al., 2010)). **B** : L'intersection entre les 785 fragments de ModEncode et les 413 fragments identifiés dans notre étude ne comprend que les 7 régions décrites dans ce tableau.



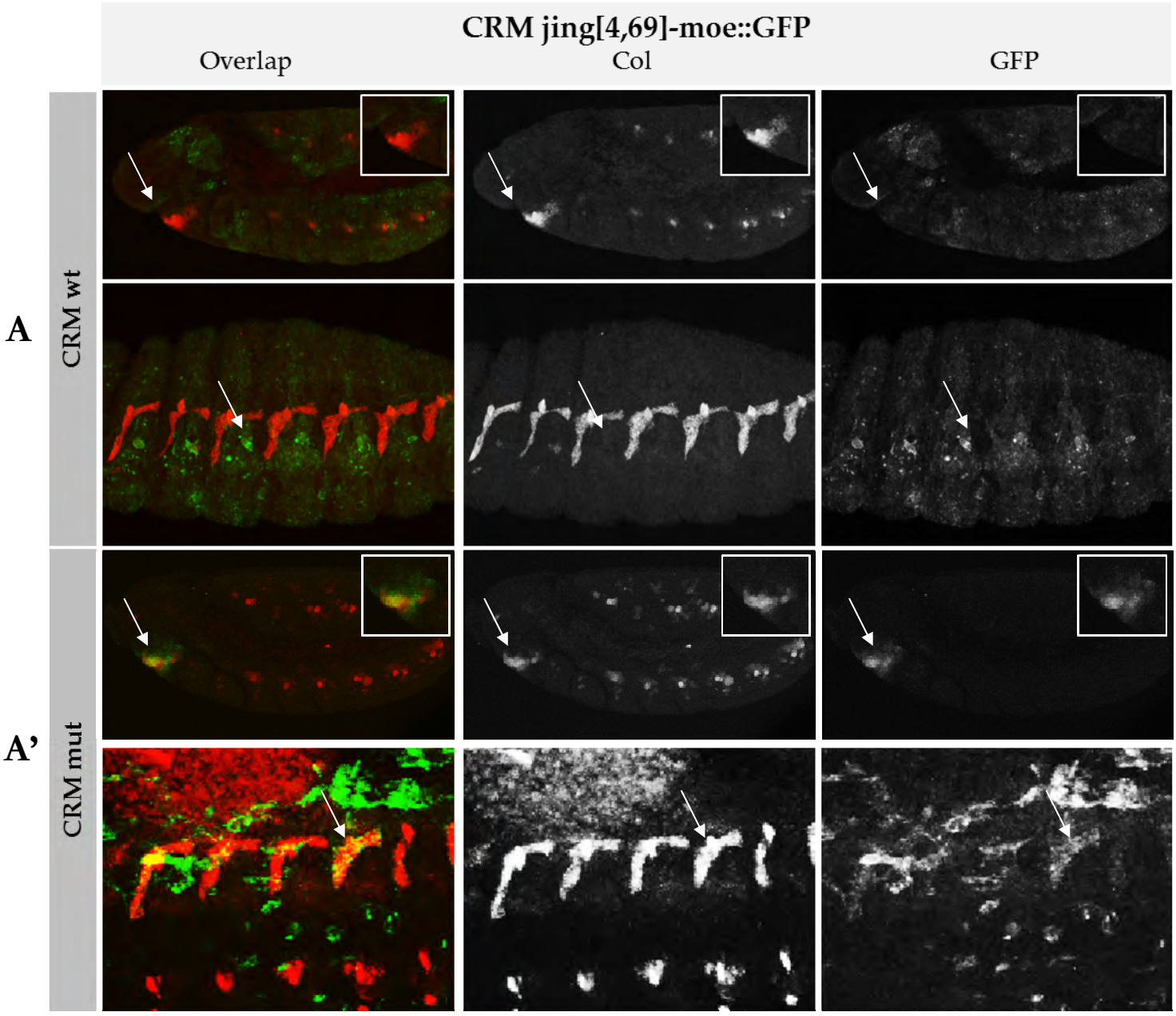
Annexe 4. Profil d'expression des CRM ama[3.25] et cg6234[7.45], avant et après mutation du site prédit de fixation de Col

A : Double immunocoloration Col/GFP d'un embryon CRM ama[3.25]-moeGFP au stade 12. L'expression recouvrant celle de Col dans la tête disparait après mutation du site Col (**A'**). **B** : Double immunocoloration Col/GFP d'un embryon CRM cg6234[7.45]-moeGFP au stade 15. L'expression observée dans les muscles DA3 n'est pas retrouvée lorsqu'on mute le site Col (**B'**).



