



HAL
open science

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

Elodie Darbo

► **To cite this version:**

Elodie Darbo. Découverte d'éléments cis-régulateurs impliqués dans l'activation transcriptionnelle du génome zygotique dans l'embryon précoce de *Drosophila melanogaster*. Bio-Informatique, Biologie Systémique [q-bio.QM]. Université de la Méditerranée - Aix-Marseille II, 2011. Français. NNT : . tel-00644865v1

HAL Id: tel-00644865

<https://theses.hal.science/tel-00644865v1>

Submitted on 25 Nov 2011 (v1), last revised 14 Dec 2011 (v2)

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

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

Université de la Méditerranée
Faculté des Sciences de Luminy
Ecole Doctorale des Sciences de la Vie et de l'Environnement



Thèse
présentée en vue de l'obtention du grade de

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

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

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

Date Prévue le vendredi 16 Décembre

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

*"Au som deu malh, que i a ua lutz,
Qu'i cau guardar los uelhs dessùs,
Que'ns cau traucar tot lo segàs,
Tà ns'arrapar, sonque las mans,
Lhèu veiram pas jamei la fin,
La libertat qu'ei lo camin,
Après lo malh, un aute malh,
Après la lutz, ua auta lutz ..."*

D'après Los de Nadau

Résumé

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

Table des matières

Table des figures	v
Liste des tableaux	ix
Acronymes et Abréviations	xi
1 Introduction	1
1.1 La régulation transcriptionnelle	2
1.1.1 Mise en place de la machinerie basale de transcription	2
1.1.2 Les facteurs de transcription	2
1.1.2.1 Définition	2
1.1.2.2 Caractérisation des sites de liaison protéine-ADN	3
1.1.2.3 Représentations de la spécificité de liaison des facteurs trans- criptionnels : les motifs de liaison	6
1.1.2.4 Découverte de motifs	7
1.1.2.5 Prédiction de sites	12
1.1.2.6 Prédiction de modules cis-régulateurs	14
1.1.2.7 Bases de données	17
1.1.3 Les co-régulateurs	18
1.1.3.1 Le complexe médiateur	18
1.1.3.2 Les régulateurs de la chromatine	18
1.2 L'activation du génome zygotique (AGZ) durant l'embryogenèse de la drosophile	20
1.2.1 L'embryogénèse précoce chez les métazoaires : mécanismes communs	21
1.2.1.1 Activation de l'œuf	21
1.2.1.2 Transition mid-blastuléenne (TMB)	21
1.2.1.3 Transition maternelle-zygotique (TMZ)	21

TABLE DES MATIÈRES

1.2.2	L'embryogénèse précoce chez <i>Drosophila melanogaster</i>	23
1.2.2.1	Transition mid-blastuléenne (TMB)	23
1.2.2.2	Mise en place de la polarité de l'embryon	23
1.2.2.3	Transition maternelle-zygotique (TMZ)	27
1.2.3	Apport des études transcriptomiques	32
1.2.4	Objectif de la thèse	33
2	Identification de gènes co-exprimés pendant l'activation transcriptionnelle du gé-	35
	nome zygotique (AGZ)	
2.1	Mise en place d'un protocole d'analyses de séries temporelles à partir des don-	
	nées de Pilot et al.	36
2.1.1	Normalisation	36
2.1.2	Sélection de gènes différentiellement exprimés	38
2.1.3	Développement d'une approche pour la sélection de signatures trans-	
	criptomiques discrètes	40
2.1.3.1	Approches classiques de clustering	40
2.1.3.2	Clustering sur base de la discrétisation des profils de transition	42
2.2	Modèles de régulation de l'activation des gènes : Lu et al. (2009)	45
2.3	Contributions maternelle et zygotique : De Renzis et al. (2007)	49
2.4	Analyse de la composition des clusters	51
2.5	Conclusion du chapitre	52
3	Analyse fonctionnelle des gènes et des régions non-codantes associées	55
3.1	Analyse des clusters primaires	55
3.1.1	Enrichissement fonctionnel	55
3.1.2	Éléments cis-régulateurs	58
3.1.2.1	Choix du modèle de background pour la détection de mots	
	sur-représentés	59
3.1.2.2	Classification des clusters transcriptionnels en fonction des	
	motifs cis-régulateurs découverts	62
3.1.2.3	Classification des mots et dyades	62
3.1.2.4	Sélection des gènes pour l'étude de l'AGZ	66
3.2	Analyse du cluster AGZ	67
3.2.1	Analyse fonctionnelle avec Gene Ontology	67

TABLE DES MATIÈRES

3.2.2	Analyse des éléments cis-régulateurs	71
3.2.2.1	Découverte de motifs	71
3.2.2.2	Sélection des motifs parmi les matrices découvertes	71
3.2.2.3	Deux motifs connus retrouvés par l'approche <i>de novo</i> et par CisTargetX (Zelda et Trl)	72
3.2.2.4	Autres motifs connus	72
3.2.2.5	Motifs inconnus	74
3.2.2.6	Détection de modules cis-régulateurs (CRM) potentiels	74
3.3	Matériels et Méthodes	76
3.3.1	Récupération des séquences non-codantes	76
3.3.2	Paramètres pour la découverte de motifs	78
3.3.2.1	<i>oligos-analysis</i>	78
3.3.2.2	<i>dyad-analysis</i>	78
3.3.2.3	<i>matrix-from-pattern</i>	78
3.3.3	Paramètres pour la prédiction de sites et de modules cis-régulateurs (CRM)	79
3.3.4	Bases de données de PSSMs	79
3.3.4.1	FlyFactorSurvey	79
3.3.4.2	Jaspar core insect	80
3.4	Conclusion du chapitre	80
4	Analyse génomique des profils d'occupation et d'accessibilité de la chromatine	83
4.1	Utilisation des pics	85
4.1.1	Chevauchement entre et régions non codantes des gènes AGZ	85
4.1.2	Analyse des pics avec <i>peak-motifs</i>	86
4.2	Utilisation des densités de "reads"	88
4.2.1	Les contrôles positifs et négatifs	88
4.2.2	Méthodologie	91
4.2.2.1	Calcul de densité intégrée	91
4.2.2.2	Construction des courbes ROC	92
4.2.2.3	Combinaison d'expériences	95
4.2.3	Résultats	95
4.2.3.1	La chromatine dans tous ses états	96

TABLE DES MATIÈRES

4.2.3.2	Localisation de Zelda	103
4.3	Conclusion du chapitre	104
5	Conclusion	107
5.1	Proposition d'un mécanisme de régulation de l'AGZ	107
5.1.1	Avant l'AGZ	107
5.1.2	Activation de la transcription	109
5.1.2.1	Fin de la répression	109
5.1.3	Rôles de Trl et modèles alternatifs	110
5.1.3.1	Trl est un facteur de transcription et un co-activateur	111
5.1.3.2	Trl pourrait recruter CBP	111
5.1.3.3	Trl est impliqué dans la pause des ARN polymérase II	111
5.1.4	Utilisation du modèle pour la sélection de CRM	111
5.2	Conclusion générale	112
	References	115
6	Annexes	121
6.1	Résultats supplémentaires	121
6.2	Articles	121
6.2.1	From Peaks to Motifs : A complete workflow for full-sized ChIP-seq (and like) dataset	121
6.2.2	Evolution of major histocompatibility complex by "en bloc" duplication before mammalian radiation.	162
6.2.3	Integration of Evolutionary Biology Concepts for Functional Annotation and Automation of Complex Research in Evolution : The Multi-Agent Software System DAGOBAH	179

Table des figures

1.1	Représentation schématique des principales méthodes pour la collection de sites de liaison de facteurs de transcription (FT)	4
1.2	Code d’ambiguïté IUPAC	7
1.3	Construction de la matrice poids-positions de Hunchback à partir d’une collection de sites obtenus par expérience de DNase1 footprinting disponible dans la base de données Jaspar	8
1.4	Méthodes de découvertes de motifs	11
1.5	Transformation d’un alignement de mots chevauchant en matrices avec <i>matrix-from-patterns</i>	13
1.6	Calcul et distribution de scores de sites attribués en considérant une matrice donnée	15
1.7	Distribution théorique des scores de probabilité des matrices Hb (A) et découverte et construire à partir de l’alignement de mots chevauchants (B)	16
1.8	Vue comparative de l’embryogenèse précoce chez les métazoaires.	22
1.9	Développement précoce de l’embryon de la drosophile entre 0h et 3h après la fécondation.	24
1.10	Mise en place des axes embryonnaires durant l’oogenèse.	26
1.11	Modèle de la spécification antéro-postérieure (AP) par les gènes à effet maternel.	27
2.1	Répartition temporelle des données transcriptomiques utilisées pour l’étude de l’AGZ.	37
2.2	Distribution des log2 ratio pour chaque transition entre classes temporelles consécutives.	41
2.3	Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression.	43

TABLE DES FIGURES

2.4	Méthode de discrétisation des profils de transition.	44
2.5	Clusters résultant du regroupement de profils discrets identiques.	46
2.6	Application et résultats de la discrétisation des profils de transitions obtenus à partir de des données de Lu et collaborateurs.	49
2.7	Comparaison des groupes de gènes co-exprimés en fonction de la significativité (binomiale corrigée pour les multi-tests) de leur chevauchement.	53
3.1	Schéma d'analyses fonctionnelles pour la prédiction d'éléments et modules cis-régulateurs potentiels impliqués dans la régulation de l'AGZ.	56
3.2	Composition en nucléotides de différents types de séquences.	60
3.3	Heatmap représentant les résultats d'oligo-analysis sur les séquences en amont du TSS des gènes AGZ en utilisant divers modèles de background.	61
3.4	Clustering des oligonucléotides et des groupes de gènes en fonction de la significativité de la surreprésentation.	63
3.5	Correspondances entre motifs découverts et motifs connus.	73
3.6	Liste des motifs sélectionnés pour la prédiction de CRM.	75
3.7	Représentation schématique des options utilisées dans <i>retrieve-ensembl-seq</i>	77
4.1	Visualisation de données provenant d'une expérience de liaison de Trl entre 0 et 8h après la fécondation (ChIP-seq).	87
4.2	Résumé de la découverte de motifs effectuée sur les pics CBP, Trl, H3K4me1 et accessibilité à la DNase1.	89
4.3	Résumé de la découverte de motifs effectuée sur les pics Zelda.	90
4.4	Représentation schématique du calcul de la densité en reads intégrée sous une région données (A) et obtention d'une courbe de ROC (B).	93
4.5	Courbe de ROC correspondant au classement des CRM AGZ (prédits dans les région en amont du TSS) parmi des régions non codantes choisies au hasard en fonction des densité intégrée obtenues avec les donnée de ChIP-seq du facteur Zelda (3h).	94
4.6	Distribution des valeurs d'AUC calculées en fonction des intensités de liaison, d'occupation ou d'accessibilité de la chromatine sous les CRM AGZ et contrôles.	96
4.7	Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques d'ouverture de la chromatine (DNase1) et de liaison de l'ARN Polymérase II (polII).	97

TABLE DES FIGURES

4.8	Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques associées aux régions activatrices.	98
4.9	Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques associées au promoteurs actifs.	99
4.10	Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles de la liaison de Trl.	100
4.11	Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques répressives.	101
4.12	Résultats de l'analyse de la combinaison de différentes marques.	102
4.13	Analyse de l'enrichissement des CRM AGZ et contrôles de la liaison de Zelda.	103
4.14	Intervalle temporels de d'expériences de localisation génomique de facteurs, d'histones et d'accessibilité de la chromatine.	105
5.1	Présentation d'un modèle spéculatif pour la régulation de l'activation transcriptionnelle du génome zygotique (AGZ).	108
5.2	Analyse positionnelle de l'enrichissement des CRM en signaux de liaison de Trl et sites Trl par rapport au TSS.	112
5.3	Visualisation de la région en amont du gène <i>crocodile</i>	113
6.1	Exploration des clusters groupant les gènes potentiellement activés pendant l'AGZ selon Pilot et collaborateurs obtenus par la méthode de détection MAS5.0.	122
6.2	Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression.	123
6.3	Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression à partir des profils de transitions.	124
6.4	Exploration des clusters groupant les gènes potentiellement activés dépendamment du ratio NC pendant l'AGZ selon Lu et collaborateurs.	125
6.5	Profils temporels et de transition des clusters les plus importants ne correspondant à aucun modèle de régulation connu obtenus à partir des données de Lu et collaborateurs	126
6.6	Clustering des dyades et des groupes de gènes en fonction de la significativité de la surreprésentation des dyades.	129
6.7	Résumé des résultats CisTargetX pour le cluster AGZ.	130
6.8	Action de Su(H) dans la définition du méssectoderme.	131

TABLE DES FIGURES

Liste des tableaux

1.1	Description de bases de motifs de liaison générales (TRANSFAC, JASPAR) et spécialisées pour la drosophile (Fly Factor Survey, DMMPMM).	19
2.1	Récapitulatif des interprétations biologiques des profils discrets présentant au moins 10 membres dans les données de Pilot (2006)	47
2.2	Récapitulatif des interprétations biologiques des profils discrets présentant au moins 10 membres dans les données de Lu (2009).	50
3.1	Comparaison de la composition de gènes entre le cluster ZGA et les classes de la Gene Ontology.	70
4.1	Significativité du chevauchement entre pics et séquences non codantes des gènes AGZ	86
6.1	Codes des indications pour l'annotation des gènes dans GO et composition des classes	127
6.2	Top 10 des 7-mères surreprésentés suivant divers modèle de background.	128

ACRONYMES ET ABRÉVIATIONS

Acronymes et Abréviations

ADN	Acide DésoxiriboNucléique
AGZ	Activation du Génome Zygotique
AP	antéro-postérieur
ARN	Acide RiboNucléique
ARNm	ARN messagers
ARNpolIII	ARN polymérase II
ChIP	Chromatine Immuno-Precipitation
CRER	Cis-Regulatory elements Enriched Region

CRM	Cis-Regulatory Module
DV	dorso-ventral
FN	faux négatifs
FP	faux positif
FT	Facteur de Transcription
GO	Gene Ontology
MT	microtubule
NC	Nucléo-Cytoplasmique
PIC	Pre-Initiation Complex
TFBS	Transcription Factor Binding Site (site de liaison de facteur de transcription)
TMB	transition mid-blastuléenne
TMZ	transition maternelle-zygotique
Trl	Trithorax-like
TSS	Transcription Start Site (site d'initiation de la transcription)
UTR	UnTranslated Region
Zld	Zelda

ACRONYMES ET ABRÉVIATIONS

1

Introduction

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

1. INTRODUCTION

1.1 La régulation transcriptionnelle

1.1.1 Mise en place de la machinerie basale de transcription

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

1.1.2 Les facteurs de transcription

1.1.2.1 Définition

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

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

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

1.1.2.2 Caractérisation des sites de liaison protéine-ADN

Afin d'identifier les séquences reconnues par les FT, plusieurs méthodes ont été mises au point et permettent de construire des collections de sites qui pourront, par la suite, être alignés afin d'obtenir une représentation des motifs représentant de façon générique la spécificité des séquences reconnues par chaque facteur. La figure 1.1 présente une liste non exhaustive des techniques utilisées pour détecter les sites de liaison des FT à l'ADN.

La technique de recherche d'empreinte de liaison protéine à l'ADN grâce à la DNase1 (4) sur l'ADN ("*DNase footprinting*" en anglais) requiert la définition préalable d'une région régulatrice relativement courte (généralement une partie de promoteur). Ainsi, une seule séquence à la fois est analysée. Brièvement, la région à analyser est amplifiée par PCR puis marquée par fluorescence. Le FT soupçonné de se lier dans cette région est rajouté à une partie des échantillons (" positifs "), les autres servant de contrôle négatif. La DNase1 est ajoutée à chaque échantillon (positifs et négatifs) et va digérer l'ADN. Les nucléotides directement liés à la protéine sont protégés de la dégradation et laissent ainsi une " empreinte " révélée par la comparaison de la migration sur gel des deux échantillons. Enfin, la technique de Maxam et Gilbert (5) permet de lire directement la séquence des segments d'intérêt sur le gel.

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

1. INTRODUCTION

Technique	Caractéristiques	Principe
Empreinte à la DNase1	Séquence régulatrice unique correspondant à un promoteur donné Détection de sites de liaison de différentes affinités Peut-être peu précis <i>In vitro</i>	
SELEX	Banque d'oligonucléotides aléatoires Détection de sites à haute affinité <i>In vitro</i>	
Système bactérien à hybridation simple	Banque d'oligonucléotides aléatoires Détection de sites à haute affinité Limitation de la longueur des oligonucléotides testés Problème de compatibilité eucaryote/procaryote <i>In vitro</i>	
ChIP + hybridation sur puce (1)	Localisation sur le génome entier Détection de sites de liaison de différentes affinités Production d'anticorps spécifiques <i>In vivo</i>	
ChIP + séquençage à haut débit (2)		

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

itérativement (typiquement 4 ou 5 cycles d'amplification-sélection) et permet de sélectionner un petit nombre de séquences ayant une forte affinité et spécificité avec la protéine. Les oligonucléotides sélectionnés sont ensuite séquencés, plusieurs méthodes à haut ou à bas débit sont alors utilisées (7). Cette méthode ne permet donc pas d'identifier tous les sites de liaison possibles du FT : d'une part la méthode repose sur des oligonucléotides artificiels, et n'indique pas la position génomique des sites réels. D'autre part, les collections de sites obtenues par SELEX présentent généralement une affinité "optimale" pour le facteur, et ne représentent pas la diversité des séquences reconnues *in vivo* par la protéine.

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

Les méthodes combinant l'immuno-précipitation de la chromatine (ChIP) et la détection de la séquence nucléotidique par hybridation sur puce à ADN (11) (ChIP-chip) ou par séquençage à haut débit (12) (ChIP-seq), permettent de sélectionner tous les sites liés par le facteur d'intérêt en procédant à l'extraction de protéines liées à l'ADN à partir d'une cellule vivante. Cette

1. INTRODUCTION

méthode est probablement celle qui permet d'obtenir la plus grande diversité de séquences reconnues par le facteur d'intérêt. De plus, l'utilisation de cellules vivantes permet d'étudier la liaison du facteur dans différents tissus et à différents stades du développement. Cependant, ces techniques reposent sur l'utilisation d'anticorps spécifiques qui ne sont pas forcément disponibles et dont la production est coûteuse.

A titre d'illustration, la figure 1.3 montre les sites résultant d'expérience de DNase1 footprinting avec le facteur Hunchback (Hb) sur les séquences promotrices du gène *even-skipped* (13). Hb est un facteur impliqué dans la polarisation antéro-postérieure de l'embryon.

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

À partir des collections de sites générées par une des méthodes citées ci-dessus, ici DNase1 footprinting, nous pouvons procéder à l'alignement 1.3A de ces sites afin d'obtenir un motif de liaison, qui représente la spécificité de liaison d'un facteur transcriptionnel. Un type familier de représentation est la séquence consensus (figure 1.3B) qui donne une idée de la conservation et des variants possibles au moyen du code IUPAC (figure 1.2). Cependant la génération de telles séquences est relativement arbitraire (3, 14), en particulier les pourcentages retenus pour considérer qu'une position donnée sera représentée par un nucléotide unique, une paire ou un triplet. Dans le cas de Hb, il est intéressant de noter que la séquence consensus (figure 1.3B) ne permet d'identifier aucun des 16 sites de la collection (figure 1.3A). Pour pouvoir l'utiliser à des fins de reconnaissance, il faudrait réduire le consensus et le rendre plus générique (en admettant plus de variations), mais il perdrait alors sa spécificité et détecterait des sites non liés par le facteur.

Une alternative pour la représentation des motifs de liaison est la matrice occurrence-position (figure 1.3C) dans laquelle nous pouvons voir le nombre de chaque nucléotide (ligne) à chaque position (colonne) d'un alignement de sites. On peut en dériver la matrice poids-positions (figure 1.3D). Le poids d'un nucléotide est calculé en fonction de sa fréquence relative au sein de l'alignement, corrigée par un pseudo-poids pour éviter les biais dus aux faibles nombres d'observations, divisée par la probabilité a priori de chaque nucléotide (fréquence de chaque nucléotide dans le modèle de background, par exemple l'ensemble des séquences non-codantes d'un génome). La formule est présentée dans la figure 1.5. Ce motif peut être représenté sous forme de logo, (figure 1.3E), grâce à l'algorithme weblogo (15) (<http://weblogo.berkeley.edu/>). Le logo consiste en un empilement des lettres A,C,G,T dont la hauteur

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

FIGURE 1.2: Code d'ambiguïté IUPAC -

totale représente le contenu informationnel (la " conservation ") associé à chaque position et la hauteur de chaque nucléotide dans la pile représente sa fréquence relative (16). La représentation d'un motif par une matrice est beaucoup plus informative que la séquence consensus, et le logo en découlant donne une bonne impression subjective de l'importance de chaque résidu pour la liaison protéine-ADN. Nous pouvons enfin estimer le contenu informationnel de chaque cellule d'une matrice en multipliant le poids par la fréquence relative. Le contenu informationnel de la matrice est calculée en additionnant le l'information de tous ses cellules. Il indique la capacité du motif, représenté par la matrice, à discriminer des sites réels d'un le modèle de background (estimé sur base de la fréquence de chaque nucléotide dans un jeu de séquences de référence) (17). Nous verrons par la suite que certains algorithmes optimisent cette grandeur pour découvrir des motifs surreprésentés dans des jeux de séquences particuliers.

1.1.2.4 Découverte de motifs

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

Les méthodes basées sur l'optimisation tentent de collecter une matrice qui optimise un score donné (contenu informationnel ou autre) étant donnée une série de paramètres fournis par l'utilisateur (longueur de matrice, nombre de sites à aligner). Pour chaque motif généré, le score est calculé et les motifs ayant les meilleurs scores sont retenus. Le nombre de matrices

1. INTRODUCTION

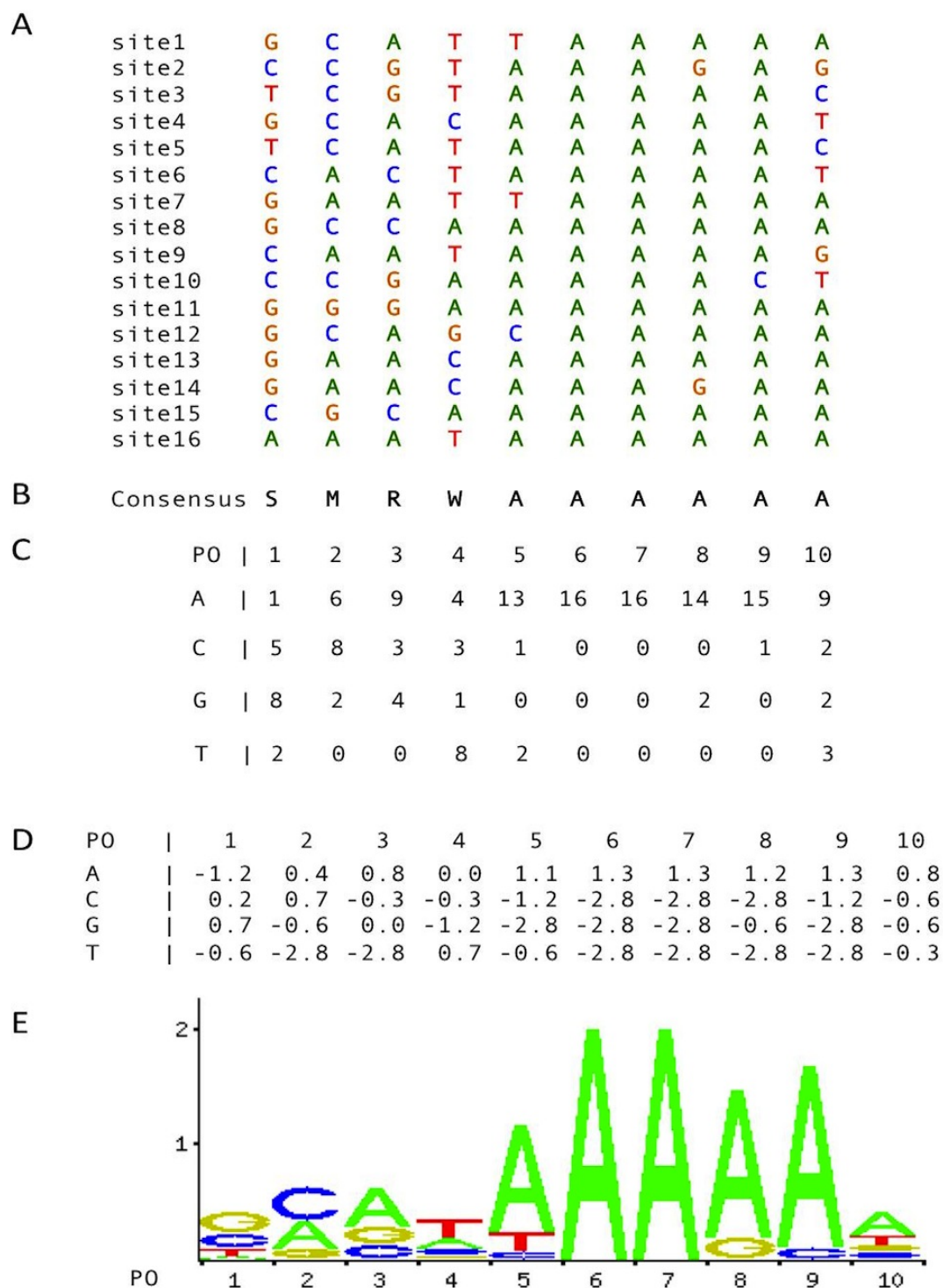


FIGURE 1.3: Construction de la matrice poids-positions de Hunchback à partir d'une collection de sites obtenus par expérience de DNase1 footprinting disponible dans la base de données Jaspar - (identifiant de la matrice : P05084). A. Alignement de sites. B. Séquence consensus. C. Matrice de comptage en format tabulé. D. Matrice poids-positions en format tabulé. E. Logo. PO : positions des nucléotides.

possibles dépend du nombre et de la taille des séquences d'entrées, de la taille du motif recherché et du nombre et de la répartition attendus des sites. Le nombre de matrices possibles augmente de façon exponentielle en fonction du nombre de séquences dans le jeu si l'on cherche exactement un site par séquence (par exemple pour les collections de sites), mais ce nombre devient encore plus énorme si l'on estime qu'une séquence peut contenir zéro ou plusieurs sites (ex. séquences non-codantes de gènes co-exprimés), en effet, ceci revient à chercher toutes les combinaisons possibles d'une longueur donnée (taille de la matrice) parmi toutes les positions d'un jeu de séquences. Afin de pallier ce problème, les algorithmes basés sur les matrices ne testent qu'une partie des possibilités et retourne alors le meilleur score pour cette partie de possibilités. Je vais brièvement décrire trois algorithmes : MEME (18), CONSENSUS (19) et *info-gibbs* (20). MEME (Multiple Expectation Maximisation for Motif Elicitation) va tester chaque oligonucléotide d'une taille donnée en construisant une matrice "semence" et va scanner les séquences afin de détecter les meilleurs sites (en fonction du poids). À chaque itération, un score de vraisemblance de la matrice par rapport à un modèle de référence (modèle de Markov) est calculé. Pour chaque position des séquences d'entrée, on calcule un score de vraisemblance d'appartenir soit au motif soit au modèle de fond. Une nouvelle matrice est alors construite en collectant les sites les plus vraisemblables. Les itérations continuent jusqu'à convergence vers un score localement maximal. Le programme CONSENSUS repose sur un algorithme glouton : il crée toutes les matrices de taille w possibles à partir des deux premières séquences du jeu puis va retenir les matrices les plus informatives. À l'étape suivante, toutes les combinaisons possibles entre ces matrices et tous les sites de la séquence suivante sont calculées et comme précédemment, les matrices les plus informatives sont gardées et ainsi de suite jusqu'à que toutes les séquences aient été prises en compte. Cet algorithme est sensible à l'ordre des séquences en entrée. *info-gibbs* est basé sur l'optimisation du score de vraisemblance, mais cette fois-ci, contrairement à MEME, cet algorithme va incorporer une composante stochastique, qui réduit le problème de convergence vers des maxima locaux. À l'étape suivante, un site ayant servi à construire la matrice est retiré et est remplacé par un autre site tiré au hasard avec une probabilité proportionnelle à sa vraisemblance d'appartenir au motif. La matrice est actualisée et les scores des positions des séquences sont recalculés. Cette méthode permet d'éviter les optimums locaux mais n'analyse qu'une partie des possibilités.

Les méthodes basées sur le comptage de mots sont présentées dans la figure 1.4. Le principe de ces méthodes est de compter les occurrences de chaque oligonucléotide d'une taille donnée et de détecter ceux qui ont une fréquence exceptionnellement élevée (ou faible, selon les options

1. INTRODUCTION

choisies). Les mots les plus fréquents ne sont pas forcément les plus pertinents puisqu'ils reflètent les biais de composition des séquences analysées (par exemple une richesse générale en A+T). Afin de détecter les mots surreprésentés, il faut comparer les fréquences obtenues à des fréquences attendues au hasard. Pour définir un modèle de background, deux stratégies alternatives peuvent être utilisées. La première consiste à estimer un modèle de Markov sur base des séquences analysées elles-mêmes. Le principe consiste à calculer la fréquence attendue d'un mot de taille k (la longueur des oligonucléotides analysée) en fonction de la composition des séquences en mots de taille plus petite. Alternativement, on peut utiliser un jeu de séquences de référence (ex. l'ensemble des promoteurs d'un organisme lors de l'étude de promoteurs de gènes co-exprimés) à partir duquel on calcule la fréquence attendue de chaque oligonucléotide de taille k . La significativité de la sur-représentation est calculée sur base d'une distribution binomiale qui repose sur l'hypothèse d'indépendance entre les positions successives. Cette hypothèse est invalide, et pose des problèmes pour les répétitions de mots auto-chevauchants ('GGGGGG', 'TATATATA' etc.). La solution apportée à ce problème est de ne compter que les occurrences renouvelantes, en excluant les occurrences chevauchantes. La p -valeur binomiale est alors calculée puis corrigée pour le multi-testing par la multiplication avec le nombre de tests effectués (le nombre d'oligonucléotides analysés). La significativité est alors calculée en transformant la e -valeur en fonction du logarithme de base 10 ($\text{sig} = -\log_{10}(e\text{-valeur})$) ce qui permet d'apprécier directement la surreprésentation du motif.

position-analysis (figure 1.4B) estime le biais positionnel des occurrences de chaque motif en comparant le nombre d'occurrences observées par rapport à un point de référence (centre, début, fin de séquences) au nombre d'occurrences attendues selon une distribution homogène. *local-word-analysis* (figure 1.4C) combine les approches de *oligo-analysis* et *position-analysis* en appliquant le test binomial pour détecter des mots exceptionnels dans des fenêtres de tailles fixes ou variables. Comme précédemment, les occurrences chevauchantes sont écartées du comptage. Pour chacun des algorithmes de comptage de mots *oligo-* et *dyad-analysis*, le choix du modèle de background est crucial, ceci sera discuté dans la partie Résultat de ce manuscrit.

La découverte de motifs retournant des matrices permet d'apprécier la variabilité de certaines positions des séquences reconnues par un FT, alors que les mots ne le permettent pas. Cependant, les algorithmes de découverte de mots détectent la plus part du temps des mots surreprésentés mutuellement chevauchants, qui révèlent différents fragments d'un même motifs. Par ailleurs, il est fréquent d'obtenir des mots qui diffèrent par une seule lettre, reflétant

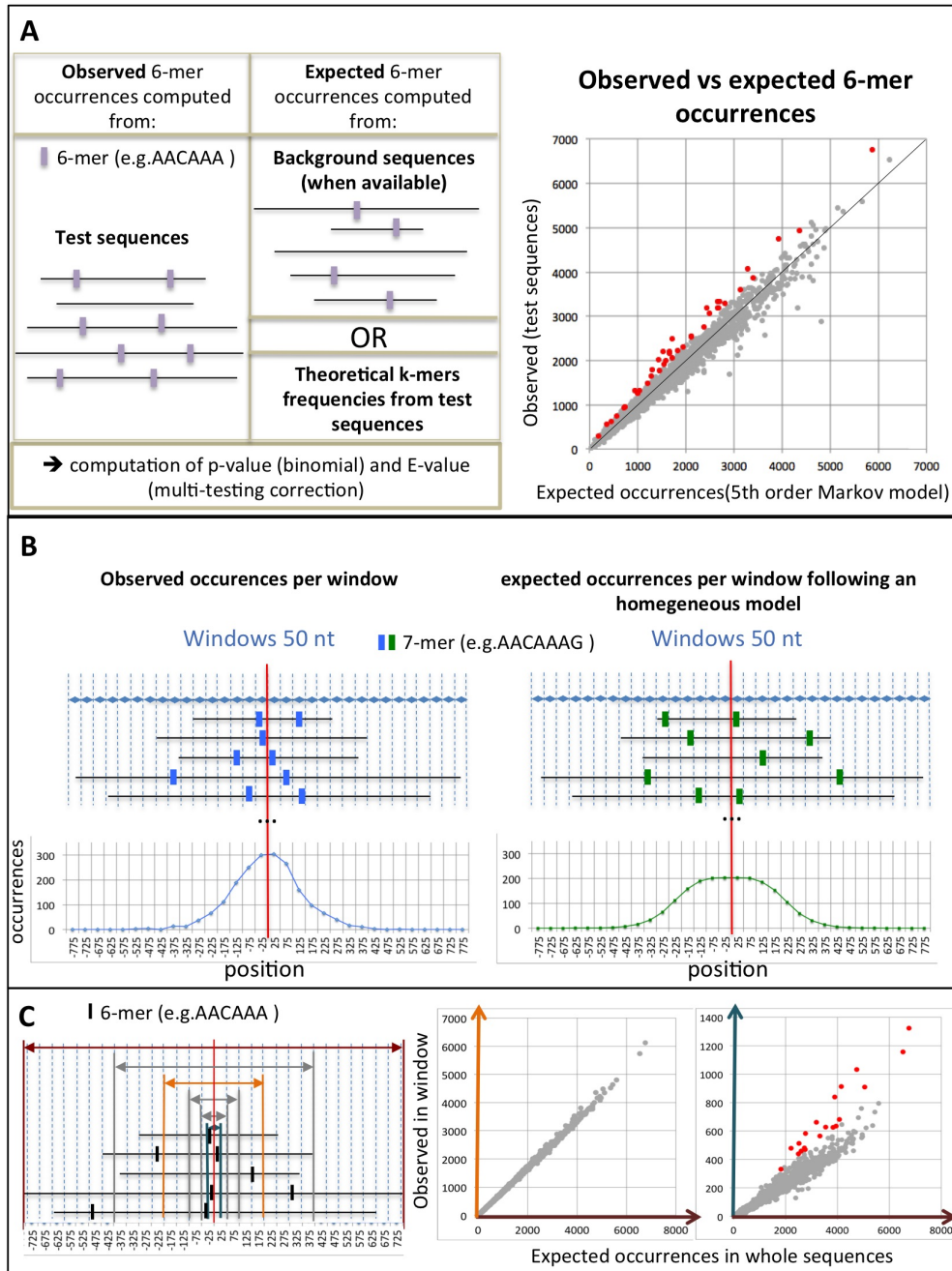


FIGURE 1.4: Méthodes de découvertes de motifs - A. *oligo-analysis*. Gauche : vision schématique du principe du test de sur-représentation pour un mot donné. Droite : Nombre d'occurrences observées pour chaque mot dans le jeu test (axe des ordonnées) comparé au nombre d'occurrences attendues selon le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

1. INTRODUCTION

les positions variables. Les mots retournés par *oligo-analysis*, *position-analysis* ou *local-word-analysis* peuvent être assemblés (*pattern-assembly*, figure 1.5A) afin de construire des matrices de significativité, à chaque nucléotide à chaque position, est assignée la significativité du mot le plus significatif contenant ce nucléotide (figure 1.5B). Les séquences sont scannées afin d'obtenir des matrices de comptage en appliquant un seuil sur le poids des sites retenus pour sa construction avec l'algorithme *matrix-from-patterns* (figure 1.5). Ici le poids de chaque site est calculé par rapport à la fréquence de bases dans le jeu de séquences testées. Par défaut, le poids minimum accepté pour que le site participe à la construction de la matrice est de 7. La séquence 1 (Seq1) de la figure 1.5G représente le site ayant le score maximum qui peut être obtenu. Nous pouvons noter que les matrices de significativité formées avec des mots faiblement significatifs ne détecteront aucun site pour leur construction et seront ainsi écartées. La séquence 2 (Seq2) est un exemple de site ayant un score inférieur au seuil de détection et qui ne sera donc pas utilisé pour la construction de la matrice de comptage. La prédiction des sites à partir des matrices ne peut pas seulement être basée sur un seuil sur le poids, nous verrons cela dans la partie "prédiction de sites".

1.1.2.5 Prédiction de sites

Nous pouvons prédire des sites à partir de motifs de liaison connus ou découverts sous forme de mots ou sous forme de matrices. Afin de prédire les sites correspondant aux motifs recherchés, il faut calculer un score à chaque position testée qui indiquera la qualité du site testé. Pour les motifs sous forme de mots, le score consistera à compter le nombre de substitutions qu'il faut appliquer au site pour correspondre au motif. Pour la prédiction à partir de matrices, le score de chaque site testé correspondra à la somme des poids des nucléotides présents dans la matrice. Dans le cas de la construction de matrices à partir d'assemblage de mots surreprésentés, le score du site correspondra alors la somme des significativités. Le scan des séquences avec l'une ou l'autre représentation des motifs va retourner un nombre important de faux positifs. Nous avons vu précédemment, dans l'exemple de la figure 1.3B, que les séquences consensus n'étaient pas efficaces pour la détection de sites et Day et McMorris (1992) (14) ont montré que le choix de la méthode de génération de séquences consensus devait être fait cas par cas suivant la question posée. Je ne parlerai ici que de la prédiction de sites à partir de matrices.

J'ai déjà un peu abordé ce sujet lorsque j'ai parlé de la transformation d'alignement de mots en matrice de comptage (figure 1.5). Durant cette étape, les sites étaient retenus si ils avaient

1.1 La régulation transcriptionnelle

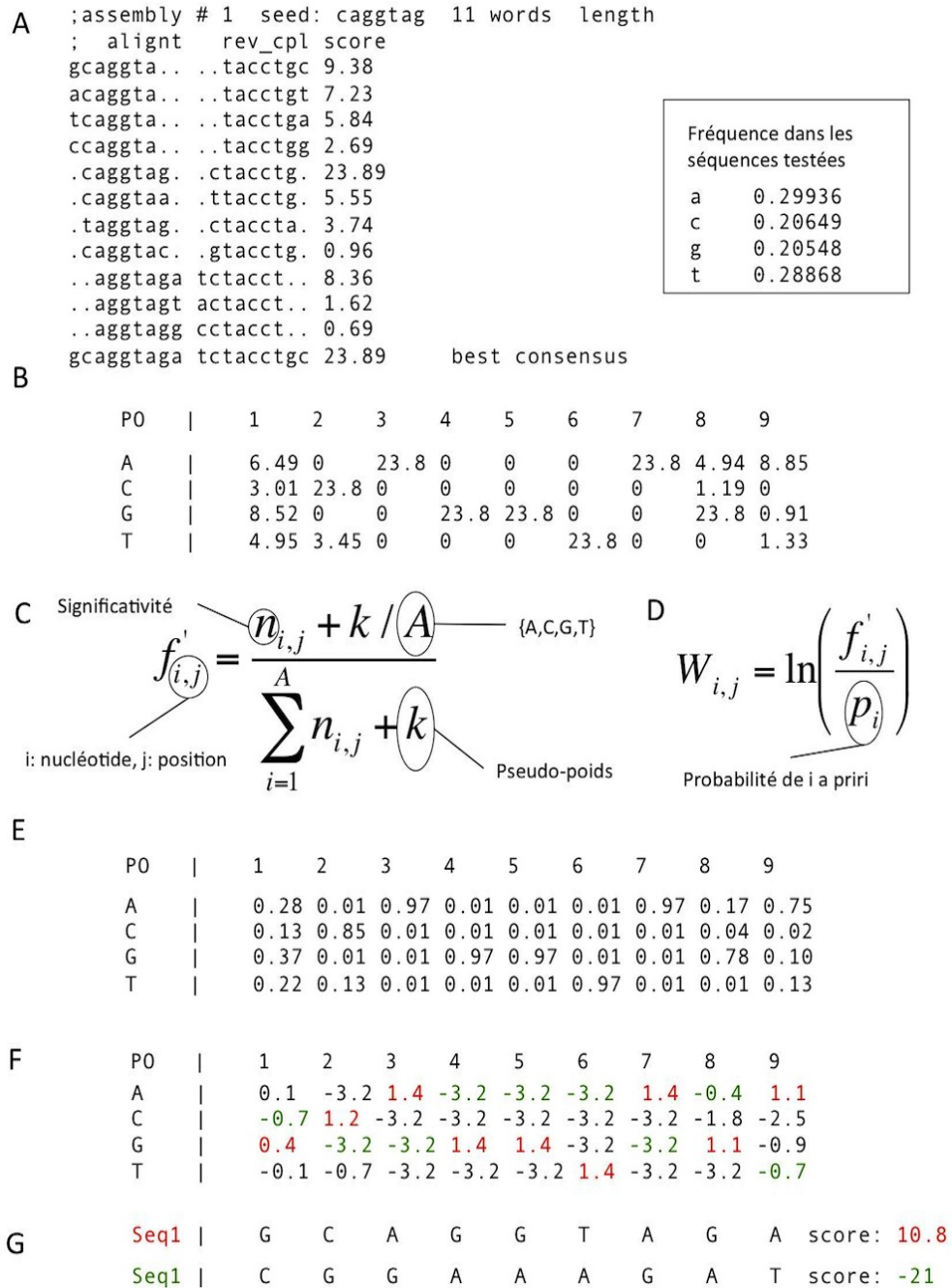


FIGURE 1.5: Transformation d'un alignement de mots chevauchant en matrices avec *matrix-from-patterns* - A. alignement de mots découverts chevauchants dont la significativité est indiquée dans la troisième colonne. B. Matrice de significativité. C. Calcul de la fréquence corrigée avec un pseudo-poids. D. Calcul du poids de chaque nucléotide i à chaque position j. E. Matrice poids-position. G. Seq1 et Seq2 sont deux exemples de séquences dont le score est indiqué en fin de ligne en rouge et vert respectivement. Le score est calculé en additionnant les nombres de même couleur dans la matrice correspondant aux significativités de chaque nucléotide respectivement aux séquences.

1. INTRODUCTION

un score au moins de 7 en utilisant un modèle de référence de Bernouilli qui assume l'indépendance entre les positions successives. Cependant, nous recherchons des séquences particulières, l'utilisation de modèle se basant sur la probabilité de dépendance entre les positions est alors plus appropriée. Pour cela, il convient d'utiliser des modèles de markov permettant de calculer la probabilité de trouver un nucléotide donné sachant la séquence précédente. Ainsi le poids calculé à chaque position est dépendant des positions précédentes. Le poids du site correspond alors au produit des probabilités de chaque position (figure 1.6B). La figure 1.6C montre la distribution des scores obtenus par la matrice de Hb sur la séquence de 2 Kb en amont du TSS du gène *even-skipped* (*eve*) qui contient trois des sites ayant servis à construire la matrice (étoiles rouges).

L'utilisation d'un seuil sur le poids est délicat lors la recherche de sites et d'autant plus si l'on cherche par la suite des régions enrichies en sites. En effet, les matrices courtes ou avec peu de positions informatives auront tendance à détecter des sites avec de plus faibles scores que des matrices plus longues. Il est donc préférable de se baser sur des critères statistiques s'appuyant sur la distribution théorique des scores de chaque matrice (*matrix-distrib*). Ainsi nous pouvons limiter le nombre de faux sites (faux positifs) détecter par la matrices. La figure 1.7 montre la distribution théorique de deux matrices. Nous pouvons voir que les deux matrices ont des comportements très différents. Par exemple, le score maximal que l'on peut obtenir avec chaque matrice, la matrice Hb (figure 1.7A) atteint un score de 7.9 alors qu'avec la seconde (figure 1.7B), le score maximal est de 12.7. Mais plus important, si l'on considère un score de poids de 5, nous pouvons voir que la p-valeur correspondante varie d'un facteur proche de 1000. Une p-valeur de 10^{-3} indique que l'on attend un faux positif toutes les 1000 positions. Il est donc préférable d'utiliser un seuil sur la p-value plutôt que sur le poids. Plusieurs algorithmes de détection de motifs existent dont patser (17), MotifLocator (21), matrix-scan (22) etc.

1.1.2.6 Prédiction de modules cis-régulateurs

Les facteurs de transcriptions n'agissant jamais seuls pour la régulation de la transcription de gènes, en se liant sur les régions promotrices distales restreintes appelées modules cis-régulateurs (CRM) (23, 24). Nous pouvons identifier de telles régions en recherchant des enrichissements en sites de liaison de FT (soit du même facteur, ce seront alors des CRM homotypiques, soit de différents facteurs, ce seront alors des CRM hétérotypiques). De nombreux algorithmes ont été développés afin de détecter les CRM. Beaucoup d'algorithmes se basent

A

P0		1	2	3	4	5	6	7	8	9	10
A		0.1	0.4	0.5	0.2	0.8	1.0	1.0	0.8	0.9	0.5
C		0.3	0.5	0.2	0.2	0.1	0.0	0.0	0.0	0.1	0.1
G		0.5	0.1	0.2	0.1	0.0	0.0	0.0	0.1	0.0	0.1
T		0.1	0.0	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.2

B

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

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

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

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

$$W_s = \ln \left(\frac{P(S|M)}{P(S|B)} \right)$$

W_s poids: log ratio des deux probabilités ci-dessus

C

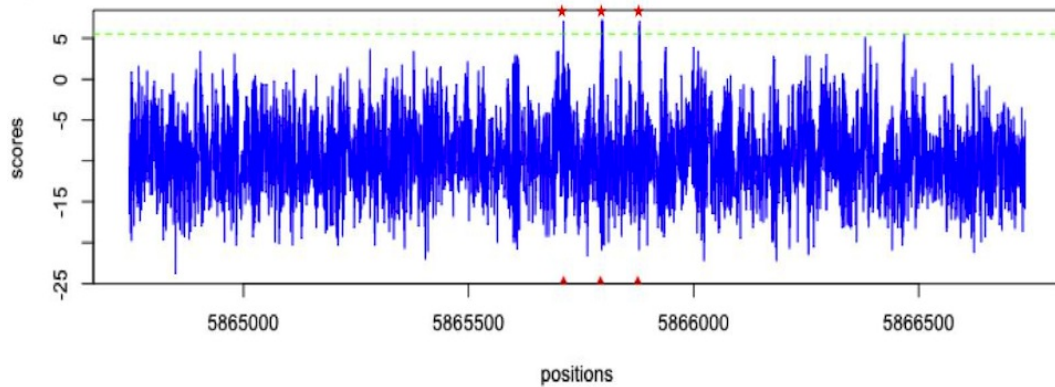


FIGURE 1.6: Calcul et distribution de scores de sites attribués en considérant une matrice donnée - A. Matrice de fréquences corrigées de Hunchback obtenue à partir d'une collection de sites provenant d'expérience de DNase1 footprinting sur le promoteur de *eve* (cf. 1.3). B. Calcul du poids d'un site étant donné un modèle de référence. C. Distribution des scores obtenus pour chaque position de la séquence de longueur 2kb en amont du TSS du gène *eve*. l'axe des abscisses indique les positions génomiques. Les étoiles indiquent les sites identifiés expérimentalement. La ligne verte pointillée indique un seuil de p-valeur de 10^{-3}

1. INTRODUCTION

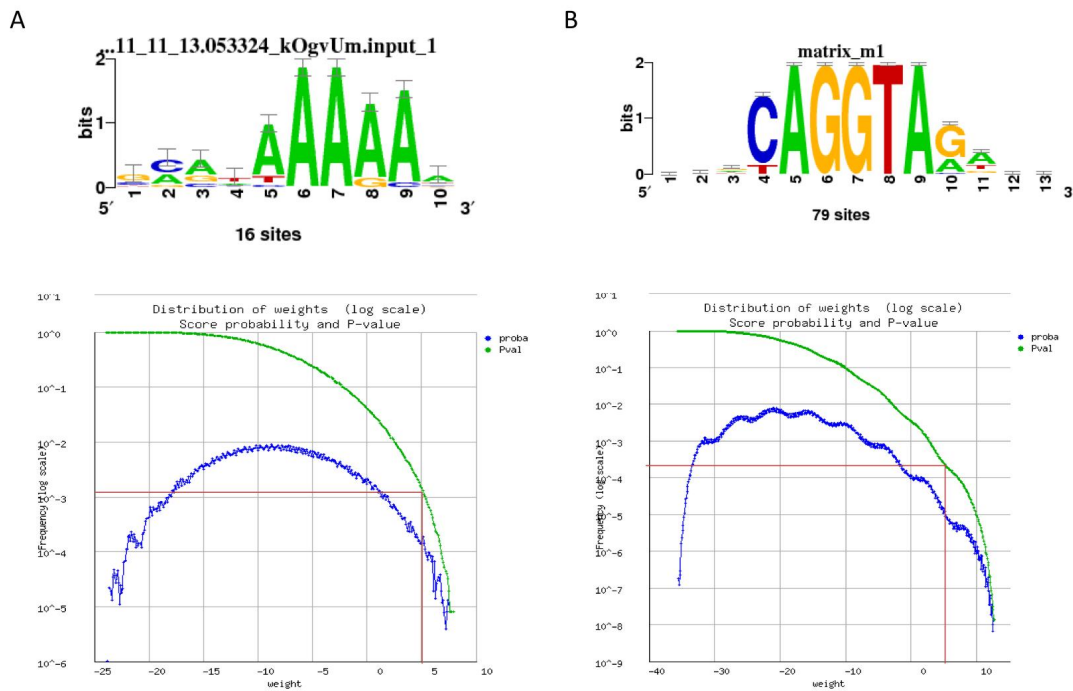


FIGURE 1.7: Distribution théorique des scores de probabilité des matrices Hb (A) et découverte et construite à partir de l'alignement de mots chevauchants (B) - La ligne rouge indique un score de 5.

sur la conservation entre espèces afin de prédire des CRM fonctionnels (Modulefinder (25), Target Explorer (26) ; TFBScluster (27), eCis-analyst (24, 28) etc. D'autres, en revanche, sont basés sur la significativité d'enrichissement de sites par rapport à un modèle de référence calculé sur les séquences à analyser (ClusterBuster (23) et matrix-scan (22)). Les deux derniers algorithmes diffèrent dans le choix des statistiques utilisées.