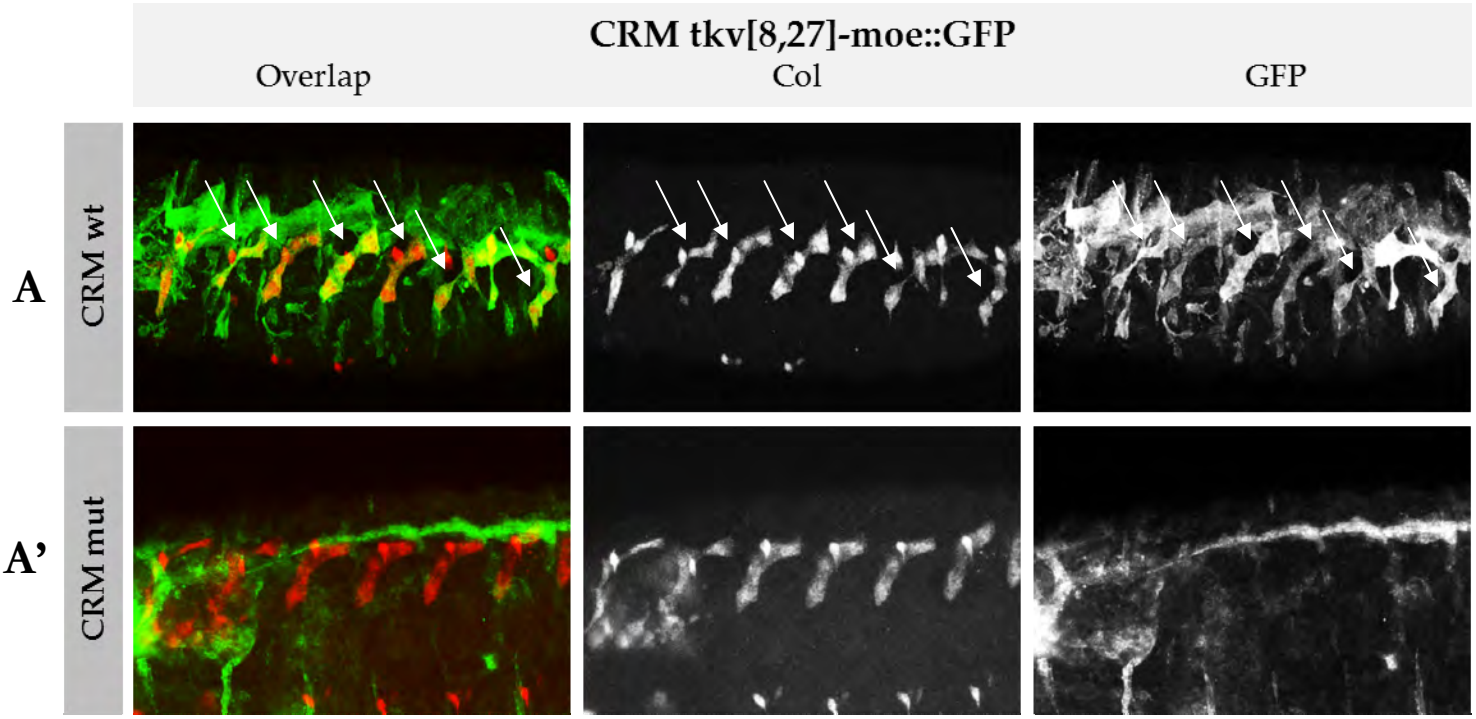
Annexe 5. Profil d'expression des CRM eya[3.91] et phyl[22.41], avant et après mutation du site prédit de fixation de Col

A : Double immunocoloration Col/GFP d'un embryon CRM eya[3.91]-moeGFP au stade 15. L'expression observée dans le muscle DA3 (marqué en rouge par l'anticorps anti-Col) disparait après mutation du site Col (**A'**). **B** : Double immunocoloration Col/GFP d'un embryon CRM phyl[22.41]-moeGFP au stade 15. L'expression observée dans le muscle DA3 (en rouge) et dans d'autres muscles DL est perdue lorsqu'on mute le site Col du CRM (**B'**).



Annexe 6. Profil d'expression du CRM jing[4.69] avant et après mutation du site prédit de fixation de Col

A : Double immunocoloration Col/GFP d'un embryon CRM jing[4.69]-moeGFP aux stades 12 (ligne supérieure) et 15 (ligne inférieure). Alors qu'on n'observe pas d'expression de la GFP ni dans la tête (st. 12), ni dans le muscle DA3 (st. 15), le CRM jing[4.69]-moeGFP muté (**A'**) est actif dans la tête au niveau du domaine Col au stade 12 et de manière stochastique dans le muscle DA3 au stade 15. Au contraire, l'expression du CRM jing[4.69]-moeGFP observée au stade 15 dans d'autres muscles DL (flèche) est perdue lorsqu'on mute le site Col.