1.1.2.7 Bases de données

Plusieurs bases de données mettent à disposition, de façon publique ou commerciale, des jeux de motifs la plupart du temps sous forme de matrices occurrence-position¹. Il existe deux bases de données générales contenant des motifs de liaison pour plusieurs organismes.

TRANSFAC (29), qui est une base de données commerciale avec un accès public à certaines données, qui rassemble 1308 matrices (version 2010.1) dont 398 sont publiques, appartenant à six grands groupes taxonomiques (vertébrés, plantes, insectes, nématodes, champignons, bactéries). Les insectes sont peu représentés dans cette base puisque seulement 68 matrices sont disponibles (dont 38 publiques). Le format de matrices fourni par TRANSFAC est très informatif, en effet, différents champs sont renseignés en plus de la matrice elle-même. Les principaux champs sont : un identifiant unique indiquant le groupe taxonomique, le nom du facteur se liant au motif et la qualité des sites qui dépend de leur provenance ; un numéro d'accèsion ; une description courte du facteur ; les sites à partir desquels la matrice a été construite (pas toujours renseigné) ; et toute une série de champs renseignant sur la publication (auteurs, titre) d'où proviennent les données. Les sites proviennent préférentiellement d'expériences de DNase1 footprinting. JASPAR (30) est une base de données publique regroupant également des motifs provenant de plusieurs groupes taxonomiques. Elle contient 457 motifs dont 123 correspondent à des facteurs de drosophile. La source des sites utilisée est variée (voir tableau 1.1) et les motifs découlent de la découverte de motifs opérée avec l'algorithme MEME. D'autres bases de données spécialisées regroupent des motifs provenant d'espèces particulières. Durant mes travaux, j'ai utilisé particulièrement les bases de données spécialisées pour la drosophile. Fly Factor Survey (31) et DMMPMM (32) utilisent des sites provenant d'expériences diverses et ont appliqué des algorithmes de découverte de motifs différents (ChIP-Munk et MEME ou

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

1. INTRODUCTION

Consensus, respectivement) afin d'en extraire les motifs. Des bases spécifiques pour d'autres organismes sont disponibles (RegulonDB (33) pour *Escherichia coli* par exemple).

1.1.3 Les co-régulateurs

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

1.1.3.1 Le complexe médiateur

Le complexe médiateur a été identifié chez la levure par Kelleher et collaborateurs au début des années 1990 (43). Ce complexe est un adaptateur entre les FT liés aux régions régulatrices distales et l'ARNpolIII. Ce complexe est conservé chez les eucaryotes. Il est constitué de plusieurs sous-unités protéiques (30 chez l'homme, 21 chez la levure et au moins 25 chez la drosophile) organisées en trois modules distincts : module "tête", module "milieu" et module "queue". Le module "tête" interagit directement avec l'ARNpolIII, le module "queue" représente une plateforme d'interaction avec les FT et le module "milieu" transmet le signal via deux sous modules med9 (requis pour l'activation) et med10 (requis pour la répression) (44). Il existe plusieurs complexes médiateurs dont la composition varie en fonction du FT par lequel il a été recruté et du type cellulaire. Par exemple, chez la drosophile, le complexe médiateur contenant la sous-unité MED31 n'est pas requis pour la régulation globale de la transcription mais plutôt pour une régulation adaptée des gènes spécifiques de la segmentation au tout début du développement embryonnaire (45).

1.1.3.2 Les régulateurs de la chromatine

Dans le noyau, l'ADN est enroulé autour de complexes d'histones formant ainsi la chromatine. La chromatine peut arborer différents niveaux de compaction. Luger et collaborateurs (46) ont défini la structure du premier niveau de compaction consistant en 146 pb d'ADN entourant un octamère d'histones et appelé "nucléosome". Les nucléosomes sont les déterminants primaires de l'accessibilité de l'ADN. Les octamères d'histones sont composés de deux copies du tétramère H2A-H2B/H3-H4. Les extrémités N- et C- terminales (ou queues) des histones passent entre les tours d'ADN étant ainsi accessibles pour interagir avec d'autres protéines.

TABLE 1.1: Description de bases de motifs de liaison générales (TRANSFAC, JASPAR) et spécialisées pour la drosophile (Fly Factor Survey, DMMPMM). Les nombres entre parenthèses indiquent le nombre de motifs publics dans la base de données TRANSFAC. HMM : Hidden Markov Model ; B1H : Bacterial one-hybrid. 1. PAZAR, <http://www.pazar.info/>, Portales-Casamar (2007) (34, 35) ; 2. FlyReg, <http://www.flyreg.org/>, Bergman et al. (2004) (36) ; 3. Bergman Lab, <http://bergmanlab.smith.man.ac.uk/>, Down et al. (2007) (37) ; 4. RedFly, <http://red-fly.ccr.buffalo.edu/>, 5. Halfon et al. (2008) (38) ; 6. Noyes et al. (2008) (39), 7. Noyes et al. (2008) (40) ; 8. BDTNP, bdtnp.lbl.gov/, (41, 42).

Base de données	Espèce(s) / Taxon(s)	Nombre de matrices	Autres informations
TRANSFAC 7.0 Version 2010.1	Vertébrés	913 (293)	- sites compilés depuis la littérature, préférence pour les expériences DNase1 footprint - familles de motifs (HMM) - format de matrices descriptif
	Champignons	194 (31)	
	Plantes	124 (31)	
	Insectes	68 (38)	
	Nématodes	7 (4)	
	Bactéries	2 (1)	
	Total	1308 (398)	
JASPAR 4.0	Vertébrés	130	- sites : compilés depuis la littérature (DNase1), SELEX, B1H, ChIP-chip, ChIP-seq, familles (PAZAR ¹) - découverte de motifs avec MEME
	Champignons	177	
	Plantes	21	
	Insectes	123	
	Nématodes	5	
	Urochordés	1	
	Total	457	
Fly Factor Survey Version 09/05/11	<i>Drosophila melanogaster</i>	470	- sites : DNase1 footprint and SELEX ^{2,3,4} , B1H ^{5,6} - découverte de motifs avec MEME ou Consensus - redondance
DMMPMM/iDMMPMM	<i>Drosophila melanogaster</i>	41/39	- sites : DNase1 footprint ² , SELEX ^{2,7} , B1H ⁴ , ChIP-chip ⁸ - découverte de motifs avec ChIPMunk

1. INTRODUCTION

Les interactions ADN/histone ou histone/histone dépendent du repliement de chaque histone qui peut-être modifié par l'acétylation, la méthylation ou la phosphorylation d'acides aminés présents à leurs queues. La stabilité de ses interactions peut être également altérée par l'action de complexes de remodelages dépendant de l'ATP qui permettent le glissement des nucléosomes. Ce niveau de compaction permet un changement rapide entre l'état répressif et actif de la chromatine. La chromatine est alors appelée euchromatine. Le niveau de compaction supérieur implique la reconnaissance de certaines modifications d'histones par les protéines HP1 (hétérochromatin protein 1) (47). À ce niveau la chromatine est appelée hétérochromatine. La compaction de la chromatine affecte toutes les étapes de la transcription depuis la liaison des FT et la formation du PIC jusqu'à l'élongation. La compaction de la chromatine est un processus dynamique et implique de nombreux facteurs. L'impact de différentes modifications d'histones sur l'accessibilité de la chromatine sera présenté dans le chapitre 4 de ce manuscrit. J'y détaillerai également l'action de différents facteurs.

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

Durant la première moitié du 20^{ème} siècle, les biologistes ont effectué des observations anatomiques détaillées du développement embryonnaire de plusieurs espèces. L'embryologie comparée a mis en lumière de nombreuses similarités dans les étapes du développement précoce des embryons. En effet, après la fécondation, l'embryon va se segmenter ce qui va mener à la formation du blastoderme. Le développement des métazoaires passe toujours par une étape où l'embryon est formé de trois feuillettes, la gastrula comprenant l'ectoderme, l'endoderme et le mésoderme, nettement caractérisés, de chacun desquels dériveront des parties analogues chez tous les animaux. L'utilisation d'organisme modèle est alors pertinente pour étudier ces différents phénomènes. En 1978, Ed Lewis a montré, grâce à l'analyse de mutants de drosophile, que des gènes guidaient les mécanismes du développement (48). En 1980 Erik Wieschaus et Christianne Nüsslein-Volhard (49) ont caractérisés 120 gènes essentiels à une segmentation normale. L'embryon précoce des métazoaires connaît des changements morphologiques et une reprogrammation transcriptionnelle drastiques. La compréhension de ces phénomènes reste encore un enjeu fondamental et la drosophile semble être un bon système modèle pour leur étude. Durant ce chapitre, je vais décrire les mécanismes impliqués particulièrement dans la repro-

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

grammation transcriptionnelle de l'embryon, ce phénomène est appelé "transition maternelle-zygotique", et a constitué le cadre général ayant motivé cette thèse.

1.2.1 L'embryogénèse précoce chez les métazoaires : mécanismes communs

1.2.1.1 Activation de l'œuf

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

1.2.1.2 Transition mid-blastuléenne (TMB)

Chez les métazoaires, après la fécondation, les cycles mitotiques sont rapides et dépourvus de phases gap. La TMB correspond à un allongement de ces cycles par l'introduction des phases gap. Ce phénomène est suivi par la gastrulation, c'est à dire la migration et l'invagination des trois feuilletts embryonnaires (endoderme, mésoderme et ectoderme).

1.2.1.3 Transition maternelle-zygotique (TMZ)

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

1. INTRODUCTION

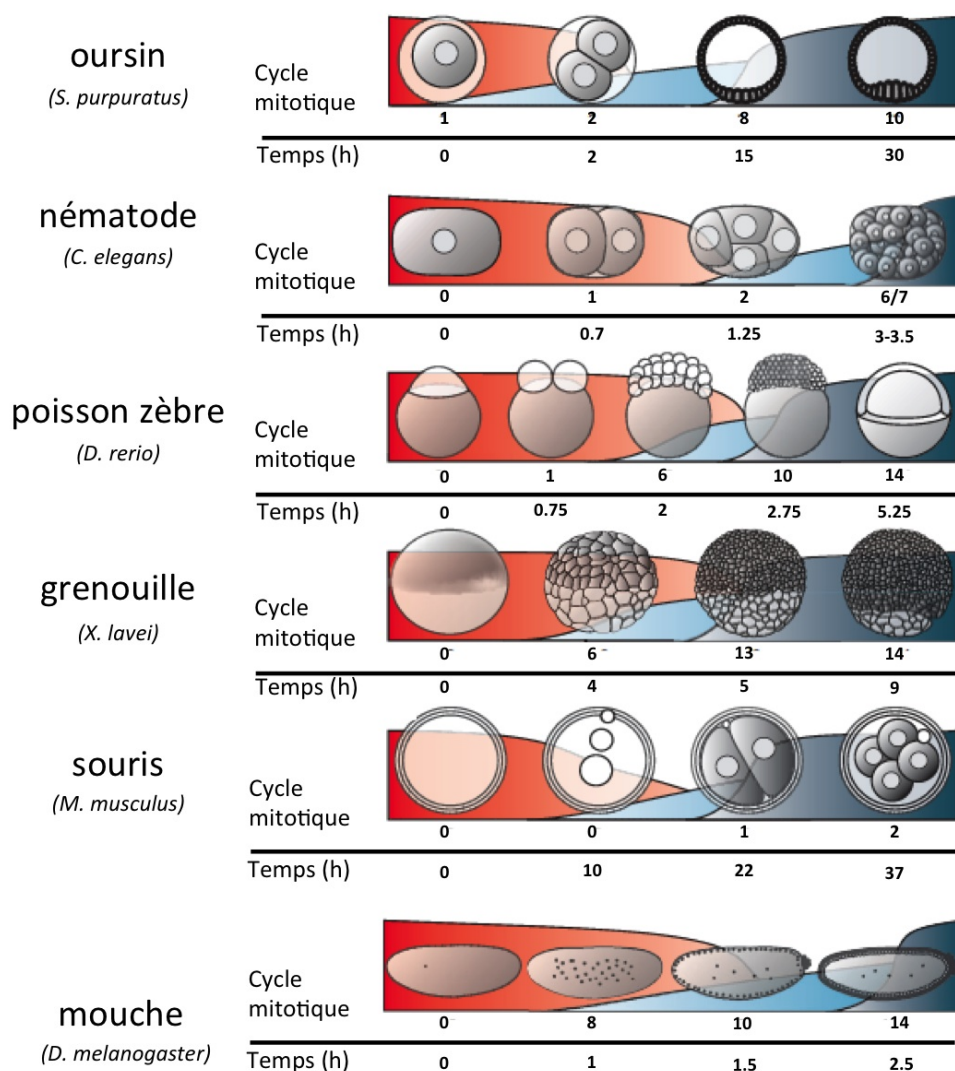


FIGURE 1.8: Vue comparative de l'embryogenèse précoce chez les métazoaires. - Représentation schématique des étapes embryonnaires clés pour plusieurs organismes modèles (oursin, nématode, poisson zèbre, grenouille, mouche et souris) précisant les cycles mitotiques et le temps après la fécondation correspondant. Quantités d'ARNm maternels (profil rouge) et d'ARNm nouvellement synthétisés lors des vagues mineure (profil bleu clair) et majeure (profil bleu foncé) d'activation du génome zygotique (AGZ). Le temps 0h correspond à la fécondation. (Figure modifiée à partir de Tadros et al. (2009) (50)).

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

1.2.2 L'embryogénèse précoce chez *Drosophila melanogaster*

1.2.2.1 Transition mid-blastulienne (TMB)

Chez la drosophile, dès la fécondation, l'embryon présente deux zones identifiables : le vitellus, sombre et opaque, qui occupe quasiment la totalité de l'embryon et le périplasma, plus clair, formant une couche périphérique. Les clivages mitotiques des noyaux ne sont pas accompagnés de cytokinèse (division du cytoplasme), l'embryon se développe donc sous forme d'un syncytium (cytoplasme contenant plusieurs noyaux) durant 13 cycles. Les dix premières divisions mitotiques durent environ huit minutes et sont synchrones. Dès le cycle 8, les noyaux jusque là répartis de façon homogène dans le vitellus, entament une migration radiale vers le périplasma. Les noyaux à la périphérie de l'embryon forment le blastoderme syncytial au cycle 10. Cette transition est marquée par la cellularisation des cellules polaires (précurseurs de la lignée germinale). À partir du cycle 7, on observe un ralentissement progressif des divisions nucléaires (10 minutes pour le cycle 10 jusqu'à 20 minutes au cycle 13) qui se produisent maintenant de façon métasynchrone (différences de vitesse de division entre les noyaux aux pôles (rapides) et dans le reste de l'embryon (51)). Ce ralentissement est dû à l'activation des points de contrôle de la réplication assurée par Chk2 et Chk1 et à la dégradation des ARNm *cycline B* (51, 52). L'introduction des phases gap au cours du cycle 14 entraîne un ralentissement important des mitoses en allongeant considérablement l'interphase. Cette pause permet la cellularisation des noyaux périphériques par invagination de la membrane plasmique. (figure 1.9A) On observe alors, à la fin du cycle 14, un épithélium, le blastoderme cellulaire, qui subira dès le début du cycle 15 la gastrulation. Il est à noter que la TMB est le premier changement morphologique assuré par le génome zygotique.

1.2.2.2 Mise en place de la polarité de l'embryon

Durant la première heure de développement, le génome de l'embryon de la drosophile est silencieux (figure 1.9B) et le développement est donc assuré par les ARNm dits « maternels », accumulés dans l'oocyte durant l'oogenèse. Les gènes produisant ces ARNm sont appelés gènes à « effet maternel ». Les gènes à effet maternels les plus étudiés sont ceux responsables de la mise en place des axes embryonnaires. La polarisation de l'embryon est initiée dès les premiers stades de développement de l'oocyte par quatre groupes de gènes : groupe A (spécification de la tête et du thorax), le groupe P (spécification de l'abdomen), le groupe T comprenant les gènes transcrits au pôle antérieur (acron), postérieur (telson) et dans les régions terminales

1. INTRODUCTION

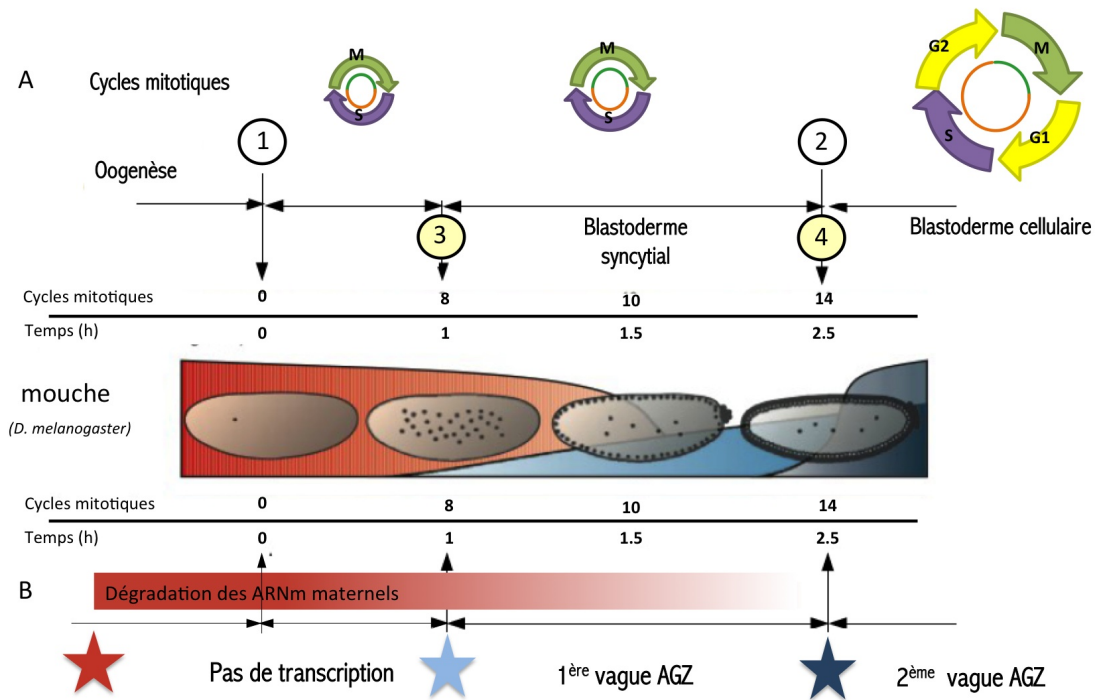


FIGURE 1.9: Développement précoce de l'embryon de la drosophile entre 0h et 3h après la fécondation. - A. La première ligne indique les phases des cycles mitotiques aux cours du temps. M : mitose (vert clair) ; S : réplication (violet) ; les phases gap G1 et G2 (jaune). Les temps de mitose et d'interphase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

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

respectivement, et enfin le groupe D impliqué dans la mise en place de l'axe dorso-ventral (DV), les groupes A et P étant impliqué dans la spécification de l'axe antéro-postérieur (AP). Le groupe T permet la différenciation des régions terminales mais n'entre pas directement dans la spécification des axes embryonnaires. La segmentation antéro-postérieure commence dans l'oocyte avec trois acteurs principaux *gurken* (groupe D), *oskar* (groupe P) et *bicoid* (groupe A) (figure 1.10). Le premier acteur *Gurken*, seul ARNm produit par le noyau de l'oocyte, est déterminant pour la mise en place de l'axe AP durant les stades 2 à 6 du développement de l'oocyte permettant la différenciation des cellules terminales en cellules postérieures (figure 1.10A). Les cellules postérieures vont induire à leur tour la réorganisation des microtubules (MTs) du cortex de l'oocyte ce qui va entraîner la migration du noyau au futur coin antéro-dorsal (figure 1.10B). Une deuxième vague de production de *Gurken* par le noyau va permettre la spécification de la face dorsale en inhibant la synthèse de *Pipe*, qui va être sécrété par les cellules épithéliales de la face ventrale et va permettre l'accumulation d'un produit encore non-identifié dans l'espace péri-vitellin (figure 1.10C). *Pipe* est le premier signal de ventralisation, le second se produira après la fécondation. Ainsi l'axe DV est mis en place. Les MTs, polarisés sur l'axe AP, permettent le transport des ARNm *bicoid* et *oskar* grâce à leur association à des moteurs moléculaires (respectivement dynéines et kinésines). Les protéines *Oskar*, au pôle postérieur, permettent la localisation restreinte des protéines *Nanos* (groupe P) en évitant le blocage de leur traduction induit par *Smaug* dans le reste de l'oocyte (53). Les ARNm *hunchback* et *caudal*, qui sont déterminants dans la segmentation AP de l'embryon sont également accumulés dans l'oocyte, cependant ceux-ci y sont distribués de façon homogène (figure 1.11A).

À la fécondation, ces ARNm peuvent être traduits. Les protéines *Bicoid*, qui sont des facteurs de transcription mais aussi des répresseurs de la traduction, vont diffuser en un gradient de concentration depuis le pôle antérieur et vont empêcher la traduction des ARNm *caudal* restreignant la synthèse des protéines *Caudal* au pôle postérieur (57). De façon opposée, la traduction des ARNm *nanos* induit un gradient postéro-antérieur des protéines *Nanos*, ces dernières, associées aux protéines *Pumilio* (groupe P), inhibent la traduction des ARNm *hunchback* au pôle postérieur (figure 1.11B et C).

La fécondation fournit aussi le second signal de spécification de l'axe DV. En effet, une cascade de phosphorylation va être initiée par l'activation des récepteurs *Toll* (groupe D) présents sur la face ventrale de l'embryon par les produits déposés préalablement dans l'espace péri-vitellin. Cette cascade mène à la libération de la protéine *Dorsal* dans le cytoplasme qui

1. INTRODUCTION

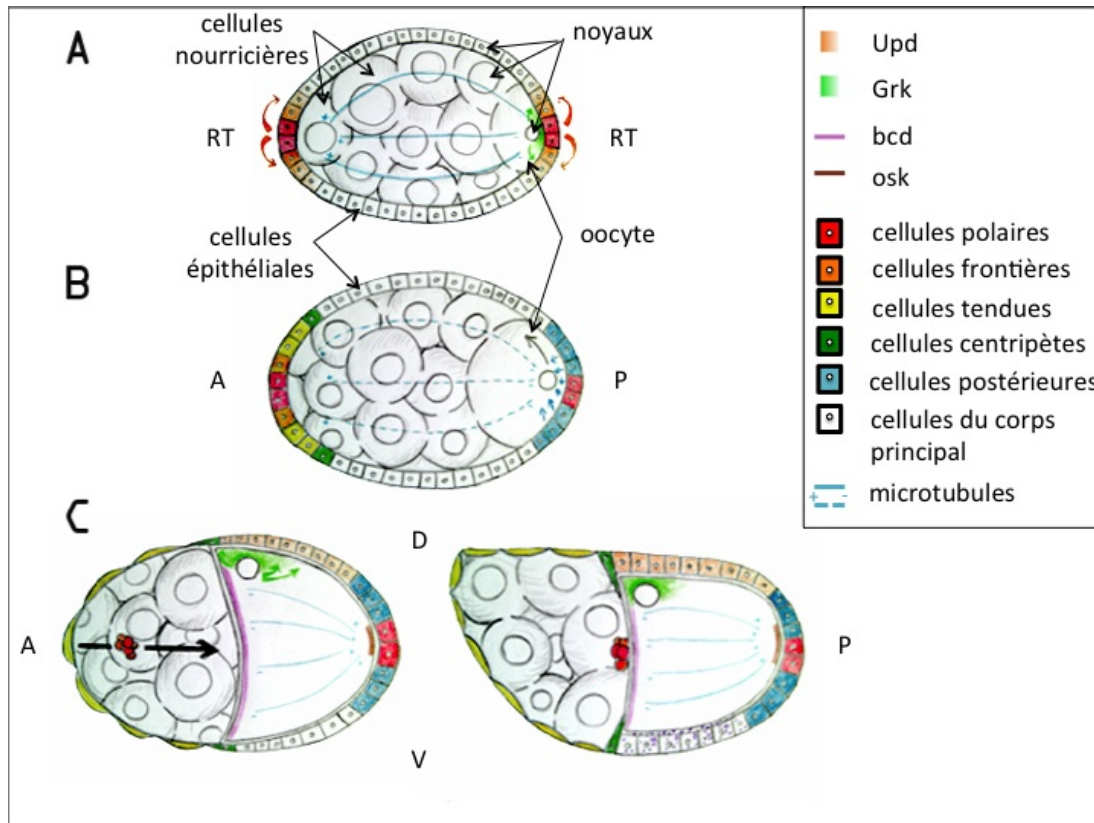


FIGURE 1.10: Mise en place des axes embryonnaires durant l'oogenèse. - (A) Stade 2-6 du développement de l'oocyte. Différents signaux moléculaires permettent la différenciation des cellules épithéliales de la région terminale (RT) : Upd en orange (voie JAK/STAT (54)), Gurken (Grk, vert clair). Le réseau de MT, résultant s'étend d'un pôle à l'autre du follicule permettant le transport des ARNm depuis les cellules nourricières (extrémité +) dans l'oocyte (extrémité -) (55). (B) Stade 7 du développement de l'oocyte : réorganisation des MT, différenciation des cellules RT en trois types (cellules frontières en orange, cellules tendues en jaune, cellules centripètes en vert) (56). (C) À gauche : début du stade 9, à droite : stade 10b. Stade 9 : migration des cellules polaires (flèche noire), spécification de l'axe dorso-ventral via Grk et qui induit la répression de la synthèse de Pipe (points violets) à la face dorsale de l'embryon. Déstabilisation et réorganisation des MTs du cortex de l'oocyte en un réseau orienté sur l'axe AP dont l'extrémité - se situe au pôle antérieur. Les ARNm de Bicoid (Bcd, rose) et Oskar (Osk, marron) vont être transportés de façon polarisée sur le réseau de MTs respectivement au pôle antérieur par les dynéines (+ -> -) et au pôle postérieur par les kinésines (+ -> -).

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

pourra alors être transportée dans les noyaux suivant un gradient ventro-dorsal. Bicoid, Hunchback, Dorsal et Caudal sont des facteurs de transcription appelés aussi morphogènes car ils vont pouvoir activer ou réprimer leurs cibles en fonction de leur concentration.

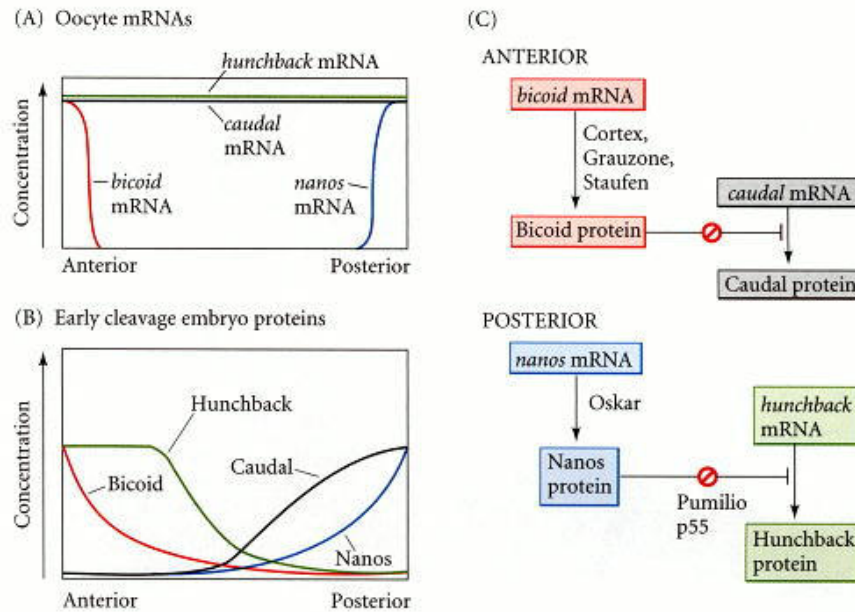


FIGURE 1.11: Modèle de la spécification antéro-postérieure (AP) par les gènes à effet maternel. - (A) Les ARNm *bicoid*, *nanos*, *hunchback* et *caudal*, déposés dans l'ovocyte par les cellules nourricières durant l'oogenèse. *bicoid* et *nanos* sont respectivement transportés au pôle antérieur et postérieur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

1.2.2.3 Transition maternelle-zygotique (TMZ)

Les premiers gènes activés, durant la première vague d'AGZ, sont les gènes gap (ex. *krüppel*, *knips*, *caudal*, *giant*) qui codent pour des FT à l'origine de la première vague d'activation transcriptionnelle des gènes impliqués dans la segmentation AP de l'embryon. L'activation des gènes gap est restreinte spatialement par les gènes à effet maternel. Bosveld et collaborateurs (2008) (45) ont mis en évidence l'implication d'un complexe médiateur comportant spécifiquement la sous-unité dMED31, synthétisée au niveau maternel, dans l'activation des

1. INTRODUCTION

gènes gap (*knips* et *kriipple*) et des gènes pair-rule (*fushi-tarazu* et *even-skipped*). Ce complexe médiateur constitue l'interface entre les morphogènes maternels et la machinerie basale de transcription et guidant ainsi l'expression zygotique des gènes primordiaux pour la détermination de l'identité des segments. Ce complexe est d'autant plus intéressant qu'il est conservé chez les métazoaires et pourrait jouer un rôle similaire dans l'AGZ de ces organismes. Cependant, l'action de ce médiateur n'explique qu'en partie la régulation de l'AGZ puisqu'il semble cibler spécifiquement les gènes de segmentation. Or plusieurs centaines de gènes sont activés durant les vagues de l'AGZ (58, 59, 60, 61, 62).

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

– La dégradation des ARNm maternel

Avant l'activation de l'oeuf, les transcrits d'environ 50% des gènes codant des protéines (environ 6000 gènes) sont présents l'oocyte (59, 61, 63). Thomsen et collaborateurs (2010) (64) ont récemment estimé que presque 4000 espèces d'ARNm sont dégradées durant la TMZ, dont un millier dès l'activation de l'oeuf (63). Deux voies indépendantes mais non exclusives sont impliquées dans la dégradation des ARNm. Une voie catalysée par les facteurs maternels et une voie zygotique catalysée par les nouveaux ARNm produits lors de l'AGZ. La dégradation de certains ARNm peut dépendre de l'action des deux voies (64). Le facteur maternel le plus connu, Smaug, a été identifié pour son implication dans le blocage de la traduction des ARNm nanos à l'activation de l'oeuf. Tadros et al. 2007 lui ont attribué la responsabilité de la dégradation d'environ 2/3 des ARNm dégradés dès l'activation de l'oeuf. La dégradation induite par Smaug semble être essentielle dans le ralentissement des cycles mitotiques syncytiaux et participerait à l'activation du point de contrôle de la réplication à partir du cycle 10 et la dégradation de la *cycline B* (52). Bushati et al. (2008) (65) ont mis en évidence l'élimination de 410 ARNm par un système purement zygotique impliquant le cluster de microARNs (miARNs) miR-309. Parmi ceux-ci, 138 ARNm sont dégradés dès l'activation de l'oeuf dont 92 sont des cibles de Smaug. De plus, dans les mutants smg, environ 85% des transcrits dégradés par cette classe de miARNs sont stabilisés (52). Ceci suggère que Smaug collabore avec les

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

facteurs zygotiques pour l'activation de l'expression zygotique de ces miRNAs, probablement indirectement via la dégradation de suppresseurs. Smaug et le cluster miR-309 n'expliquent pas la totalité de la dégradation des ARNm maternels. De Renzis et al. (2007) et Thomsen et al. (2010) ont mis en évidence l'implication de la protéine Pumilio, connue pour être impliqué dans le blocage de la traduction des ARNm maternels dans l'oeuf activé non-fécondé en s'associant avec Nanos (figure 1.11), mais aussi dans leur dégradation (66). D'autre part, les travaux menés par ces deux groupes et ont également mis en évidence l'enrichissement d'une famille de motifs contenant des éléments riches en AU (AU Rich Elements, AREs) dans les régions 3'UTR des gènes maternels. Ce type de motif est reconnu par des miARN et a été impliqué dans la dégradation de certains ARNm dans des cellules S2. L'action des miARNs serait exclusivement zygotique (63, 64).

La dégradation des ARNm maternels est un processus dynamique et spécifique (64) qui commence dans l'oocyte dès son activation. Trois raisons ont été proposées pour expliquer l'importance de la dégradation des ARNm maternels. Dans un premier temps, la dégradation pourrait compenser un dosage anormal des ARNm dans l'embryon (cependant cela devrait impliquer une dégradation globale plutôt que spécifique). Un deuxième modèle suggère que la dégradation spécifique de certains ARNm maternels distribués de façon ubiquitaire permettrait une expression localisée pour la mise en place de régions spécialisées (59, 64). Le dernier modèle propose une dégradation instructive plutôt que permissive. Par exemple, le niveau des ARNm impliqués dans le cycle cellulaire est responsable du ralentissement des cycles mitotiques jusqu'à la pause en interphase du cycle 14 qui permet la cellularisation. L'exemple du gène *string*, codant pour un régulateur des cycles mitotiques, permet d'illustrer ces 2 derniers modèles. En effet, l'élimination des ARNm maternels *string* à la fin de la TMZ laisse place à une expression zygotique régionalisée correspondant aux domaines mitotiques de l'embryon en gastrulation, ce qui est consistant avec le modèle permissif. Edgar et Datar (1996) (67) ont montré qu'une augmentation (ou une diminution) des ARNm *string* et *twine* entraîne une augmentation (ou une diminution) du nombre de cycles nucléaires avant la cellularisation. La dégradation est alors instructive.

– Régulation temporelle de l'AGZ

Trois modèles ont été proposés pour la régulation de la répression et l'activation de l'AGZ au niveau temporel : le ratio nucléo-cytoplasmique (diminution du rapport ADN/cytoplasme au fil des cycles mitotiques dans le syncytium), l'horloge maternelle (temps absolu après la

1. INTRODUCTION

fécondation) et l'avortement de la transcription dû à la rapidité des premiers cycles mitotiques (cycle 1 à 8, cf. figure 1.9A).

Jusqu'à la TMB, le volume global de l'embryon est stable alors que la quantité d'ADN augmente exponentiellement, modifiant ainsi le ratio ADN/cytoplasme (ratio NC). Edgar et al. (1989) (68) ont montré que la diminution de moitié du contenu en ADN dans des embryons mutants haploïdes engendrait un cycle mitotique supplémentaire avant la TMB, qui se produit alors au cycle 15 plutôt que 14. À partir de ces observations, les auteurs ont suggéré que la titration de facteurs maternels par la quantité d'ADN en constante augmentation pourrait expliquer un tel phénomène. Pritchard et Schubiger (1996) (69) ont montré que la répression de *fushi tarazu* (*ftz*) dépend de la quantité du répresseur maternel Tramtrack (*Ttk*), elle-même dépendante du ratio N/C. Ils ont également suggéré qu'il existait d'autres répresseurs maternels titrés expliquant la répression/activation d'autres gènes (comme *kriippel* qui n'est pas régulé par *Ttk* mais semble être soumis à ce ratio). Lu et collaborateurs (2009) sont partis de ces observations et ont mis en place une étude à grande échelle, basée sur l'utilisation de puces à ADN, pour détecter les gènes dont l'activation dépend du ratio NC. Pour cela, ils ont comparé l'abondance des ARNm dans des embryons sauvages et des embryons haploïdes durant la cellularisation qui se produit respectivement entre les cycles 13-14 et 14-15. Ils se sont particulièrement intéressés aux ARNm ayant une composante exclusivement zygotique (en partant des gènes purement zygotiques rapportés par (59)) et ont ainsi identifié 88 gènes purement zygotiques dont l'activation dépend du ratio NC. Cependant, Lu et collaborateurs n'ont pas analysé la présence d'éléments ou modules cis-régulateurs pour expliquer la régulation de ces gènes. Une grande majorité des gènes exprimés lors de l'AZG ne répond pas au ratio NC : au contraire, leur activation dépend du temps absolu après l'activation de l'oeuf et/ou après la fécondation, indépendamment du caractère diploïde ou haploïde de l'embryon. Ceci correspond au modèle impliquant l'horloge maternelle, dont *Smaug* semble être un des principaux acteurs (52, 63). Benoit et collaborateurs ont suggéré que *Smaug* est responsable de la dégradation de répresseurs de l'activation zygotique, qui corrèle avec l'augmentation de la transcription pendant la TMZ, indépendamment des points de contrôle de la réplication, permettant par exemple l'accumulation de la forme phosphorylée de l'ARN polymérase II.

Enfin, la rapidité des cycles cellulaires ne permettrait pas de terminer la transcription des ARNm. L'application d'inhibiteurs du cycle cellulaire résulte en une activation prématurée de la vague majeure d'AGZ. De plus, il a été démontré qu'effectivement la progression du cycle

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

mitotique arrête les transcriptions en cours (50). De Renzis et al. 2007 ont montré que les premiers transcrits zygotiques synthétisés lors de la première vague (cycles ralentis mais encore rapides) sont relativement courts et dépourvus d'introns. L'arrêt en interphase du cycle 14 est donc nécessaire pour la transcription d'ARNm plus long. Le ralentissement des cycles mitotiques semble indispensable mais pas suffisant pour expliquer l'activation de la transcription. En effet, lorsque les cycles sont bloqués artificiellement, l'AGZ ne peut être initiée avant le cycle 10 (68). Ceci indique que d'autres mécanismes sont impliqués pour empêcher l'activation trop précoce du zygote. Ces trois modèles ne sont pas exclusifs et semblent se compléter. Particulièrement, les cycles mitotiques sont régulés à la fois par le ratio NC et l'horloge maternelle. Le ratio NC est effectivement impliqué dans la titration de la machinerie de réplication (70) et de la Cycline B (51). Ceci permettrait l'introduction de points de contrôle, assurés par les protéines Chk1 et Chk2, entre la fin de la phase S de synthèse d'ADN et le début des phases mitotiques à partir du cycle 8 (70) et induirait ainsi le ralentissement des cycles mitotiques. D'autre part, Smaug est en partie responsable de la dégradation de la Cycline B qui réprime Chk1 et Chk2. En fait la dégradation de certains activateurs du cycle cellulaire, comme nous venons de le voir, et de la transcription, comme Ttk, dépend des deux mécanismes.

– Les facteurs activateurs de l'AGZ

Comme nous l'avons vu précédemment, l'activation d'un nombre restreint de gènes impliqués dans la segmentation a pu être en partie expliquée par le recrutement du complexe médiateur contenant dMED31 (45). Les régions régulatrices impliquées dans l'activation des gènes de la segmentation ont été intensivement étudiées à grande échelle (ex. (24, 28)) ou de façon spécifique (ex. régions régulatrices d'*even-skipped* (71)). Il en va de même pour les gènes impliqués dans la spécification de l'axe dorso-ventral (ex. cibles de dorsal (72)). Cependant, ces régions activent les gènes d'une façon très localisée, et ne peuvent pas expliquer l'activation ubiquiste du grand nombre de gènes durant l'AGZ.

Plusieurs études ont mis en évidence l'action d'un facteur maternel, Zelda, dans l'activation des gènes durant la première vague de l'AGZ (59, 73, 74). Ce facteur collabore avec Dorsal (72), avec STAT (75) et semble être plus généralement un amplificateur de l'action des FT précoces (Hunchback, Bicoid, Krüppel, Giant, Caudal, Knirps) (42, 76).

1. INTRODUCTION

1.2.3 Apport des études transcriptomiques

J'ai cité précédemment plusieurs études qui ont permis d'identifier les gènes induits durant l'AGZ (58, 59, 62). Ces études ont été menées à large échelle au moyen de puces à ADN qui permettent d'étudier les profils de transcription de tous les gènes simultanément dans différentes conditions. Ce genre d'étude permet d'évaluer des profils temporels d'expression à différents stades de développement, ou dans différents contextes génétiques (pour étudier par exemple l'effet de mutations de protéines régulatrices conduisant à une expression différentielle de gènes cibles directs ou indirects). Ce type d'expérience est particulièrement utilisé pour sélectionner des groupes de gènes coexprimés afin de découvrir des éléments fonctionnels partagés par ces gènes. Il existe plusieurs méthodes pour définir de tels clusters qui seront décrites et discutées dans la section 2.1 du manuscrit. J'ai déjà décrit brièvement l'étude menée par Lu et collaborateurs (2009) dans le sous chapitre précédent. Brièvement, ils ont analysé les profils d'expression temporels dans des embryons sauvages et haploïdes afin de détecter les gènes dont l'activation est soumise au ratio NC. L'étude menée par De Renzis et al. (2007) a consisté à analyser l'expression des gènes durant différentes classes temporelles entourant l'AGZ dans des embryons sauvages et des embryons dont un chromosome (ou d'une partie de chromosome) a été retiré. Les résultats de ces expériences ont permis de détecter la contribution maternelle et zygotique de chaque espèce de transcrits et d'effectuer classement temporel précis des gènes par rapport au deux vagues de l'AGZ. Ils ont également mis en évidence la surreprésentation spécifique du motif de liaison de Zelda dans les régions promotrices des gènes activés dans le blastoderme pré-cellulaire (59). Enfin, Pilot et collaborateurs (2006) ont produit une série temporelle transcriptomique afin de détecter les gènes impliqués dans la cellularisation. L'ensemble de ses études sera décrit plus particulièrement dans la partie 2.

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

1.2.4 Objectif de la thèse

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

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

1. INTRODUCTION

2

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

Le projet trouve son origine dans les travaux de Fanny Pilot (58), qui a produit des données transcriptomiques à partir d'embryons précoces de drosophile minutieusement échantillonnés à la loupe binoculaire et répartis en cinq classes ou stades du développement (figure 2.1B). Elle a ainsi extrait les ARNm pour cinq classes temporelles. La première classe permet d'analyser le contenu en ARNm de l'embryon avant l'AGZ, ces ARNm correspondent aux ARNm maternels représentés en rouge dans la figure 2.1A. Les classes temporelles suivantes représentent le contenu en ARNm après l'activation majeure du génome zygote, ces ARNm correspondent aux transcrits nouvellement synthétisés à partir du génome zygotique (représentés en bleu foncé dans la figure 2.1A). Le but de cette étude était d'identifier des gènes variant de façon similaire entre classes temporelles consécutives. J'ai tout d'abord utilisé des approches « classiques » de clustering sur les profils temporels d'expression des gènes. Ces analyses ne m'ont pas permis d'extraire des groupes de gènes bien définis, me laissant juger arbitrairement de la définition de chaque cluster, et ne permettant pas d'évaluer directement les variations d'expression entre classes temporelles consécutives. J'ai alors décidé de développer une approche adaptée en discrétisant, sur base de critères statistiques robustes, les profils de transition. Cette méthode m'a permise d'obtenir des clusters compacts et cohérents biologiquement. Durant mon doctorat, Lu et collaborateurs (62) ont analysé l'impact du ratio nucléo-cytoplasmique (NC) sur l'AGZ en étudiant les profils temporels d'expression des gènes dans des embryons diploïdes sauvages et

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

des embryons haploïdes mutants (figure 2.1D). Leur étude ne se basant que sur un sous groupe de gènes, j'ai décidé de traiter ces données en adaptant la méthode de discrétisation développée précédemment. J'ai ainsi pu identifier les gènes dont l'activation suivait le modèle impliquant ratio NC ou l'horloge maternelle. Un troisième groupe de gènes dont l'activation ne suit aucun des deux modèles connus à également été identifié. Enfin, De Renzis et collaborateurs (59) ont étudié la contribution des ARNm maternels et zygotiques dans l'abondance des transcrits observée dans l'embryon au moment des deux vagues de l'AGZ (figure 2.1C). Pour cela ils ont utilisé des données temporelles couplées à des expériences de délétions chromosomiques. La méthode proposée précédemment n'étant pas adaptée à ce schéma expérimental, j'ai utilisé les clusters qu'ils avaient définis.

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

2.1.1 Normalisation

La première étape dans le traitement de données issues des expériences de puces est la normalisation des données. Les puces Affymetrix fournissent pour chaque sonde un couple de valeurs PM ("*Perfect Match*") et MM ("*MisMatch*"). Les valeurs MM permettent d'estimer le bruit de fond dû à des liaisons non spécifiques. Dans les puces de type Affymetrix, chaque gène est représenté par huit sondes. Plusieurs algorithmes de traitement existent, dont les différences reposent principalement sur la façon d'identifier les hybridations non spécifiques. La méthode de normalisation MAS5.0 (77), incluse dans les logiciels fournis par la firme Affymetrix, repose sur la soustraction des valeurs MM aux valeurs PM, en adoptant un traitement spécifique pour les cas où les MM ont des valeurs plus élevées que les PM (gènes considérés comme « absents », plus précisément non-détectés). L'utilisation des valeurs PM-MM pour l'analyse de l'expression différentielle des gènes produit beaucoup de faux positifs particulièrement pour des échantillons contenant une faible concentration de molécules et pour les valeurs faibles (78). Des labels de détection (absent : 'A', présent : 'P' et marginal : 'M') ont été calculés (79) à partir des valeurs PM/MM permettant d'écarter les sondes non fiables. En effet, le retrait de ces valeurs diminue sensiblement le taux de faux positifs (FP) mais conduit à une augmentation des faux négatifs (FN) (80)¹. Ces résultats montrent que les étiquettes « A »

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

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

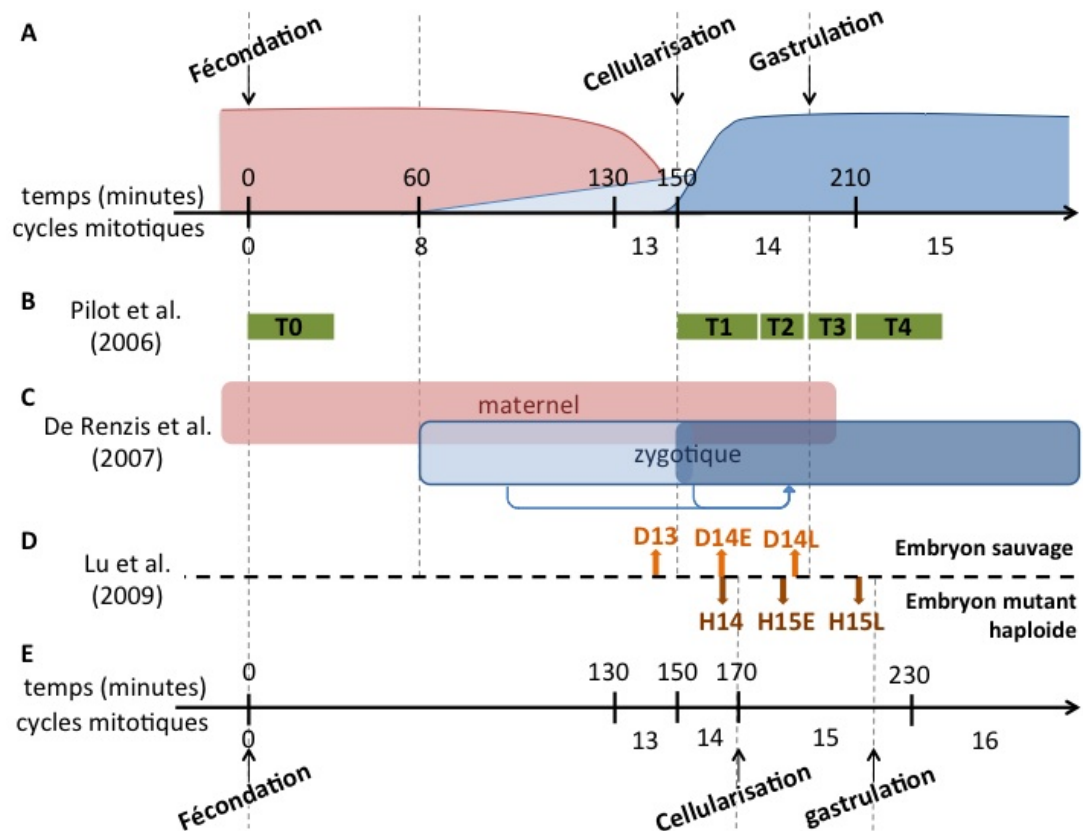


FIGURE 2.1: Répartition temporelle des données transcriptomiques utilisées pour l'étude de l'AGZ. - A. Quantité des ARNm maternels (profil rouge) ; ARNm zygotiques synthétisés durant la première et la seconde vague d'AGZ (profil bleu clair et foncé, respectivement). B. Intervalles temporels étudiés par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

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

sont ambiguës et ne permettent pas de déterminer si la faible hybridation est due à un artefact technique (mauvaise d'hybridation) ou à l'absence réelle du transcrit. D'autres méthodes ont été proposées, en particulier les méthodes RMA (81) et gcRMA (82), où les valeurs MM ne sont pas utilisées et où les PM sont corrigées par le bruit de fond ajusté et transformées en logarithme base 2, alors que l'ensemble des puces¹ est normalisé par les quantiles afin d'assigner à chaque puces une distribution de valeurs identique à une distribution « étalon » calculée sur l'ensemble des puces. Les méthodes RMA rapporte un faible taux de FP (78, 80) mais l'association des labels de détection à la méthode MAS5.0 réduit le nombre de FP et donne même de meilleurs résultats que RMA sur ce critère. Cependant le taux de FN est sensiblement le même dans les deux approches. Harr et collaborateurs (83) ont comparé ces différentes approches en considérant les quatre étapes de normalisation de puces à ADN Affymetrix : correction du bruit de fond, normalisation entre puces, correction PM-MM, obtention d'une valeur unique d'expression par jeu de sondes. Ils ont conclu que les méthodes RMA/gcRMA étaient les meilleures méthodes de normalisation pour la détection de gènes différentiellement exprimés. J'ai finalement choisi d'utiliser la méthode RMA qui semble mieux correspondre à la détection de gènes différentiellement exprimés. De plus, la méthode de labélisation appliquée à MAS5.0 identifie des sondes comme absentes ou « non fiables » pour palier le défaut de MAS5.0 alors que l'interprétation de ce label est ambiguë. Pilot et collaborateurs (58) avait utilisé les labels de détection MAS5.0 et d'autres critères pour identifier les gènes différentiellement exprimés puis déterminer les gènes co-exprimés. L'analyse des clusters obtenus par Pilot (annexe 6.1) a montré les limites des labels de détection. J'ai préféré baser les analyses suivantes sur un jeu de données complet et robuste et appliquer par la suite des filtres basés sur des tests statistiques disponibles dans les packages R (84).