Annexe 7. Profil d'expression du CRM tkv[8.27] avant et après mutation du site prédit de fixation de Col

A : Double immunocoloration Col/GFP d'un embryon CRM tkv[8.27]-moeGFP au stade 15. L'expression observée dans le muscle DA3 (marqué en rouge par l'anticorps anti-Col) disparait après mutation du site Col (**A'**), de même que l'expression observée dans le muscle DA2, plus dorsal.

Summary

The COE (**Co**llier/**E**arly B cell Factor) family is a metazoan-specific family of transcription factors (TF) that are involved in the control of numerous biological processes, including hematopoiesis, neurogenesis and muscle identity. Mutant analysis of COE TFs across several organisms showed defects in the specification of different cell types, like neuron subtypes or, in mammalians, B lymphocytes and brown adipocytes. However, the COE target genes are mostly unknown. *Drosophila* (fruit fly) is an excellent model to study the functional diversity of COE TFs. Collier (Col), the only COE member in this insect, controls several processes during embryogenesis: intercalary segment formation in the head, specification of somatic muscle identity, of subtypes of neurons both in the central and peripheral nervous system, and specification of the larval hematopoietic "niche". A gene candidate approach identified a few Col target genes, which appear specific and different in each of these tissues, but the molecular basis of this specificity remains unknown. The *Drosophila* embryonic musculature is composed of 30 different muscles per trunk hemisegment, each muscle constituted by a single multinucleate fiber. It is now well established that the combination of TFs expressed in the founder myoblast of a muscle controls the identity, i.e. morphological and functional properties which are characteristic of this muscle. Col acts as an identity TF in dorso-lateral muscles, in particular the DA3 muscle. My PhD project was to set up a genome wide approach to identify direct Col target genes in the mesoderm which control DA3 muscle morphology.

I first studied *col* transcriptional regulation in the DA3 muscle lineage by characterizing in more details the 2 Cis-Regulatory Modules (CRM) controlling *col* expression in the mesoderm, respectively during the specification and realization phases of muscle identity. The dissection of the Early CRM (^{E}CRM), initially identified by bio-informatics, allowed me to define a shorter fragment solely responsible for Col expression in the promuscular cluster. The late CRM (^{L}CRM) was previously characterized in the lab. I tested the function of *in silico* predicted and/or *in vivo* bound motifs for the mesodermal TF Twist, and homeotic Hox proteins, by point mutations of these motifs in a reporter gene assay. Unfortunately, this assay did not give us better insight into the direct regulatory control of *col* transcription via the ^{L}CRM. I therefore started developing a novel strategy to analyze CRMs in their genomic context, based on BAC recombineering, a strategy that will also serve to identify Col direct target genes in a tissue-specific way. In parallel, I contributed analyzing the combinatorial control of DA3 muscle identity by Col and Nautilus (MyoD), showing that each TF regulates different properties of this muscle.

The core of my PhD work was the identification of Collier direct target genes in the DA3 muscle lineage, and the characterization of the corresponding CRM to better understand how COE proteins activate specific target genes in a tissue-dependent manner. I performed chromatin immuno-precipitation on whole embryos followed by systematic sequencing of the immuno-precipitated fragments (ChIPseq). By bio-informatics, I identified Col *in vivo* binding motif and showed that Col binding *in vivo* is context-dependent. Several candidate genes were validated by *in situ* hybridizations and functional analysis of the Col binding CRM. TF are over-represented among these targets. All together, the results reveal an unexpected complexity of gene regulatory networks that control muscle identity in *Drosophila* and confirm the critical role for Col in several transcription regulatory networks in the embryo. Considering the evolutionary conservation of COE proteins and their *in vivo* DNA binding properties, these results bring new insight into the complexity of COE function in other organisms, including mammals.

Key words : Drosophila, Collier/EBF, DA3 muscle, ChIPseq, CRM, transcriptional regulatory networks

AUTEUR : Mathilde de TAFFIN de TILQUES

TITRE : Contrôle transcriptionnel de l'identité musculaire chez la drosophile : Modules Cis-Régulateurs et gènes cibles directs de Collier