2.1.2 Sélection de gènes différentiellement exprimés

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

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

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

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

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

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

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

où \tilde{m} est médiane des valeurs de transition au temps j pour tous les gènes. Nous pouvons noter que j'ai utilisé les estimateurs robustes de la tendance centrale (médiane) et de la dispersion (IQR standardisé¹) ce qui permet d'être moins sensible aux valeurs extrêmes. J'ai ensuite calculé, pour chaque gène i et à chaque transition j , la p-valeur nominale $Pval_{i,j}$ basée sur la distribution normale ajustée. La p-valeur a été corrigée pour les multi-tests en multipliant la p-valeur par le nombre de gènes G .

$$Eval_{i,j} = Pval_{i,j} * G \quad (2.3)$$

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

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

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

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

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

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

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

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

2.1.3.1 Approches classiques de clustering

Afin de regrouper les gènes présentant des profils d'expression similaires, j'ai testé les principales approches pour l'identification de groupes de gènes co-exprimés. Pour cela, j'ai utilisé l'outil MeV (85) (Multiexperiments Viewer) qui propose une série d'algorithmes pour le clustering, la visualisation et l'analyse statistiques de données transcriptomiques.

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

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

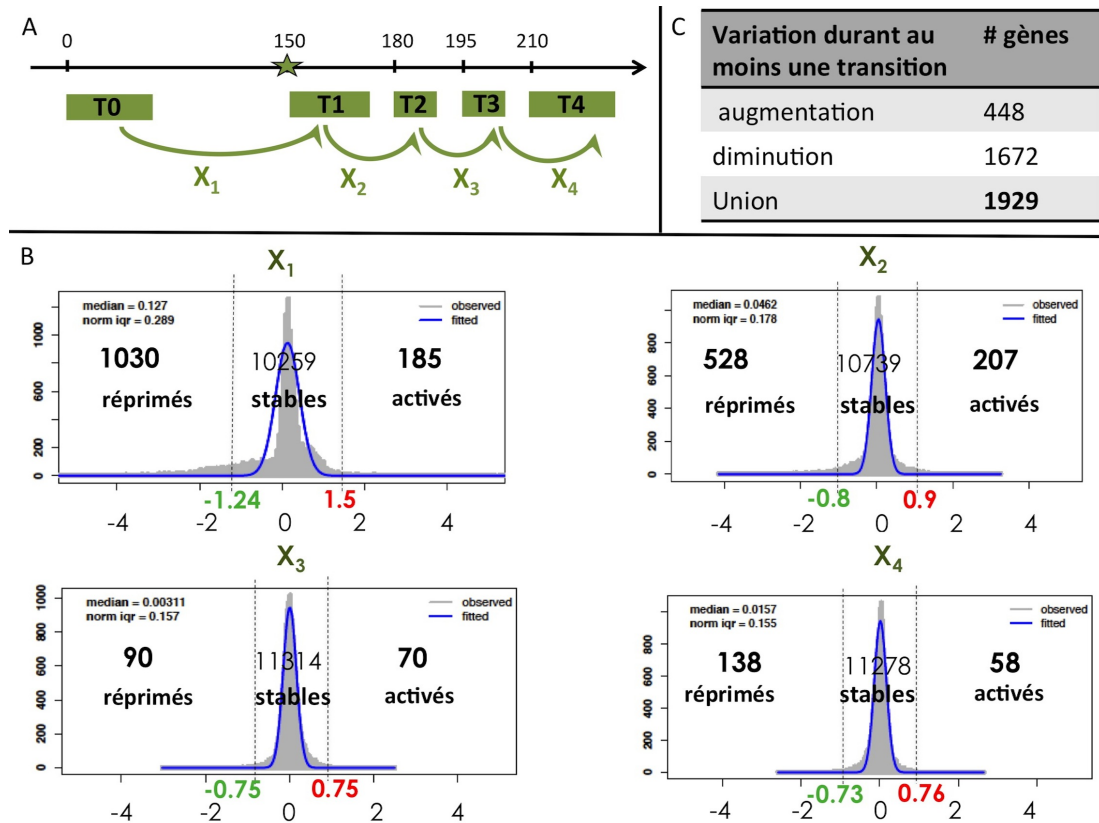


FIGURE 2.2: Distribution des \log_2 ratio pour chaque transition entre classes temporelles consécutives. - A. Représentation du schéma temporel des expériences de Pilot (58). L'axe horizontal représente le temps en minutes après la fécondation. L'étoile indique la cellularisation. B. Pour chaque distribution l'abscisse représente la valeur des transition $X_{i,j}$ (\log_2 ratio entre deux classes temporelles consécutives), l'ordonnée représente le nombre de gènes, les lignes pointillées représentent les seuils négatifs (en vert) et positifs (en rouge) calculés à partir de l'e-value maximale à 0,01. La courbe normale ajustée à la distribution des \log_2 ratio est indiquée en bleu. C. Résumé du filtrage.

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

au produit scalaire moyen qui permet de différencier la corrélation (caractérisée par des signes positifs) de l'anti corrélation (signes négatifs), ce que ne permet pas la distance euclidienne. Le produit scalaire moyen prend également en compte l'amplitude des variations, ce qui permet de différencier entre les profils « plats » des profils présentant de fortes variations (ce que ne permet pas la corrélation de Pearson). Visuellement, le dot product donnait des regroupement au moins aussi satisfaisants que les autres métriques, en terme de cohérence des groupes. Cependant, le problème qui restait, était dans le choix du niveau de l'arbre à couper pour obtenir des clusters cohérents. Ce choix est assez arbitraire et délicat. La figure 2.3A montre un exemple de choix de cluster visuellement intéressant. Si l'on regarde le profil temporel médian, il y a effectivement une tendance d'augmentation du signal tout au long des points temporels. Cependant les profils de transition individuels apparaissent peu cohérents et informatifs puisque, par exemple, la transition $X_{.1}$ regroupe des valeurs de log ratio allant de - 1.5 à 3, ce qui correspond à un continuum entre la diminution et l'augmentation significative du signal à $X_{.1}$. En ce qui concerne la méthode des k-means, le choix préalable du nombre de clusters influence beaucoup les résultats. La figure 2.3B montre un cluster résultant d'un test avec la méthode k-means. L'observation de la heatmap donne déjà une indication sur l'hétérogénéité de la composition du cluster. Cette représentation est trompeuse et parfois peu intuitive car les passages entre noir-rouge et vert-noir, par exemple, correspondent tous deux à une augmentation de l'intensité entre deux points temporels consécutifs. Les profils de transition regroupés par k-means manifestent donc, comme précédemment, un manque de cohérence par rapport à la question posée, à savoir regrouper les gènes dont l'intensité varie de la même façon tout au long du temps.

2.1.3.2 Clustering sur base de la discrétisation des profils de transition

Sur la base de ces constats, j'ai décidé de travailler directement avec les profils de transition qui permettent une meilleure représentation des variations de signaux entre points temporels consécutifs. J'ai décidé de me servir du calcul de significativité opéré sur les log ratios durant l'étape de filtrage pour la sélection des gènes différenciellement exprimés. Pour chaque gène et chaque transition, j'ai discrétisé les z-scores $Z_{i,j}$ (équation 2.2) en considérant le seuil $\Theta_{0,01}$ défini par la e-valeur à 0,01 (figure 2.4A). L'échelle discrétisée admet trois valeurs possibles= u ("up-regulated"), d ("down-regulated") et s ("stable"), définies comme suit :

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

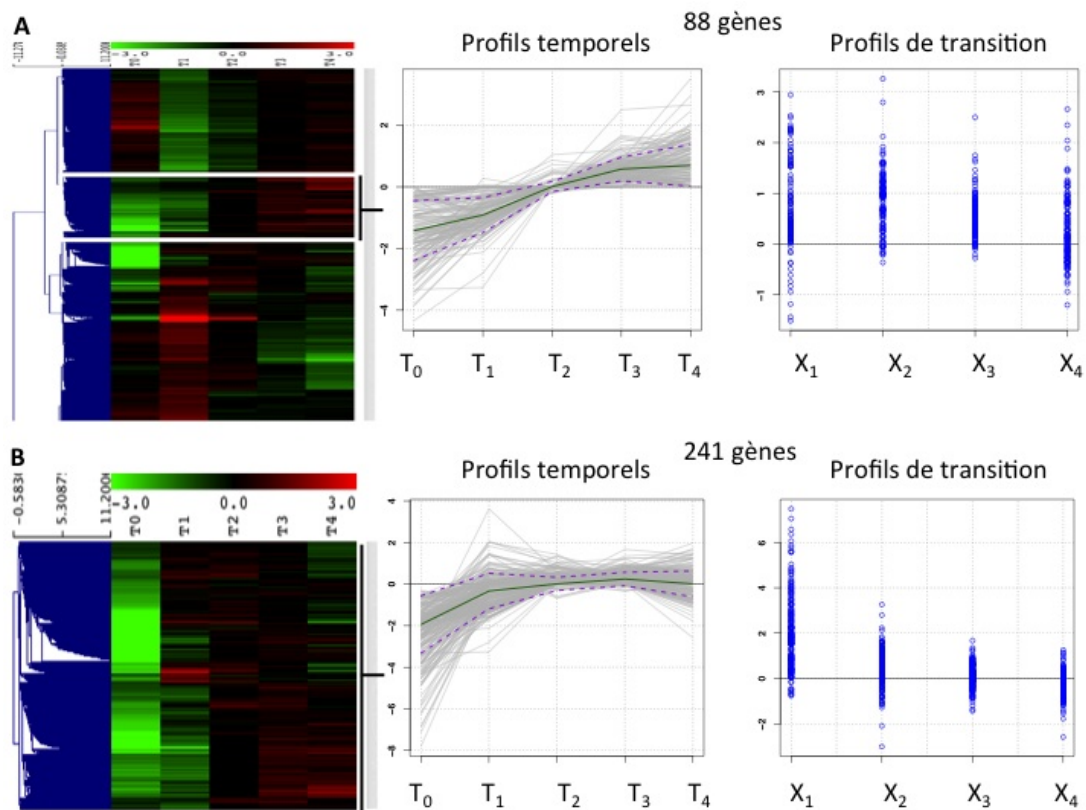


FIGURE 2.3: Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression. - À gauche : heatmap des profils temporels entre T_0 et T_4 , vert, noir et rouge : intensité inférieure, égale ou supérieure à l'intensité médiane du gène au travers de tous les points temporels, respectivement. Au milieu : Profils temporels, où l'ordonnée indique la valeur d'intensité (\log_2) pour chaque gène (chaque courbe grise correspond à un gène) standardisée par l'intensité médiane du gène dans l'ensemble des classes temporelles. La ligne verte représente le profil médian et les lignes pointillées représentent la 1 écart-type par rapport à la médiane. À droite : Profils de transition (X_{i1} à X_{i4}) entre points temporels consécutifs. L'ordonnée indique la valeur du log ratio et chaque point représente un gène. A. regroupement par clustering hiérarchique, distance : produit scalaire moyen, règle d'agglomération : complète, la heatmap ne représente qu'une partie des données. B. K-means, un des dix clusters obtenus après 50 itérations.

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

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

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

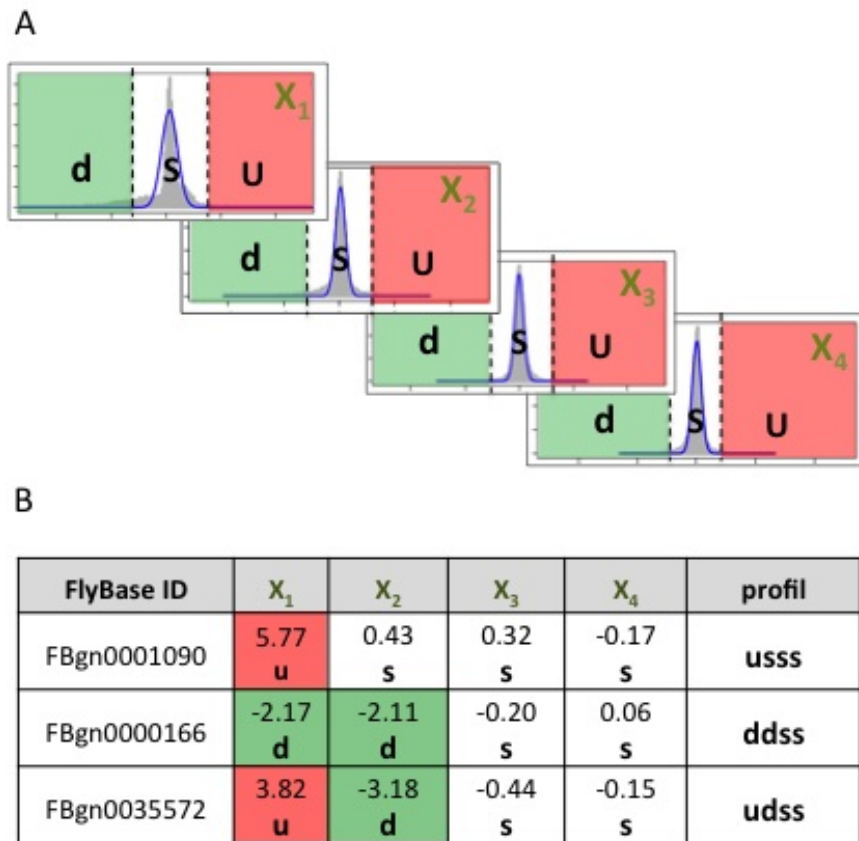


FIGURE 2.4: Méthode de discrétisation des profils de transition. - A. Distribution des log-ratios pour chaque transition ($X_{i1} - X_{i4}$) et indication de l'assignation des labels selon la significativité de la variation. Les lignes pointillées correspondent au seuil $\Theta_{0.01}$ comme indiqué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

Pour des profils comportant un nombre restreint de points temporels, la discrétisation des valeurs de transition définit un nombre raisonnable de profils possibles. Par exemple, les données de Pilot comportent 5 classes temporelles (T_0, \dots, T_4) convertis en 4 transitions (X_{i1}, \dots

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

X_{i4}). Chaque gène i est donc représenté par un vecteur de 4 valeurs discrètes (D_{i1} , D_{i2} , D_{i3} , D_{i4}) pouvant prendre trois valeurs (d, s, u). On peut donc a priori définir 3^4 profils discrets de transition distincts. Parmi ces 81 clusters, 46 profils ont été prédits.

Ce faible nombre de profils permet donc d'effectuer le clustering par simple regroupement des gènes ayant des profils discrets identiques. La figure 2.5A montre les profils de transition obtenus, qui s'avèrent particulièrement facile à interpréter de façon intuitive (chaque couleur peut être associée à un type de variation : vert pour d ; rouge pour u et noir pour s (figure 2.5 B)). Par ailleurs, les profils temporels sont remarquablement compacts et cohérents (figure 2.5 C-E). À titre de comparaison, j'ai testé les méthodes classiques de clustering sur les profils temporels (annexe 6.3). Les clusters sont beaucoup moins compacts et sont plus difficiles à définir.

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

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

En 2009, Lu et collaborateurs (62) ont étudié l'impact du ratio NC sur l'activation des gènes au moment de la deuxième vague de l'AGZ (figure 2.1D). Ils ont procédé à un clustering hiérarchique sur un sous-ensemble de données d'expression (Affymetrix) pour les gènes purement zygotiques (59). Ils ont ainsi identifié 88 gènes dont l'activation dépend du ratio NC. Afin de retirer le maximum d'information de ce jeu de données, j'ai décidé de retraiter la totalité des données en appliquant le protocole défini précédemment : normalisation RMA, sélection des gènes différentiellement exprimés sur base d'un seuil de significativité par échantillon (chip-wise) et groupement des gènes co-exprimés sur base des profils de transition discrets. La particularité des données (deux génotypes différents) et des interprétations possibles (implication ou non du ratio NC) m'ont amenée à modifier légèrement la procédure. J'ai donc appliqué un autre filtre en considérant la cohérence entre les profils de transition de deux génotypes sur base des deux modèles alternatifs de régulation du niveau d'ARNm, à savoir le ratio NC ou l'horloge maternelle. En effet, dans un embryon haploïde, la cellularisation et la pause en interphase sont retardées d'un cycle mitotique, c'est à dire qu'elles se passent au cycle mitotique 15 au lieu de 14 dans l'embryon diploïde.

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

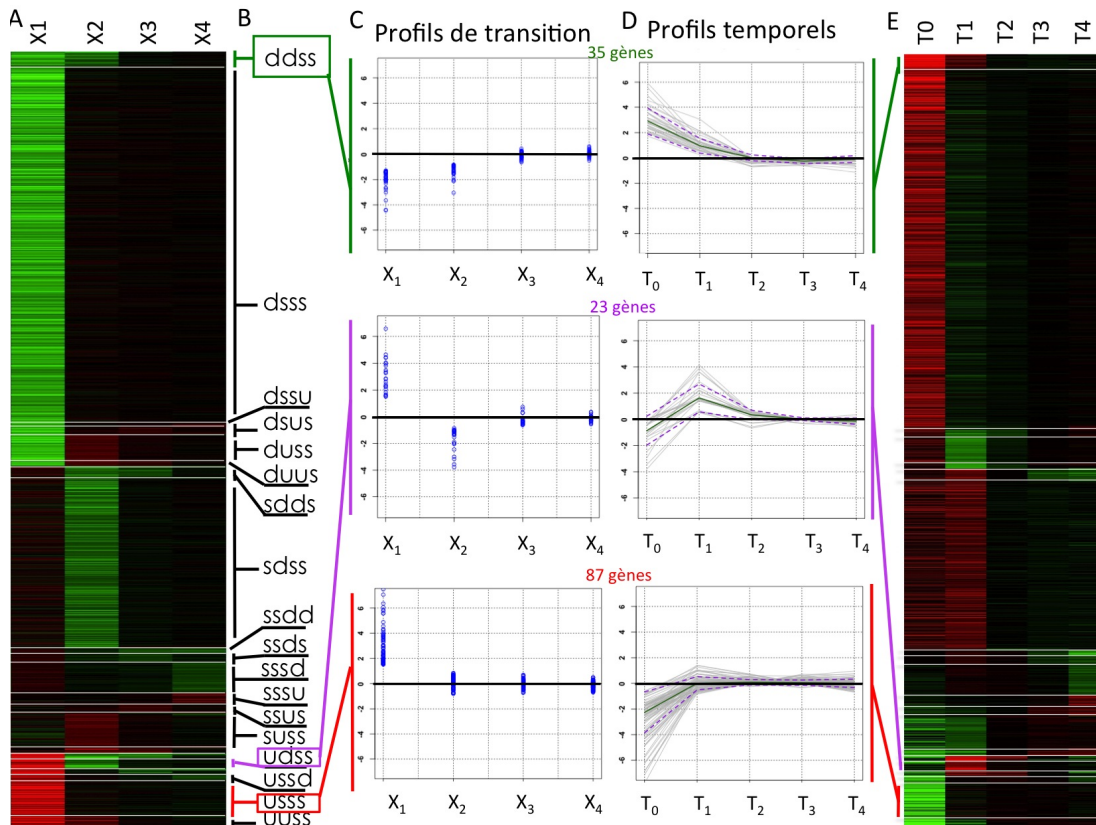


FIGURE 2.5: Clusters résultant du regroupement de profils discrets identiques. - A. Heatmap des profils de transition des gènes regroupés. Les couleurs verte, rouge et noire représentent respectivement les variations positive, négative ou négligeable durant la transition. B. Profils de transition discrets.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

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

TABLE 2.1: Récapitulatif des interprétations biologiques des profils discrets présentant au moins 10 membres dans les données de Pilot (2006).

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

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

Je vais tout d'abord définir la nomenclature utilisée pour les transitions. Ici, les classes temporelles sont définies par le génotype (D : diploïdes, H : haploïde) des embryons et le cycle mitotique pendant lequel les embryons ont été extraits. Le cycle mitotique correspondant à la cellularisation (cycle 14 et 15 chez les diploïdes et haploïdes respectivement) dure environ 1h, les auteurs ont extrait des embryons au début de ce cycle (E : "early") et à la fin (L : "late"). Ainsi, en suivant l'équation 2.2, la transition entre les classes temporelles consécutives D13 et D14E (embryons diploïdes au cycle 13 et au début du cycle 14) pour le gène i sera notée X_{D_i14E} et plus généralement pour l'ensemble des gènes X_{D14E} (figure 2.6A). À partir de ces transitions (diploïdes : X_{D14E}/X_{D14L} et haploïdes : X_{H15E}/X_{H15L}), j'ai obtenu 37 profils discrets sur les 81 possibles.

Pour les gènes régulés par le ratio NC, on s'attend à ce que les niveaux d'expression (et leurs transitions) du 14ème cycle cellulaire de l'embryon diploïde soit similaire à celui du 15ème cycle de l'embryon haploïde ($X_{D14E} = X_{H15E}$ et $X_{D14L} = X_{H15L}$). Au contraire, les gènes insensibles au ratio NC, régulés par l'horloge maternelle, varieront de la même façon quelque soit le génotype des embryons utilisés. Les transitions X_{D14L} et X_{H15E} correspondent au même temps absolu (figure 2.6A), on suppose alors : $X_{D14L} = X_{H15E}$. La figure 2.6B montre un exemple de gène pour chacun des cas. Seuls 24 de ces clusters comportent plus de 10 gènes. Parmi les 24 clusters contenant au moins dix gènes (tableau 2.2), 15 correspondent à l'un de ces deux modèles, dont trois clusters contiennent des transcrits dont la dégradation est dépendante du ratio NC, " $ds_D ds_H$ ", " $sd_D sd_H$ " et " $dd_D dd_H$ ". Les 9 clusters restants contiennent des nombres de gène non-négligeables (montrés dans l'annexe 6.5). Il serait donc étonnant que cela soit dû à des erreurs de manipulations expérimentales. Il est possible que ces gènes combinent une composante maternelle et une composante zygotique et que leur régulation soit dépendante différemment du ratio NC ou de l'horloge maternelle. Pour la suite de mes analyses, j'ai décidé d'écarter ces gènes en raison de ces ambiguïtés. Le groupes de gènes retenus sont néanmoins plus larges et diversifiés que celui des gènes purement zygotiques analysés par Lu et collaborateurs.

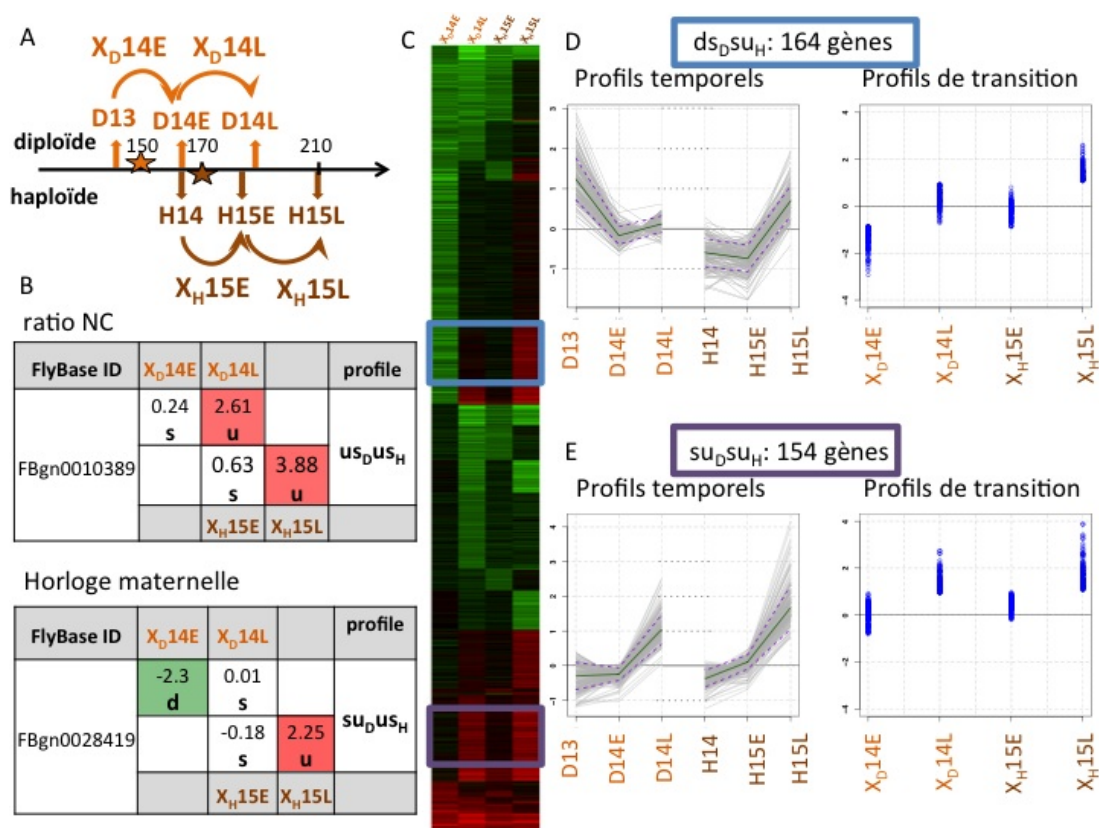


FIGURE 2.6: Application et résultats de la discrétisation des profils de transitions obtenus à partir de des données de Lu et collaborateurs. - A. Schéma expérimental. L'axe horizontal représente le temps en minutes après la fécondation, les étoiles correspondent au moment de la cellularisation observé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

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

D'autres études fournissent des données transcriptomique à grande échelle concernant l'embryogenèse précoce de la drosophile. De Renzis et collaborateurs (59) ont procédé à l'analyse du transcriptome dans des embryons sauvages pour trois classes temporelles (entre 0 et 1h : ARNm maternels, entre 1 et 2h : première vague de l'AGZ ; entre 2 et 3h : deuxième vague de

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

TABLE 2.2: Récapitulatif des interprétations biologiques des profils discrets présentant au moins 10 membres dans les données de Lu (2009).

profil	modèle	nombre de gènes	Interprétation biologique
$dd_D dd_H$	NC	41	ARNm maternels dégradés durant la cellularisation
$dd_D ds_H$	horloge maternelle	165	
$ds_D ds_H$	NC	37	ARNm maternels dégradés durant la phase lente de cellularisation
$ds_D ss_H$	horloge maternelle	406	
$ds_D su_H$	horloge maternelle	163	ARNm maternels dégradés durant la phase lente de cellularisation puis ARNm zygotiques synthétisés pendant la phase précoce de la gastrulation
$ds_D du_H$?	19	?
$dd_D ss_H$?	104	
$du_D su_H$?	48	
$sd_D sd_H$	NC	91	ARNm maternels dégradés à partir de la phase rapide de cellularisation
$sd_D dd_H$	horloge maternelle	61	
$sd_D ds_H$	horloge maternelle	97	
$ss_D sd_H$	horloge maternelle	111	ARNm maternels dégradés durant la phase précoce de la gastrulation
$sd_D ss_H$?	213	?
$ss_D ds_H$?	58	
$ss_D su_H$	horloge maternelle	164	ARNm zygotiques synthétisés pendant la phase précoce de la gastrulation
$su_D su_H$	NC	154	ARNm zygotiques synthétisés à partir de la phase rapide de la cellularisation
$su_D uu_H$	horloge maternelle	47	
$ss_D us_H$?	12	?
$su_D ss_H$?	37	
$us_D us_H$	NC	14	ARNm zygotiques synthétisés dès le début de la cellularisation
$us_D ss_H$	horloge maternelle	24	
$uu_D uu_H$	NC	60	
$uu_D us_H$	horloge maternelle	27	
$uu_D ss_H$?	13	?

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

2.4 Analyse de la composition des clusters

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

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

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

2.5 Conclusion du chapitre

L'analyse approfondie de séries temporelles provenant de données de puces m'a menée à établir une méthode simple adaptée à ce type de données puisqu'elle permet d'apprécier directement les variations de signal entre points temporels consécutifs. De plus, elle permet de regrouper des gènes suivant des profils de transition facilement interprétables. Ainsi 42 clusters contenant plus de 10 gènes ont été définis par cette méthode à partir des données de Pilot et al. (2006) et de Lu et al. (2009). La comparaison de la composition des 48 clusters (42 clusters, 1 cluster publié par Lu et al. et 5 clusters définis par De Renzis et al.) a montré que des clusters obtenus à partir de données différentes et présentant des profils de transition discrets comparables (par exemple *"dsss"* et *"ds_Dsss_H"* ou *"uuss"* et *"uu_Duu_H"* définis à partir des données de Pilot et Lu respectivement) avaient un nombre significatif de gènes en commun. Ceci suggère que la méthode de groupement basée sur les profils discrets est relativement fiable. Cependant, cette méthode n'est applicable que pour des séries temporelles relativement courtes. En effet le nombre de combinaisons possibles est exponentiel en fonction du nombre de transition (3^n).

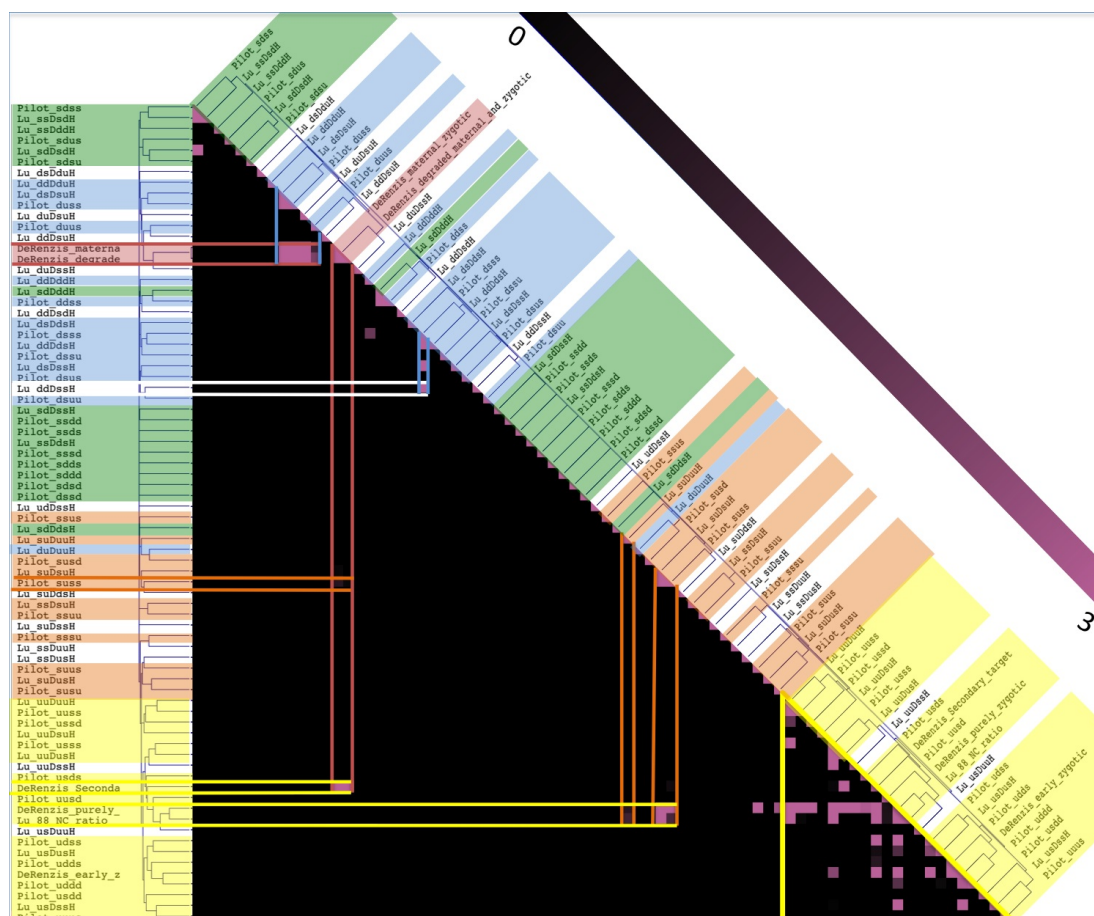


FIGURE 2.7: Comparaison des groupes de gènes co-exprimés en fonction de la significativité (binomiale corrigée pour les multi-tests) de leur chevauchement. - Les clusters de gènes induits avant ou au tout début de la cellularisation sont soulignés en jaune, ceux induit plus tardivement en orange, les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

2. IDENTIFICATION DE GÈNES CO-EXPRIMÉS PENDANT L'ACTIVATION TRANSCRIPTIONNELLE DU GÉNOME ZYGOTIQUE (AGZ)

3

Analyse fonctionnelle des gènes et des régions non-codantes associées

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

3.1 Analyse des clusters primaires

3.1.1 Enrichissement fonctionnel

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

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

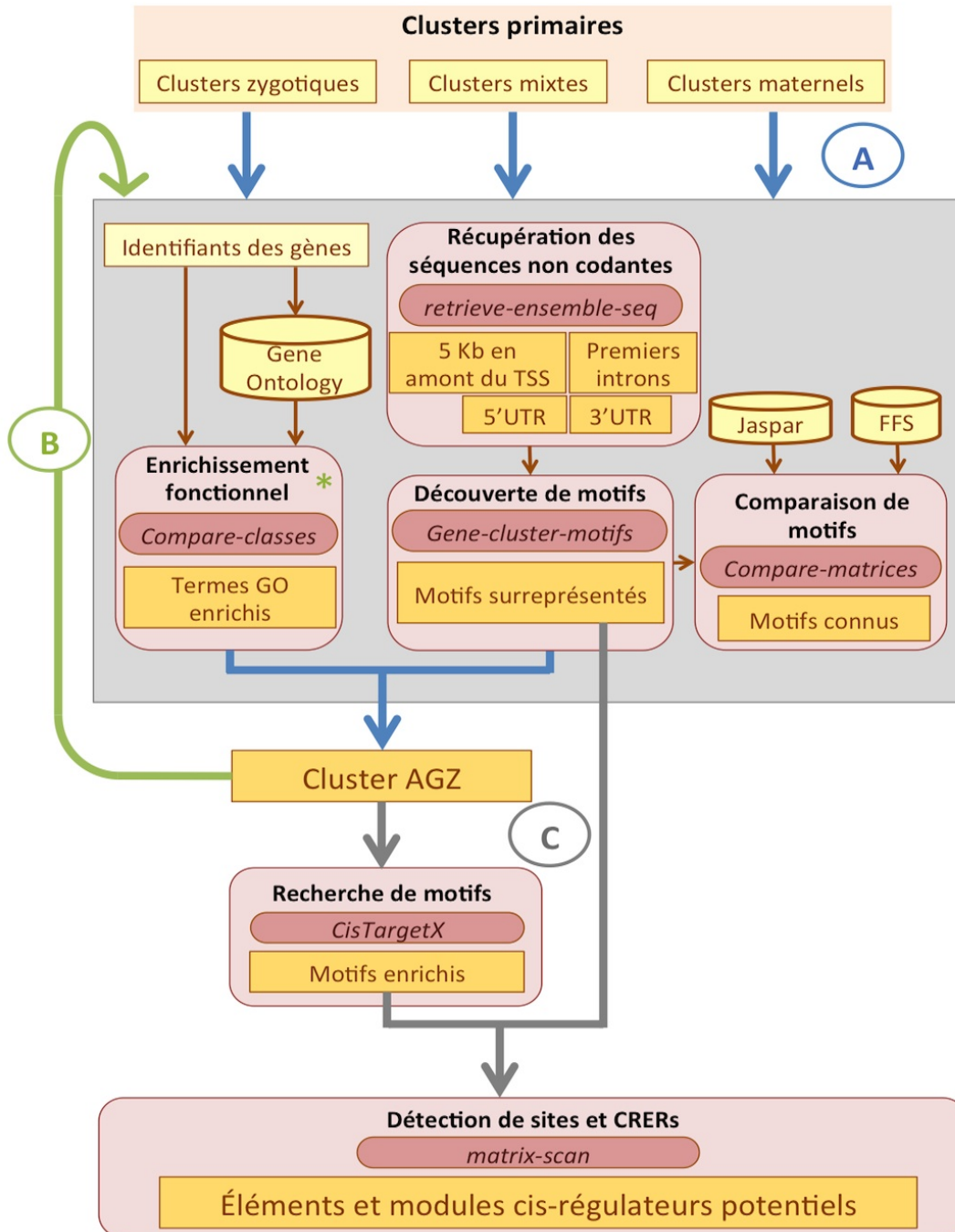


FIGURE 3.1: Schéma d'analyses fonctionnelles pour la prédiction d'éléments et modules cis-régulateurs potentiels impliqués dans la régulation de l'AGZ. - A. Analyses fonctionnelles des clusters primaires pour la définition du cluster AGZ. B. Analyses fonctionnelles du cluster AGZ. C. Complément d'analyse des éléments cis-régulateurs du cluster AGZ. L'encadré gris met en évidence les analyses qui ont été appliquées de façon itérative. L'étoile verte indique des outils qui n'ont été utilisés que lors de l'itération B. FFS : Fly Factor Survey.

l'outil *compare-classes* (86).

GO est une base de données contenant la description des gènes et de leurs produits dans un vocabulaire contrôlé commun à toutes les espèces. Les descriptions sont organisées selon trois classes : les processus biologiques dans lesquels les gènes sont impliqués, les fonctions moléculaires des produits géniques et les composants cellulaires où ils sont actifs. Les descriptions sont structurées sous forme de termes définis par un identifiant unique, un nom correspondant à la description, la classe, une définition avec ses sources citées et les termes avec lesquels il est en relation directe. En effet, la base GO est structurée sous forme de graphe orienté acyclique où chaque terme est en relation avec un ou plusieurs autres termes. Les termes présents aux classes les plus profondes de la hiérarchie correspondent aux descriptions les plus précises et vont en se généralisant jusqu'à la « racine » qui correspond au nom de la classe. Les gènes sont annotés à partir de différentes sources. Un code d'indications ("*evidences*" en anglais) est utilisé pour différencier annotation reposant sur les indications expérimentales, informatiques (inférences diverses), bibliographiques (annexe 6.1). L'explication détaillée de ce code est disponible à l'adresse <http://www.geneontology.org/GO.evidence.shtml>. J'ai décidé d'écartier les annotations dont la source n'est pas fiable ou peu informative, c'est à dire lorsque :

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

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

Sur les 48 clusters analysés, seuls 18 ont manifesté un enrichissement significatif (e-valeur < 0.1) dans au moins une classe GO ("biological process", "molecular function" ou "cellu-

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

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

3.1.2 Éléments cis-régulateurs

J'ai procédé à la découverte de motifs surreprésentés dans chacun des clusters et pour chaque type de séquences non-codantes (5kb en amont, premier intron, 5'UTR et 3'UTR, détails de l'obtention de séquences figure 3.7). Pour cela, j'ai utilisé le pipeline gene-cluster-motifs combinant différents programmes d'analyses de séquences régulatrices RSAT (87). Ce pipeline comprend une étape de découverte de motifs avec plusieurs outils dont *oligo-* et *dyad-analysis*, ainsi que la construction des matrices poids-positions à partir des mots ou dyades significativement surreprésentés à l'aide des algorithmes *pattern-assembly* et *matrixfrom-patterns* (les paramètres utilisés sont indiqués dans la partie 3.3.2). J'ai ensuite construit une matrice confrontant mots (oligos ou dyades) en confrontant tous les clusters avec tous les motifs découverts en fonction de la significativité de la surreprésentation de chaque motif (oligo-nucléotide ou dyade) dans chaque cluster. Parmi 48 clusters analysés (ayant plus de 10 membres), 33 possèdent au moins un motif surreprésenté de manière significative ($sig \geq 2$). Certains motifs sont surreprésentés de façon récurrente dans plusieurs clusters.

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

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

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

J'ai également calculé des modèles de background d'un vertébré (souris), et d'une bactérie (spirochaete).

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

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

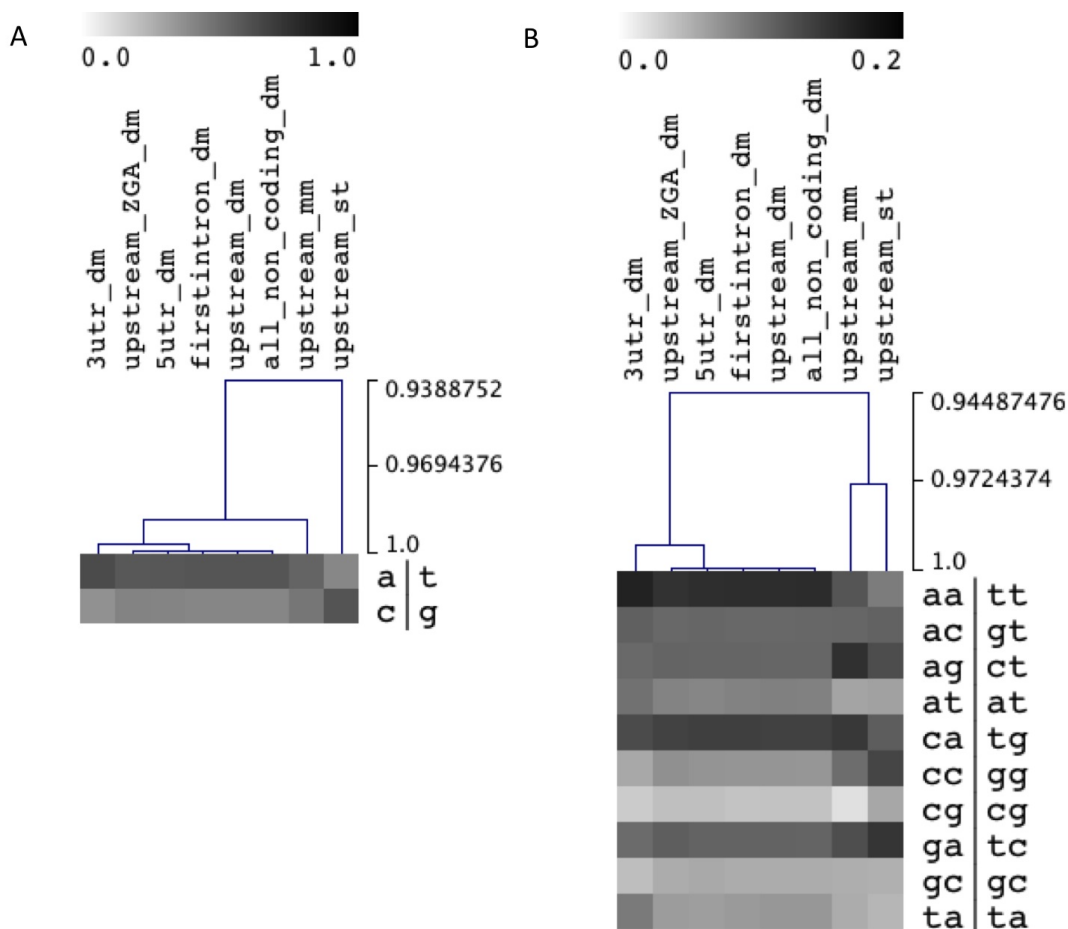


FIGURE 3.2: Composition en nucléotides de différents types de séquences. - dm : *Drosophila melanogaster*, mm : *Mus musculus*, st : *Spirochaeta thermophila*. Le clustering hiérarchique a été calculé par la méthode de Pearson non-centrée.

La figure 3.3 montre les oligomères de 7 nucléotides surreprésentés dans les séquences en amont du TSS en fonction du modèle de background. L'impact du background est flagrant lorsqu'il a été construit à partir d'espèces différentes, à partir des séquences d'entrée (modèle de Markov d'ordre 2) ou d'un jeu de référence externe (ensemble des 3'UTR). Pour les autres tests, même si l'ordre diffère légèrement, les significativités des mots sont du même ordre de grandeur (annexe 6.2). J'ai ainsi choisi d'utiliser le modèle spécifiquement calculé pour chaque type de séquences.

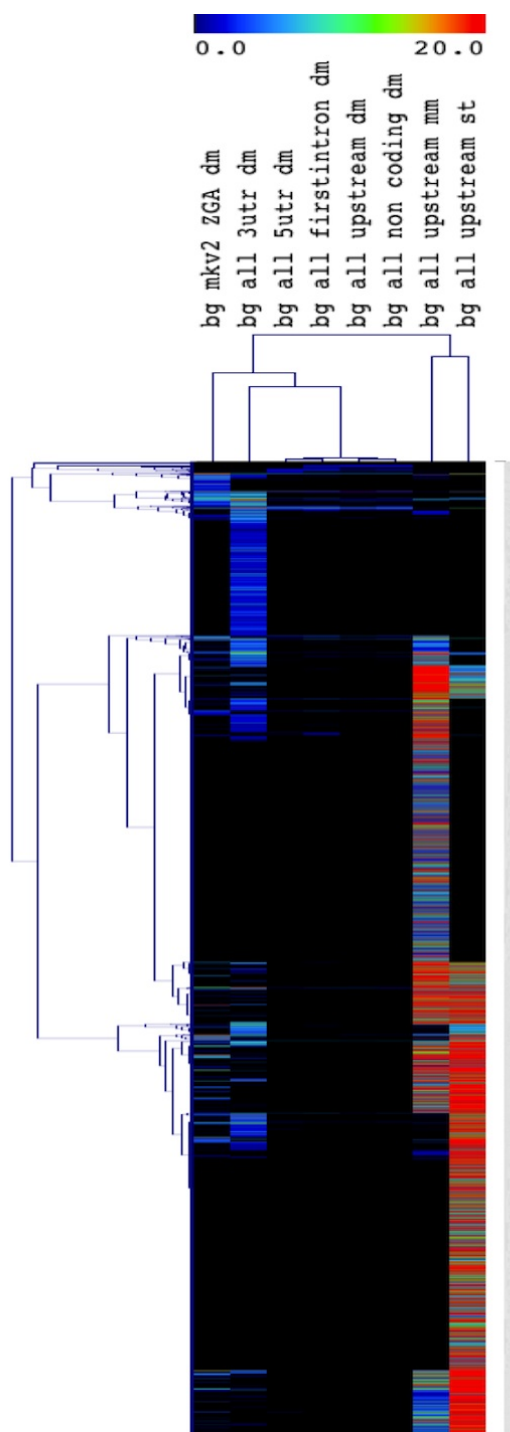


FIGURE 3.3: Heatmap représentant les résultats d'oligo-analysis sur les séquences en amont du TSS des gènes AGZ en utilisant divers modèles de background. - Les couleurs représentent la significativité des oligos (lignes). Chaque colonne correspond à un modèle différent. dm : *Drosophila melanogaster*,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

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

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

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

3.1.2.3 Classification des mots et dyades

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

3.1 Analyse des clusters primaires

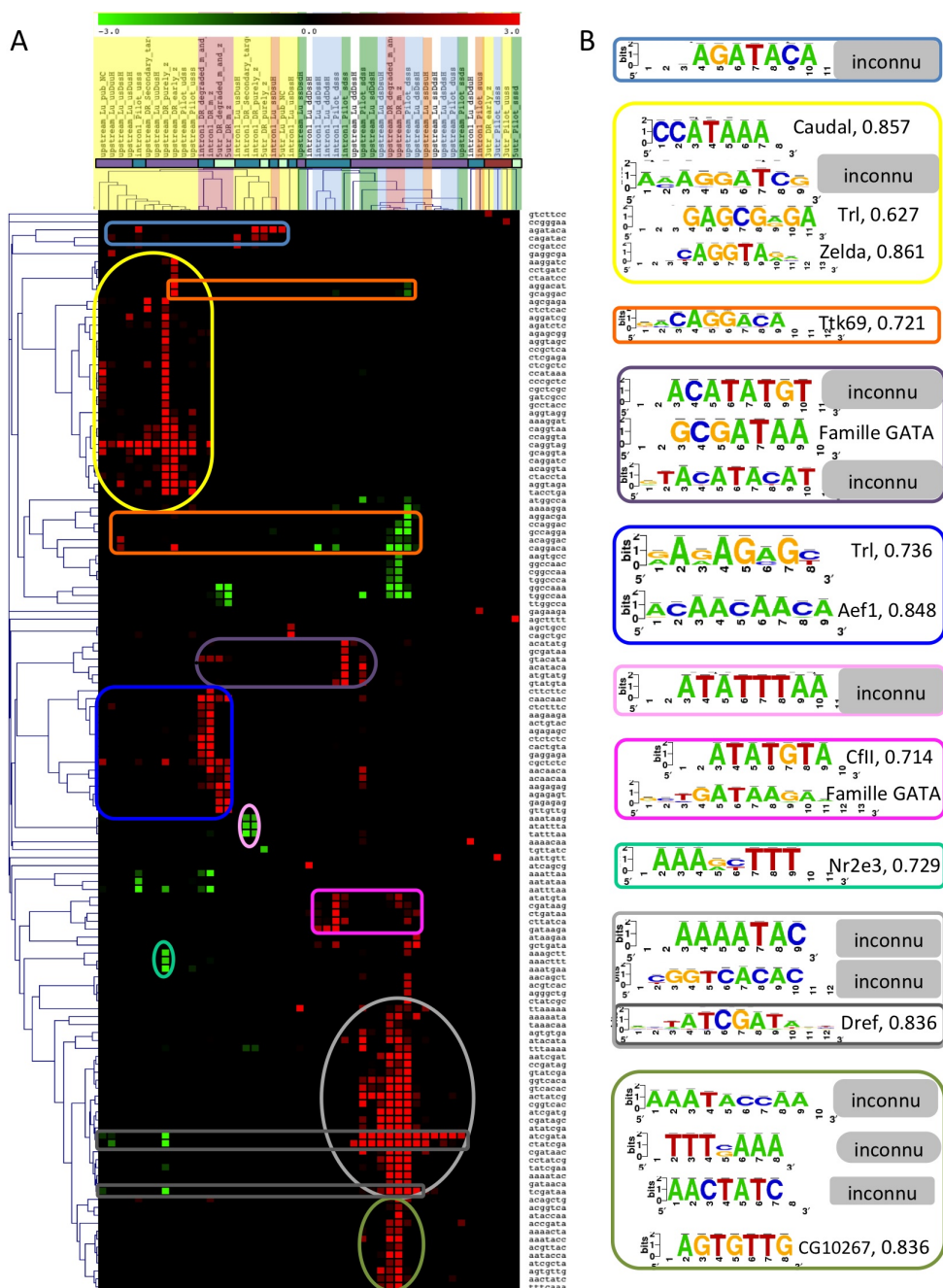


FIGURE 3.4: Clustering des oligonucléotides et des groupes de gènes en fonction de la significativité de la surreprésentation. - A. Heatmap représentant le clustering hiérarchique des clusters de co-expression (colonnes), plus exactement des différentes régions non codantes reliées aux gènes de ces clusters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

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

zygotiques, pour lesquels certains motifs issus de l'assemblage des mots chevauchants correspondent à des facteurs pertinents connus pour être actifs dans le blastoderme. En effet, nous retrouvons le motif de liaison de Zelda, un facteur impliqué dans l'activation zygotique mineure des gènes (59, 73, 74). En accord avec une étude qui vient d'être publiée (88), le motif de liaison de Zelda apparaît de façon générale dans les régions en amont des gènes tous les clusters zygotiques (et dans les premiers introns du cluster "*uss*") sans restriction aux gènes induits durant le stade pré-blastoderme cellulaire. Le mot 'CAGGTAG' (correspondant à une variante la plus utilisée du motif de liaison de Zelda) apparaît également dans les premiers introns du cluster mixte maternel-zygotique. Dans le cluster de gènes purement zygotiques, nous retrouvons également un motif proche de celui associé à Caudal, un facteur de transcription maternel connu pour son rôle dans la segmentation de l'embryon.