DIRECTEUR DE THESE : Alain VINCENT

LIEU ET DATE DE SOUTENANCE : à TOULOUSE le 29 Octobre 2013

RÉSUMÉ en français

Les facteurs de transcription (FT) métazoaires de la famille COE (Collier/Early B Cell Factor) participent au contrôle de divers processus biologiques : hématopoïèse, neurogenèse, établissement du patron des muscles. L'analyse de mutants chez divers organismes modèles montre des défauts de spécification de différents types cellulaires, dont des sous-types de neurones et, chez les mammifères, les lymphocytes B et les adipocytes bruns. Cependant les gènes cibles régulés par les facteurs COE restent majoritairement inconnus. La drosophile est un excellent modèle pour étudier la diversité fonctionnelle des FT COE. Collier (Col), le seul FT COE chez cet insecte, contrôle plusieurs processus au cours de l'embryogenèse : formation du segment céphalique intercalaire, spécification de l'identité de muscles squelettiques, de sous-types neuronaux du système nerveux central et périphérique, et de la « niche » dans l'organe hématopoïétique larvaire. Une approche gène-candidat a montré que Col régule des gènes spécifiques et différents dans chacun de ces tissus, mais les bases moléculaires de cette spécificité restaient inconnues. La musculature de l'embryon de drosophile est formée d'environ 30 muscles squelettiques différents dans chaque hemisegment, constitués chacun d'une seule fibre multi-nucléée. Il est maintenant bien établi que la combinatoire de facteurs de transcription (FT) exprimée dans les myoblastes fondateur des muscles contrôle l'identité musculaire, c'est-à-dire les caractéristiques morphologiques et fonctionnelles propres à chacun des muscles. Col est un FT identitaire, requis en particulier pour l'identité des muscles dorso-latéraux, dont le muscle DA3. Mon projet de thèse était de mettre en œuvre une recherche des gènes cibles à l'échelle génomique afin d'identifier les gènes régulés par Col au cours de la myogenèse embryonnaire et impliqués dans l'identité morphologique du muscle DA3.

Au cours de ma thèse, j'ai d'abord étudié à la régulation transcriptionnelle de *col* dans le lignage DA3, et caractérisé plus en détail les 2 Modules Cis-Régulateurs (CRM) contrôlant l'expression mésodermique de Col, respectivement aux étapes de spécification et de réalisation de l'identité musculaire. La dissection du CRM précoce (^{E}CRM) identifié par bio-informatique m'a permis d'identifier un fragment responsable uniquement de l'expression promusculaire de *col*. Le CRM tardif (^{L}CRM) avait été préalablement identifié. J'ai entrepris une étude de la fonction de sites de fixation prédits *in silico* et/ou *in vivo* pour les FT mésodermiques Twist et homéotiques Hox, par mutagenèse ponctuelle suivie de la construction de gènes rapporteurs, mais cette analyse n'a pas permis de conclusion définitive. J'ai donc commencé de développer une stratégie d'analyse des CRM de *col* dans leur contexte génomique, une stratégie qui permettra par ailleurs d'identifier les gènes cibles directs de Col selon les tissus où il est exprimé. En parallèle, j'ai contribué à l'étude du contrôle combinatoire de l'identité musculaire par Col et le FT mésodermique Nautilus (MyoD), qui montre que chaque FT contrôle des propriétés différentes du muscle DA3.

Le cœur de mes travaux de thèse a consisté à identifier les gènes-cibles directs de Collier, en particulier au cours de la myogenèse, et à caractériser les modules cis-régulateurs associés afin de comprendre les bases contextuelles de la régulation transcriptionnelle tissu-spécifique par les protéines COE. Pour cela, j'ai réalisé des expériences d'immunoprécipitation de la chromatine à partir d'embryons entiers, suivi du séquençage systématique des fragments d'ADN précipités (ChIPseq) et de leur analyse par bio-informatique. Cette analyse m'a permis d'identifier le motif ADN sélectivement reconnu par Col *in vivo*, et de montrer que cette reconnaissance est contextuelle. Plusieurs gènes cibles ont été validés par des expériences d'hybridation *in situ* et d'analyse fonctionnelle de leurs CRM, parmi lesquels une majorité d'autres FTs. L'ensemble des résultats révèle une complexité inattendue des réseaux de régulation transcriptionnelle contrôlant l'identité musculaire chez la drosophile et confirme que Col est un acteur majeur de différents réseaux dans différents tissus embryonnaires. Au vu de la conservation des FTs COE au cours de l'évolution, les conclusions de cette étude modèle chez la drosophile apportent un éclairage nouveau sur les études en cours sur des modèles mammifères.

MOTS-CLES : Drosophile, Collier/EBF, muscle DA3, ChIPseq, CRM, Réseaux de régulation transcriptionnelle

DISCIPLINE ADMINISTRATIVE : Biologie du Développement

INTITULE ET ADRESSE DU LABORATOIRE : Centre de Biologie du Développement – UMR 5547
Université Paul Sabatier (Toulouse III)
118, route de Narbonne - Bât. 4R3b3 - 31062 Toulouse cedex 9 - FRANCE.