Deux autres motifs également surreprésentés dans ces clusters ne ressemblent à aucun motif répertorié dans les bases de données de référence (FlyFactorSurvey (31), JASPAR5 (30)). Le motif 'AGATACA' est particulièrement surreprésenté dans les premiers introns, ainsi que dans les régions en amont du cluster "*uu_Duu_H*". Le second motif ('AmAAGGATCG') est surreprésenté dans les clusters de gènes purement zygotiques et dans le cluster "*us_Dus_H*" suggérant une implication dans le contrôle de l'activation de la transcription dépendante du ratio NC.

Le motif de liaison de Trithorax-like (Trl ou facteur GAGA) est également surreprésenté dans les régions en amont des gènes purement zygotiques, cibles secondaires (de façon moins significative) et soumis au ratio NC selon Lu et collaborateurs. Le motif de liaison de Trl apparaît aussi dans un groupe de mots associés aux 5'UTR et premiers introns des gènes des clusters mixtes (3.4, groupe bleu foncé). Trl est un facteur maternel général agissant à divers niveaux de la régulation transcriptionnelle ((co-)facteur de transcription (89), remodelage de la chromatine (90, 91, 92)). L'implication de ce facteur semble être générale au niveau de l'AGZ, son rôle sera discuté plus loin dans le manuscrit.

Toujours au sein des 5'UTR et premiers introns des clusters mixtes, nous avons mis en évidence un motif ressemblant au motif de liaison d'Aef1. Ce facteur est un répresseur maternel de la transcription dont la quantité baisse entre 2 et 4h après la fécondation (Flybase, données de RNA-seq (93)). Ceci suggère que les ARNm synthétisés à partir du gène *aef1* sont dégradés. La protéine Aef1 pourrait être titré par la quantité croissante d'ADN au fur et à mesure des divisions mitotiques et tant que son niveau est important dans l'embryon, Aef1 pourrait donc contrecarrer l'action activatrice de Trl.

Un dernier motif correspondant au motif de liaison de Tramtrack (TTK) apparaît dans les séquences en amont de gènes du cluster "Lu *usDSSH*" et du cluster regroupant les gènes précoces définis par De Renzis et al. (59) (groupe orange). TTK est connu pour être un répresseur maternel dont la titration au cours des cycles mitotique permet l'activation des gènes zygotiques (69). De plus, il a été montré que TTK est capable d'inhiber l'activation médiée par Trl (dont le motif est également surreprésenté dans ce cluster) au niveau du promoteur de *even-skipped* (*eve*) par une interaction protéine-protéine directe (94). Ce motif est sous-représenté de façon significative (significativité de 5.01) dans les séquences en amont des gènes des clusters maternels et mixtes. Comme TTK est un facteur maternel, si les gènes exprimés au niveau maternels contenaient des sites de liaison de ce facteur, ils pourraient être réprimés et du coup ne pourraient pas assurer leur(s) fonction(s).

Dans les gènes des clusters mixtes, nous trouvons également un motif sous-représenté (significativité 3.98) dans les régions en amont et dans les 5'UTR. Ce motif (ctttGGCCaaa) ne correspond à aucun des motifs annotés dans les bases de données. En outre, ces séquences comportent trois motifs surreprésentés (groupe violet) dont deux sont inconnus ("ACATATGT", "TACATACAT"), alors que le troisième ressemble au motif de liaison de la famille de facteur GATA. Quatre facteurs de cette famille existent chez *Drosophila melanogaster* : Serpent, GATAe, Grain et Pannier. Aucun des transcrits synthétisés à partir de ces gènes n'est présent au niveau maternel : les facteurs Serpent, Grain et Pannier sont purement zygotiques (Serpent est en fait activé de façon secondaire par des facteurs zygotiques), alors que GATAe n'appartient à aucun cluster et aucun transcrit n'a été détecté dans les expériences de Gelbart et collaborateurs (93). La surreprésentation de ce motif dans les 5'UTR suggère une implication potentielle dans une activation zygotique plutôt que maternelle de ces gènes. Grain et/ou Pannier pourraient donc être impliqués dans l'activation des cibles secondaires appartenant aux clusters mixtes.

Enfin, les groupes rose, gris et kaki de la figure 3.4 sont associés aux séquences non codantes des clusters maternels et aux séquences en amont des gènes des clusters mixtes. Le groupe gris comporte trois motifs surreprésentés dans la majorité des clusters. Deux d'entre eux sont inconnus ("GGTCACAC" et "AAAATAC"). Le troisième correspond au motif de liaison de Dref appelé DRE (« DNA replication-related element ») et ayant comme consensus "TATCGATA". Ce motif est par ailleurs sous-représenté dans les séquences en amont des gènes purement zygotiques et des gènes appartenant au cluster "*uuDuuH*" dépendant du ratio NC. Dref est un facteur maternel impliqué dans l'activation de gènes impliqués dans la réplication.

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

Hirose et collaborateurs (95) ont montré que la protéine Dref s'accumule dans les noyaux à partir du cycle mitotique 8 (début de la formation du blastoderme syncytial) et ont suggéré que Dref permettrait la bonne coordination de l'activation zygotique de ces gènes. D'autre part, BEAF32 (insulateur) reconnaît également sur le motif "TATCGATA" (BEAF32 reconnaît seulement le mot "tatcgata" et aucun autre variant). Hart et collaborateurs (96) ont montré que Dref et BEAF32 pouvaient entrer en compétition pour la liaison à ce motif dans les cellules en prolifération rapide. Ces résultats suggèrent que l'action de Dref et/ou BEAF32 n'est pas restreinte aux gènes reliés à la réplication. La déplétion de ce motif dans les gènes du cluster "*uu_Duu_H*" et dans les gènes purement zygotique pourrait empêcher l'activation aberrante de ces gènes.

Le groupe rose associe des motifs présents dans les régions en amont, dans les premiers introns et 5'UTR des clusters maternels "Pilot *dsss*" et "Lu *ds_Dss_H*", ainsi que dans les séquences en amont (uniquement) des gènes des clusters mixtes. Le premier motif correspond au motif de liaison de CF2, qui est connu pour être actif durant l'oogenèse et pour participer à la polarisation dorso-ventrale⁶⁵. Le second motif correspond au motif de liaison des facteurs de transcription de la famille GATA. Comme vu précédemment, les facteurs de cette famille ne sont pas actifs dans l'oeuf. Ce motif pourrait être en fait une partie du motif DRE.

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

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

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

L'analyse précédente m'a permis de mettre en évidence le fait que les clusters zygotiques partagent plusieurs motifs, j'ai donc décidé de les regrouper pour les analyses suivantes. Quant

aux gènes maternel-zygotiques, ils sont régulés d'une façon particulière ayant des motifs surreprésentés dans leurs séquences en amont partagés avec les clusters maternels du type "*dxxx*" alors que les motifs surreprésentés dans les régions 5'UTR et premiers introns sont plutôt partagés avec les clusters zygotiques (motif de liaison de Trl et de Zelda). À cause de cette dualité, j'ai décidé de ne pas étudier les gènes maternel-zygotiques pour l'analyse de la régulation de l'AGZ.

Je me suis demandé si les gènes induits plus tardivement étaient régulés de la même façon que les gènes des clusters zygotiques. Dans l'analyse de surreprésentation de motifs, seul les gènes du cluster "*Lu ss_Dsu_H*" ont des motifs surreprésentés. Ce dernier regroupe les gènes actifs durant la phase tardive de cellularisation. Le clustering de ce groupe est le même que celui des clusters mixtes. Dans l'analyse des dyades, quatre clusters contiennent au moins un motif dont la significativité est ≥ 2 et leur groupement est mal défini puisque qu'ils font partie des branches non-résolues de l'arbre obtenu par le clustering hiérarchique (annexe 6.6A). Ces clusters n'ont pas de motifs en commun avec les gènes des clusters zygotiques, ils ont en fait que très peu de motifs surreprésentés. Ces gènes ne seront donc pas inclus dans les analyses suivantes.

Pour résumer, je retiens dix clusters : "Pilot *uss*" (87 gènes), "Pilot *udss*" (23 gènes), "Pilot *uuss*" (23 gènes), "Pilot *ussd*" (16 gènes), "*Lu us_Dus_H*" (14 gènes), "*Lu us_Dss_H*" (24 gènes), "*Lu uu_Dus_H*" (27 gènes), "*Lu uu_Duu_H*" (60 gènes), "DeRenzis purement zygotiques" (325 gènes), "DeRenzis zygotiques précoces" (60 gènes). Les cibles secondaires purement zygotiques sont déjà incluses dans le cluster de gènes purement zygotiques, donc je ne rajoute pas ce cluster. J'ai obtenu ainsi un cluster de 417 membres que je nommerai le "cluster AGZ".

3.2 Analyse du cluster AGZ

3.2.1 Analyse fonctionnelle avec Gene Ontology

Afin d'étudier les fonctions des gènes présents dans le cluster AGZ, j'ai analysé l'enrichissement du cluster en termes gene ontology (GO) avec l'outil *compare-classes* (86) et les outils en ligne DAVID (<http://david.abcc.ncifcrf.gov/>) (97, 98) et GOToolBox (<http://genome.crg.es/GOToolBox/>) (99).

J'ai paramétré les programmes *compare-classes*, DAVID et GOToolBox afin d'obtenir des résultats comparables entre analyses. Cependant, chaque programme a des particularités au

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

niveau des calculs statistiques de significativité. Même si les trois algorithmes basent leur calcul de significativité sur la probabilité hypergéométrique, DAVID par exemple utilise le score EASE (100) qui consiste à modifier le test exact de Fischer (lui même basé sur la loi hypergéométrique) en procédant à l'extraction d'un élément et recalculant la probabilité exacte de Fisher. Cela permet de pénaliser la significativité des catégories contenant peu de membres. En effet, si une catégorie contient peu de membres et qu'on les retrouve tous dans un jeu d'une centaine de gènes, ceci peut-être du au hasard. Cependant, la présence des quatre membres peut avoir une pertinence biologique. Une différence majeure existe entre GOToolBox et les deux autres méthodes, en ce qui concerne le choix du groupe de gènes de référence. En effet, DAVID (par défaut) va calculer les probabilités en fonction du nombre total de gènes annotés dans au moins une des classes analysées, choix que j'ai également fait pour mes analyses avec compare-matrices. Par contre, GOToolBox utilise une référence globale (nombre total des gènes de la drosophile). Ceci entraîne une significativité va être plus importante, comme on peut l'observer dans le tableau 3.1. Le fait de prendre en compte tous les gènes est cependant contestable, car ceux ne présentant aucune annotation GO ne pourront jamais apparaître à l'intersection d'un cluster et d'une classe GO, et ne font donc pas partie de « l'univers » accessible à la distribution hypergéométrique. Pour les trois méthodes, la correction de Bonferroni pour les tests multiples a été appliquée.

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

Le cluster ZGA est donc enrichi pour beaucoup de termes. L'analyse séparée des clusters primaires ne donne pas autant de termes. Le cluster de gènes purement zygotiques défini par De Renzis et collaborateurs contient le plus grand nombre termes enrichis (90 termes). Nous retrouvons ces termes enrichis dans le cluster ZGA. Plusieurs termes enrichis sont relatifs à la

3.2 Analyse du cluster AGZ

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

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

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

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

3.2.2 Analyse des éléments cis-régulateurs

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

Afin d'analyser l'enrichissement en motifs de liaison connus, j'ai utilisé CisTargetX (101) (<http://med.kuleuven.be/cme-mg/lng/cisTargetX/>). En intégrant ces résultats et ceux de l'analyse des clusters primaires, j'ai sélectionné un jeu de matrices qui seront ultérieurement utilisées pour scanner les régions non codantes et de détecter des régions significativement enrichies en sites (CRER pour "*Cis-Regulatory element Enriched Region*") qui seront considérés comme des CRM potentiels.

3.2.2.1 Découverte de motifs

La première étape était de récupérer les séquences non codantes en amont du TSS, premiers introns et UTR. J'ai choisi de récupérer les 5kb en amont du TSS car on retrouve fréquemment des modules cisrégulateurs sur plusieurs Kb en amont des gènes,. En effet, il a été montré que les CRM pouvaient se trouver tant dans les régions inter-géniques que dans les introns (particulièrement dans le premier) et UTR (102). Pour de cela j'ai utilisé *retrieve-ensembl-seq* avec les options me permettant d'optimiser l'analyse de découverte de motifs (cf. section 3.3.1, figure 3.7). Afin de découvrir les motifs les plus significativement surreprésentés dans les séquences non codantes des gènes du cluster AGZ, comme précédemment, j'ai utilisé le pipeline *gene-cluster-motifs*. J'ai testé par ailleurs l'algorithme MEME (18), mais à cause du grand nombre de séquences à analyser, MEME n'a jamais donné de résultats (le temps de réponse de l'algorithme augmente en raison du carré de la taille des séquences). En revanche, *gene-cluster-motifs* permet de tester facilement et rapidement différents paramètres à différentes étapes du pipeline. Les détails de la paramétrisation sont dans la partie 3.3.2. La recherche de motifs connus avec CisTargetX a été opérée avec les paramètres par défaut.

3.2.2.2 Sélection des motifs parmi les matrices découvertes

108 matrices ont été découvertes (19 à partir d'*oligo-analysis* et 79 à partir de *dyad-analysis*). Ce grand nombre de matrices vient de deux facteurs. Premièrement, plusieurs assemblages alternatifs peuvent être formés à partir de groupes de mots partiellement chevauchants.

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

Ensuite, l'assemblage des dyades renvoie certains motifs très longs (environ 25 nucléotides) composés de plus de 10 dyades mais ces chevauchements n'ont pas forcément de sens, en effet la plupart des matrices résultantes sont construites à partir de très peu de sites (entre 1 et 50 sites). J'ai tout d'abord écarté les matrices construites à partir de moins de 100 sites. Ensuite, j'ai procédé à un ré-alignement mutuel des matrices restantes en utilisant *compare-matrices*. Lorsque les matrices s'alignent avec une corrélation de Pearson normalisée supérieure à 0.8, j'ai gardé la matrice la moins dégénérée. Un avantage de cette procédure est qu'elle indique que certains motifs ont été découverts dans plusieurs types de séquences. J'ai ainsi sélectionné 9 matrices que j'ai ensuite comparées à différentes bases de motifs connus (corrélation normalisée > 0,7).

3.2.2.3 Deux motifs connus retrouvés par l'approche *de novo* et par CisTargetX (Zelda et Trl)

Le motif de liaison de Zelda est le motif le plus significatif dans les deux approches (figure 3.5A et annexe 6.7). Ce facteur étant connu pour être impliqué de façon générale dans l'AGZ (88), c'est un bon indicateur de la pertinence du choix des gènes AGZ.

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

3.2.2.4 Autres motifs connus

Le motif de liaison de TTK apparaît comme surreprésenté en découverte de motifs, mais n'a pas été détecté par CisTargetX. Durant l'analyse des clusters primaires dans le chapitre , ce motif a été détecté dans le cluster regroupant les gènes zygotiques précoces de De Renzis et dans celui regroupant les gènes activés durant la phase lente de cellularisation "Lu *us_{DSSH}*", dont le mot principal "caggac" avait une significativité de 6.29 et 3.84 respectivement avec *dyad-analysis*. La significativité est plus faible (3.61) pour le cluster AGZ (figure 3.5C). Il existe un antagonisme entre Trl et TTK (94) dans la régulation du gène *eve*, il sera intéressant d'analyser les co-occurrences de Trl et TTK lors de l'analyse des régions enrichies en sites opérée par la suite avec *matrix-scan*. J'ai donc décidé de garder ce motif pour les analyses suivantes.

Comme le montre l'annexe 6.7, plusieurs motifs de liaison de FT impliqués dans la segmentation (Dorsal, Krüppel, Bicoid) sont également enrichis dans les séquences non codantes

3.2 Analyse du cluster AGZ

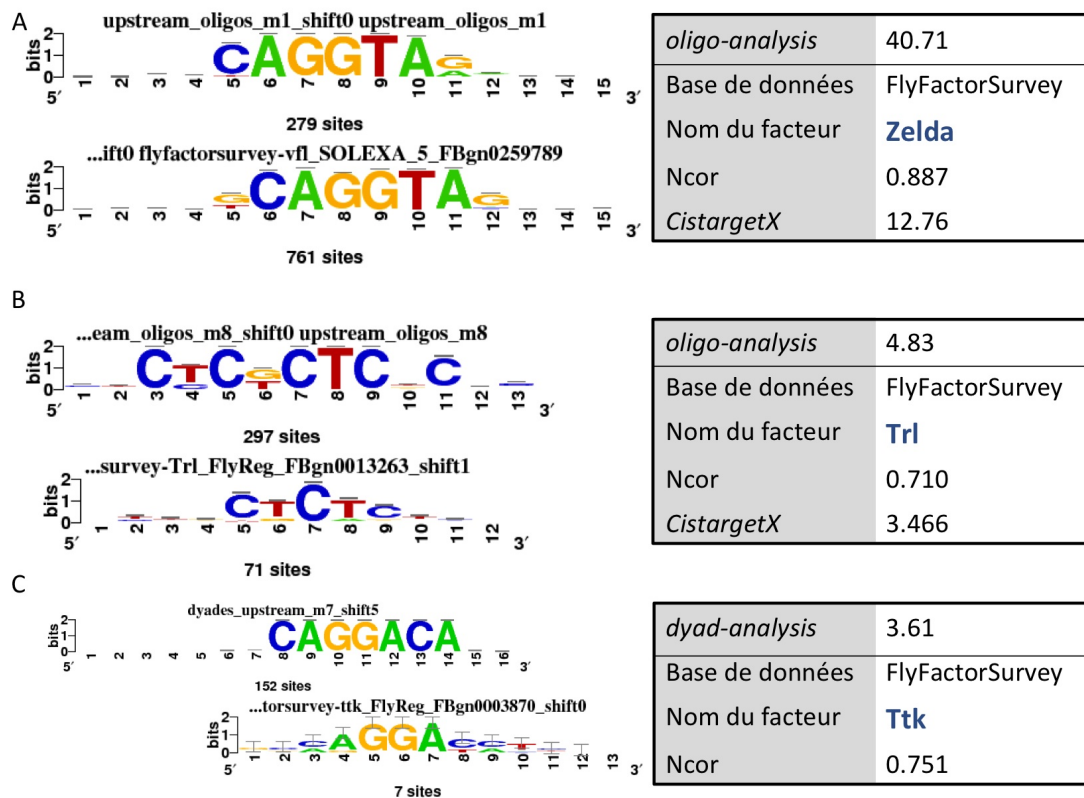


FIGURE 3.5: Correspondances entre motifs découverts et motifs connus. - Chaque ligne contient, à gauche, l'alignement des logos entre le motif découvert et le motif connu et le tableau à droite indique la significativité du motif découvert, le facteur se liant au motifs connu correspondant, la corrélation normalisée (Ncor) de l'alignement, et le score d'enrichissement de CisTargetX pour le même motif connu ou un motif similaire.

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

des gènes AGZ. Cependant, ces FT sont impliqués dans l'activation localisée spatialement au cours du processus segmentation, or je recherche des facteurs impliqués dans le mécanisme ubiquiste de l'activation zygotique. Donc, j'ai écarté les motifs reconnus par les gènes de segmentation pour les analyses subséquentes. Le motif de liaison de Su(H) ressort également. Durant la cellularisation, ce facteur est impliqué dans la définition du méséctoderme (deux lignes cellulaires entourant le neuroectoderme) en inhibant l'action de Twist et Dorsal dans le mésoderme et permettant le recrutement du fragment intracellulaire de Notch dans le méséctoderme (annexe 6.8) (103). Son action localisée n'en fait pas un bon candidat pour la régulation globale de l'AGZ.

3.2.2.5 Motifs inconnus

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

3.2.2.6 Détection de modules cis-régulateurs (CRM) potentiels

Afin d'identifier des CRM potentiels dans les régions non codantes des gènes AGZ (5kb en amont du TSS, premier intron, 5'UTR), j'ai utilisé *matrix-scan* (22) pour détecter les régions enrichies en sites (CRER : *Cis-Regulatory element Enriched Region*). Lors de cette analyse, il s'est avéré que les motifs 5 et 6 reconnaissent partiellement les mêmes sites, en dépit des différences pour les résidus A ou T centraux. Si les sites reconnus par ces matrices sont chevauchants, le nombre de sites sera dans une région donnée sur-estimé et faussera les résultats du test binomial. Cependant, lors du scan des régions non codantes pour la recherche de CRER, j'ai utilisé l'option qui permet d'écarter les sites chevauchants pour le calcul de la significativité de l'enrichissement en site de la région (cf. 3.3.3). Le maintien des deux matrices partiellement similaires ne pose donc pas de problème pratique pour la prédiction de CRM.

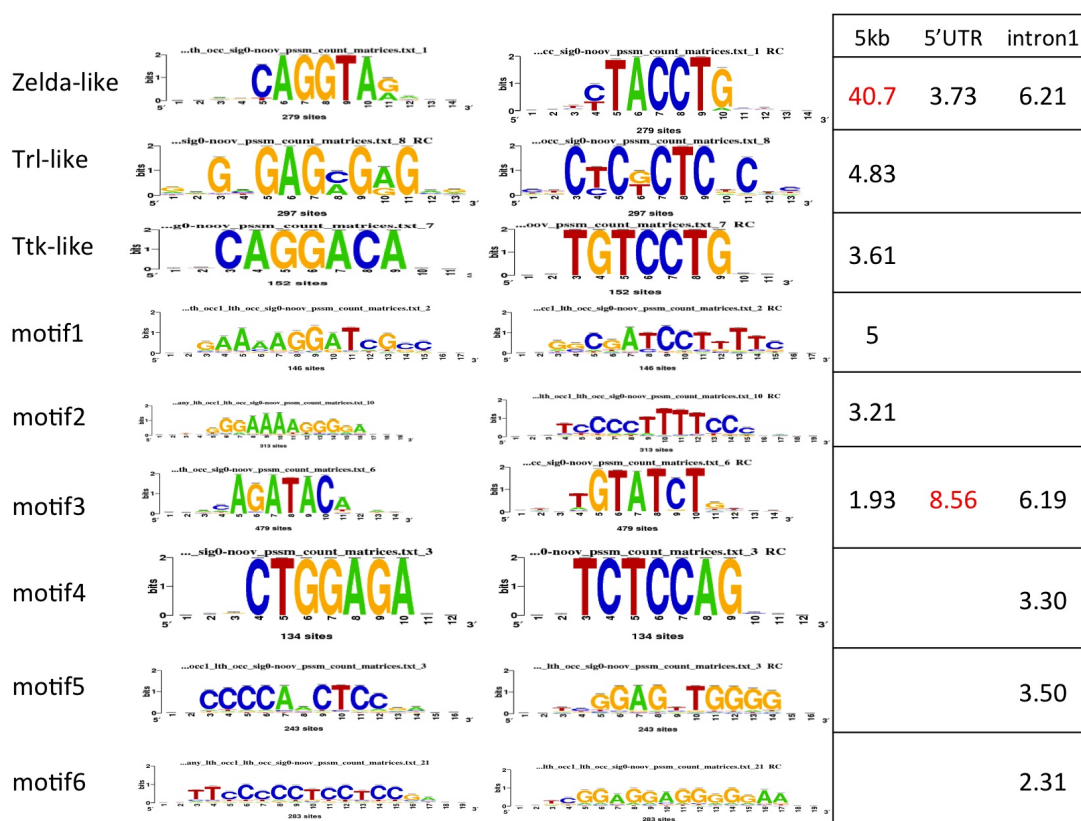


FIGURE 3.6: Liste des motifs sélectionnés pour la prédiction de CRM. - Les chiffres indiquent la significativité de la surreprésentation des motifs selon le type de séquences non-codantes.

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

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

3.3 Matériels et Méthodes

3.3.1 Récupération des séquences non-codantes

Pour récupérer les séquences non codantes j'ai utilisé l'outil RSAT *retrieve-ensembl-seq* (104) avec les options suivantes :

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

La figure 3.7 (4) montre que les différents types de séquences non codantes peuvent se chevaucher. Ceci peut poser problème lors de la découverte de motifs. En effet, j'ai calculé le background à partir des séquences non codantes de même type que les séquences analysées, cependant il apparaît une hétérogénéité. L'impact du background est discuté dans la partie 3.1.2.1.

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

3.3.2 Paramètres pour la découverte de motifs

oligo-analysis et *dyad-analysis* (105) sont des algorithmes basés sur le comptage de mots simples (*oligo-analysis*) ou de paires de mots espacés (*dyad-analysis*). Le nombre d'occurrences de chaque mot (ou dyade) d'une taille donnée est calculé dans le jeu de séquences à tester et comparé au nombre attendu au hasard, d'après un jeu de séquences de référence (modèle de background). Le choix de ce modèle va donc être capital pour identifier les mots (ou dyades) les plus pertinent(e)s. La composition en nucléotides varie entre les espèces mais aussi selon le type de séquence (codantes, intronique, intergénique, ...). La significativité du nombre d'occurrences observée est estimée sur base de la distribution binomiale. Le programme retourne une P-valeur nominale (risque de faux-positifs pour un mot donné), une E-valeur (nombre de faux-positifs attendus pour un seuil de P-valeur donné) et un indice de significativité ($sig = -\log_{10}(E - value)$). J'ai gardé les mots (et dyades) dont la significativité est supérieure à 0.

3.3.2.1 *oligos-analysis*

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

3.3.2.2 *dyad-analysis*

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

3.3.2.3 *matrix-from-pattern*

Cet algorithme extrait une série de matrices position-poids à partir d'une liste de mots sur-représentés produite par *oligo-analysis* et *dyad-analysis*. Il procède de la façon suivante :

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

Cet algorithme permet de filtrer les motifs résultants des assemblages de mots les moins significatifs grâce au seuil appliqué sur le poids des sites à utiliser pour la construction des matrices de comptage. Le nombre de matrices en sortie ne correspond donc pas toujours au nombre d'assemblages de mots.

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

La détection des sites et modules cis-régulateurs a été effectuée avec le programme *matrix-scan* (22). Ce programme supporte des modèles Markoviens, mais le temps de calcul augmente très rapidement avec l'ordre du modèle. Nous avons donc choisi un modèle d'ordre (m=2) plus faible que pour la découverte de motifs directement calculé sur les séquences analysées.

Options :

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

3.3.4 Bases de données de PSSMs

3.3.4.1 FlyFactorSurvey

FlyFactorSurvey (31) (<http://pgfe.umassmed.edu/TFDBS/>) fournit une base de motifs de liaison pour des facteurs de drosophile, en provenance de différentes sources.

- Expériences de "bacterial one-hybrid" (B1H) (39, 40)

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

- Expérience de "DNase1 footprinting" (36) directement disponibles sur le site web <http://www.flyreg.org/>.
- La base de données TIFFIN (37), composée de motifs découverts à partir de promoteurs et validés par comparaison avec les motifs définis à partir d'expériences de SELEX (JASPAR55) et de DNase1 footprinting (<http://www.sanger.ac.uk/resources/databases/tiffin/index.jsp>).

3.3.4.2 Jaspar core insect

Jaspar core insect (30) est disponible sur le site web http://jaspar.genereg.net/cgi-bin/jaspar_db.pl?rm=browse&db=core&tax_group=insects. Cette base regroupe des matrices construites en grande partie à partir d'expérience de SELEX, l'autre partie provient de la collecte de données dans la littérature.

3.4 Conclusion du chapitre

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

3.4 Conclusion du chapitre

Le regroupement des clusters zygotiques a renforcé la significativité des motifs détectés précédemment, à part dans le cas de Caudal dont le motif de liaison n'est plus détecté. Ainsi, cinq motifs ont été redécouverts dans le cluster AGZ : trois motifs de liaison de facteurs connus (Zelda, Trl et TTK) et deux motifs inconnus.

3. ANALYSE FONCTIONNELLE DES GÈNES ET DES RÉGIONS NON-CODANTES ASSOCIÉES

4

Analyse génomique des profils d'occupation et d'accessibilité de la chromatine

La découverte du motif de liaison de Trl m'a conduit à m'intéresser à la régulation de la transcription par le modelage de la structure de la chromatine. Comme nous l'avons vu dans l'introduction, la présence des FT et de leurs sites de liaison n'est pas suffisante pour expliquer la régulation de l'expression des gènes. En effet, l'accessibilité de ces régions joue un rôle important dans cette régulation. Mes recherches bibliographiques autour de Trl m'ont conduit à suspecter une implication de CBP (CREB-Binding Protein), codé par le gène *nejire*. CBP fait partie d'un complexe de remodelage de la chromatine (TAC1) incluant des membres du groupe Trithorax, dont le recrutement au promoteur du gène *hsp70* est facilité par Trl et HSP (106). Synthétisé au niveau maternel, CBP est un co-régulateur général de la transcription recruté à la chromatine par des centaines de facteurs. CBP joue un rôle décisif dans la différenciation cellulaire lors du développement (107) en participant à la régulation de plusieurs voies de signalisation (co-activateur de Mad dans la voie Dpp/Scw (108), de CI dans la voie Hedgehog (109), de Dorsal dans la voie Toll (110); répresseur de dTCF dans la voie Wnt/Wingless (111)). CBP agit au moins à deux niveaux de régulation. En effet, CBP est capable d'interagir directement ou indirectement (par l'intermédiaire des complexes médiateurs (112)) avec les facteurs généraux de transcription et l'ARN polymérase II (113) et fait ainsi le lien entre les facteurs de transcription et la machinerie basale de la transcription. CBP a également une activité acétyltransférase. En effet, CBP peut s'associer avec des membres du

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

groupe Trithorax (TRX et UTX) pour contrer les effets répresseurs des complexes Polycomb PRC1 et PRC2, qui déacétylent et triméthylent la lysine K27 de l'histone H3 (H3K27) (114). En fait, UTX déméthyle H3K27me3 et TRX tri-méthyle la lysine K4 de H3, CBP cible particulièrement les H3K4me3 (115) et acétyle H3K27. A cet égard, dans le cadre du projet modENCODE (<http://www.genome.gov/modencode/>), qui regroupe un grand nombre d'expériences de ChIP-seq à différents stades du développement de la drosophile, Nègre et collaborateurs (116), ont récemment mis en évidence que CBP était associé aux CRM activateurs à tous les stades du développement et que la présence de CBP était suffisante pour prédire de vrais CRM. Cette étude couvre également quatre marques de méthylation, deux marques d'acétylation de l'histone H3, l'ARN polymérase II (GSE16013), Polycomb-like (Pcl) (GSE23127), qui est un membre du groupe Polycomb et est associé au complexe PRC2 pour la tri-méthylation de H3K27 (117), et enfin Trl (GSM614652). Ils ont également trouvé une corrélation entre la présence de la marque H3K4me1 et des pics CBP avec l'activité des CRM. Enfin, leurs résultats suggèrent que les marques H3K4me3 et d'acétylation sont spécifiques des promoteurs actifs. Toutes les données sont disponibles dans les méthodes supplémentaires de l'article de Nègre et collaborateurs (116). Par ailleurs, j'ai récupéré des données d'accessibilité à la DNase1 générée dans le cadre du projet Berkeley Drosophila Transcription Network Project (BDTNP) (118, 119). Les auteurs fournissent les données brutes dans un format appelé SRA ("*sequence read archive*") permettant de reprendre dès le départ l'analyse de ces données, sous la forme d'un fichier contenant les densités normalisées des reads, ou encore sous la forme d'un fichier contenant les pics. Ces données sont disponibles sur le site <http://bdtnp.lbl.gov/FlyNet/browseAccess.jsp>. Enfin durant la rédaction de ce manuscrit, des expériences de ChIP-seq avec le facteur Zelda ont été publiées (88). Dans un premier temps, j'ai utilisé les pics provenant des expériences de Trl, de CBP, et de la DNase1 afin de calculer la significativité de leur chevauchement avec les séquences non-codantes des gènes AGZ. Afin d'identifier des facteurs potentiellement impliqués dans le recrutement CBP, ou des co-facteurs de Trl et Zelda, ou plus généralement des facteurs présents dans les régions actives (H3K4me1) ou accessibles (DNase1) de la chromatine, j'ai procédé à la découverte de motifs surreprésentés dans la totalité des pics avec le programme *peak-motifs*. J'ai également analysé les sous-ensembles de pics chevauchant les régions non codantes des gènes AGZ pour identifier des motifs qui serait spécifiquement reliés à l'AGZ. J'ai enfin analysé les profils de liaison et d'occupation des différentes protéines ainsi que l'accessibilité de la chromatine dans les CRM AGZ afin d'identifier si ces CRM sont spécifiquement associés à un environnement chromatinien particulier.

4.1 Utilisation des pics

J'ai tout d'abord utilisé les pics mis à disposition pour les expériences de ChIP-seq pour CBP (E0-4h) et de Trl (E0-8h) d'accessibilité à la DNaseI (stade 5), et de ChIP-seq pour Zelda (cycles mitotiques 8-9 : 1h ; cycles mitotiques 13-14 : 2h ; cycle mitotique 14 : 3h). Les auteurs des différentes expériences ont utilisé diverses méthodes afin de détecter les pics de liaison de facteurs (MACS (120)) ou d'accessibilité à la chromatine (méthode décrite dans Thomas et al. (118)). Harisson et al. (2011) ont mis au point leur propre outil de détection de pics *Grizzly-peak* (<http://eisenlab.org/software/grizzly>). L'intersection entre les régions non-codantes et les pics a été faite avec les outils fournis dans la suite BedTools (121) qui permet de manipuler des fichiers "bed" contenant la localisation génomique des régions. J'ai utilisé Galaxy (<http://main.g2.bx.psu.edu/>) afin de récupérer les séquences nucléotidiques des pics sur la base de leurs coordonnées.

4.1.1 Chevauchement entre et régions non codantes des gènes AGZ

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

De plus, en visualisant les résultats sur le genome browser de UCSC (<http://genome.ucsc.edu>) (122), je me suis aperçue que ces pics pouvaient être mal définis. En effet, la Figure 26 montre deux cas où la prédiction des pics n'est pas pertinente au regard des profils de liaisons représentés par la densité en reads normalisée. Dans le premier cas (figure 4.1A), le pic recouvre une région très étendue autour du pic réel de liaison, et sa longueur dépasse les 4,5 Kb. Le second cas (figure 4.1B) révèle que MACS, avec les paramètres utilisés par les auteurs, manque de résolution et ne prédit qu'un pic manifestement produit par la fusion de trois pics distincts. J'ai tout de même procédé à la découverte de motifs surreprésentés dans ces pics pour identifier des facteurs potentiellement impliqués dans le recrutement de CBP, de co-facteurs de Trl, ou

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

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

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

encore de motifs associés régions accessibles et/ou actives de la chromatine. Ces résultats sont présentés dans la section suivante.

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

4.1.2 Analyse des pics avec *peak-motifs*

J'ai procédé à deux analyses complémentaires. Premièrement, j'ai analysé l'ensemble des pics des jeux de données de CBP, Trl, Zelda, d'accessibilité à la DNase1 et de H3K4me1. Ensuite, j'ai analysé les pics chevauchant les régions non codantes des gènes AGZ (que je nommerai "pics AGZ") de ces mêmes jeux. En effet, je voulais identifier les motifs reliés à ces différentes marques, mais surtout identifier ceux qui sont spécifiques à ces marques dans les séquences non-codantes des gènes AGZ. A cet égard, j'ai utilisé *peak-motifs*, qui combine plusieurs outils de découverte de motifs (oligo-, dyad-, position- et local-word-analysis), ainsi qu'un outil de comparaison de motifs (compare-matrices). J'ai participé à l'évaluation de l'outil en procédant au choix et à l'analyse de cas d'études pour un protocole d'utilisation de l'outil (annexe 6.2.1). Ces analyses m'ont permis de d'apprécier l'impact des différents paramètres sur les résultats. J'ai ainsi utilisé des modèles de background appropriés à chaque cas. Pour les jeux de pics entiers contiennent plus d'un Mb de séquences, j'ai donc utilisé le modèle

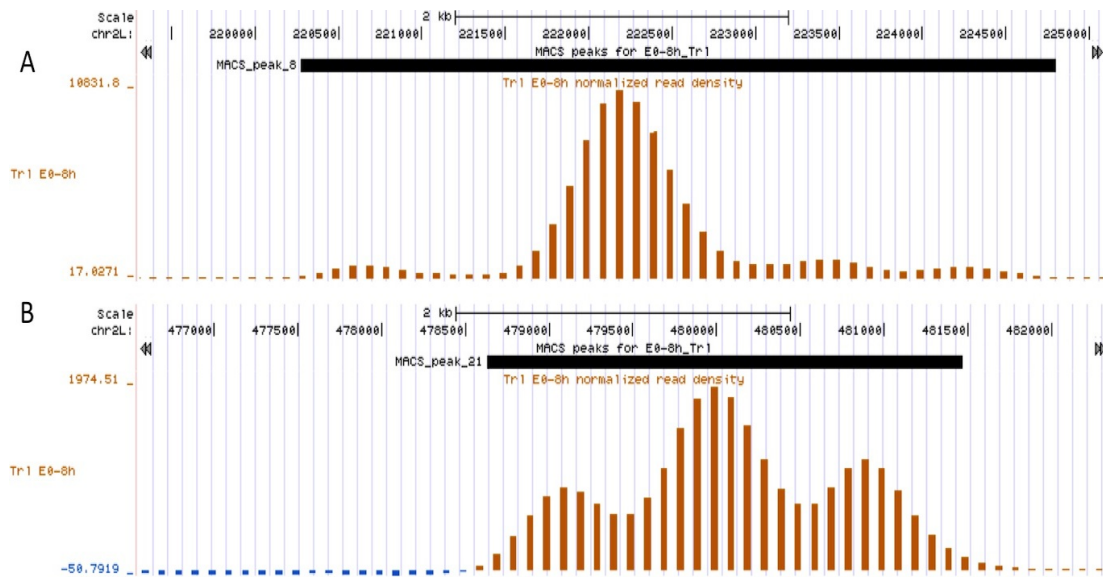


FIGURE 4.1: Visualisation de données provenant d'une expérience de liaison de Tr1 entre 0 et 8h après la fécondation (ChIP-seq). - A et B montrent les profils en densité de « reads » normalisée (orange, positive ; bleu négative) ; la piste "MACS peaks for E0-8h_Tr1" contient les pics prédits par MACS.

de markov d'ordre $n - 2$ où n est la longueur des oligo-nucléotides analysés. Pour les sous-ensembles de pics, en général inférieur à un Mb, j'ai utilisé un modèle d'ordre $n - 3$. J'ai également décidé de ne pas couper les pics malgré la présence de pics très longs. En effet, comme nous l'avons vu plus haut, les pics ne correspondent pas forcément à une seule région de liaison. Si l'on considère l'occupation des histones, l'approche ChIP-seq met en évidence des régions éventuellement occupées par plusieurs nucléosomes marqués, et les pics longs ne sont donc pas forcément aberrants.

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

De manière générale, à part pour les pics d'occupation de H3K4me1, dans les pics AGZ, le motif de liaison de Zelda (consensus "CAGGTA_g") est fortement sur-représenté. C'est bien entendu particulièrement le cas dans les pics provenant des expériences de liaison de Zelda. Les résultats montrés ont été calculés avec l'algorithme position-analysis, ce qui nous permet de préciser que ces motifs sont centrés au milieu des pics. Le motif résultant est construit à partir de 3530 sites et représente probablement un modèle de liaison plus proche du "vrai" consensus que ceux intégrés dans la base de données FlyFactorSurvey (les deux matrices présentes ont été

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

construites sur la base de 18 sites documentés et de 761 sites SELEX, respectivement). Nous retrouvons également le motif de liaison de Zelda dans le jeu entier de pics de CBP, mais de façon moins significative que pour les gènes AGZ. Dans les autres jeux entiers de pics, nous ne détectons pas ce motif. Le motif de liaison de Trl et le motif non connu « cAGATACa » apparaissent fortement surreprésentés dans l'ensemble des jeux et sous jeux de pics. Le motif de liaison de Trl est généralement plus significativement sur-représenté dans les jeux entiers de pics que dans les pics AGZ contrairement au motif « cAGATACa » qui est généralement plus fortement sur-représenté dans les pics AGZ. Le motif « cAGATACa » est cependant plus significatif dans les jeux entiers des pics Zelda 1h et 2h. Ceci suggère que ce motif est plus spécifiquement associé à la liaison de Zelda dans les séquences non codantes AGZ durant la cellularisation (Zelda 3h) que durant les intervalles de temps précédents. Les données de DNase1 nous confirment que les motifs identifiés se trouvent dans les régions ouvertes de la chromatine et donc accessibles aux TF.

4.2 Utilisation des densités de "reads"

J'ai utilisé les données de densité de reads représentant le profil de liaison et d'occupation de protéines ou d'accessibilité à la DNase1. Ces données contiennent des densités de reads discrétisées. Le génome a été fractionné en segments de taille donnée (20 ou 100 pb, pour les données d'accessibilité à la chromatine ou d'occupation et de liaison à la chromatine, respectivement) et une densité moyenne (pour les profils d'occupation d'histones) ou normalisée par le background (pour les profils de liaison de facteurs) a été assignée à chaque segment.

J'ai analysé les profils de liaison de Pcl, Trl et Zelda, les profils de recrutement de CBP, les profils d'occupation des histones H3 : tri-méthylation des lysines 4 (K4me3), 27 (K27me3) et 9 (K9me3) ; monométhylation des lysine 4 (K4me1) ; acétylation des lysines 9 (K9ac) et 27 (K27ac) et d'accessibilité de chromatine à différentes classes temporelles. J'ai ainsi analysé 35 profils sous les régions définies par les CRM prédits. Afin, d'analyser la spécificité de ces profils, j'ai également considéré plusieurs contrôles négatifs et positifs.

4.2.1 Les contrôles positifs et négatifs

Afin de vérifier la spécificité des associations observées pour les CRM prédits avec les matrices résultant de la découverte de motifs sur-représentés dans les séquences non codantes des 417 gènes AGZ (que je nommerai dans cette partie "matrices AGZ"), j'ai utilisé trois contrôles

4.2 Utilisation des densités de "reads"

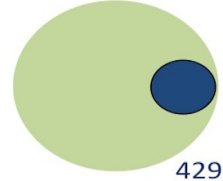






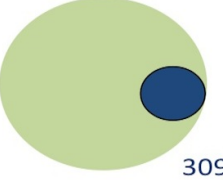





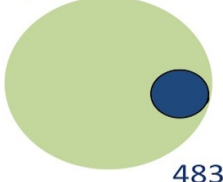



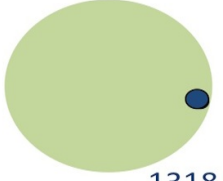





Experiments (timing)	Number of peaks	Discovered motifs in ZGA peaks	Discovered motifs in all peaks
CBP (E0-4h)	3296 	  	  
Trl (E0-8h)	3723 	  	 
H3K4me1 (E0-4h)	6162 		 
DNase1 (stade 5)	22693 	  	 

FIGURE 4.2: Résumé de la découverte de motifs effectuée sur les pics CBP, Trl, H3K4me1 et accessibilité à la DNase1. - Les cercles verts représentent l'ensemble de pics, les cercles bleus foncés représentent les pics chevauchant les séquences non-codantes des gènes AGZ. La significativité est indiquée à coté de chaque logo, l'exposant indique l'outil utilisé (oa : *oligo-analysis*, pa : *position-analysis*). Les motifs surlignés en gris correspondent aux motifs de liaison du facteur étudié.

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

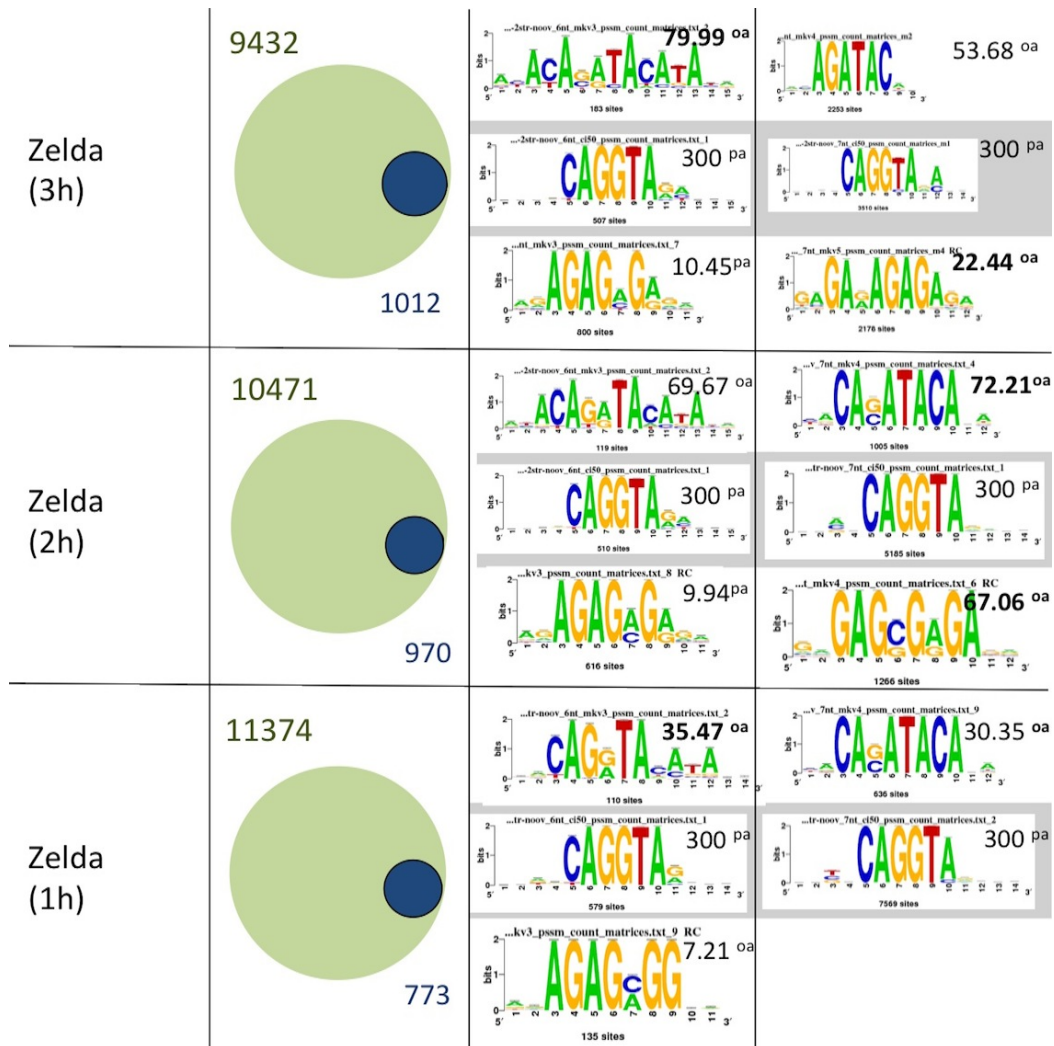


FIGURE 4.3: Résumé de la découverte de motifs effectuée sur les pics Zelda. - Les cercles verts représentent l'ensemble de pics, les cercles bleus foncés représentent les pics chevauchant les séquences non-codantes des gènes AGZ. La significativité est indiquée à côté de chaque logo, l'exposant indique l'outil utilisé (oa : *oligo-analysis*, pa : *position-analysis*). Les motifs surlignés en gris correspondent aux motifs de liaison du facteur étudié.

négatifs et un contrôle positif. Le premier contrôle négatif a consisté à prédire des CRM à l'aide des matrices AGZ sur les régions non codantes de 417 gènes choisis au hasard. Le second contrôle négatif a consisté à permuter les matrices AGZ et à utiliser les matrices résultantes pour prédire des CRM sur les séquences non codantes des gènes AGZ. J'ai également utilisé les 317 CRM non actifs dans le blastoderme disponible dans REDFly (123) comme troisième contrôle négatif. Enfin, j'ai analysé comme contrôle positif les 114 CRM répertoriés dans REDFly comme étant actifs dans le blastoderme.

4.2.2 Méthodologie

4.2.2.1 Calcul de densité intégrée

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

$$S_u = \sum_{i=1}^{n-1} \left(\frac{h_i + h_{i+1}}{2} \right) (x_{i+1} - x_i) \quad (4.1)$$

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

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

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

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

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

$$S_s = \left(\frac{h_1 + h_s}{2}\right)(x_1 - x_s) \quad (4.4)$$

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

$$S_e = \left(\frac{h_n + h_e}{2}\right)(x_e - x_n) \quad (4.5)$$

La densité intégrée S_r sous la région entière est donc :

$$S_r = S_s + S_u + S_e \quad (4.6)$$

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

$$S_{norm} = \frac{S_r}{L_r} \quad (4.7)$$

où L_r est la longueur de la région.

4.2.2.2 Construction des courbes ROC

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

J'ai ensuite normalisé les axes afin de pouvoir analyser la spécificité et la sensibilité du critère choisi (par exemple la densité en reads intégrée obtenue pour un facteur donné) pour la discrimination des CRM par rapport aux régions aléatoires. La figure 4.5 représente le résultat obtenu en classant les CRM prédits dans les régions non-codantes des gènes AGZ et les régions choisies au hasard en fonction de la densité intégrée S_r calculée à partir des densités en reads provenant de l'expérience de liaison du facteur Zelda. L'abscisse, qui indique la fraction de régions choisies au hasard présentes dans le classement, représente le taux de faux positifs (FP)

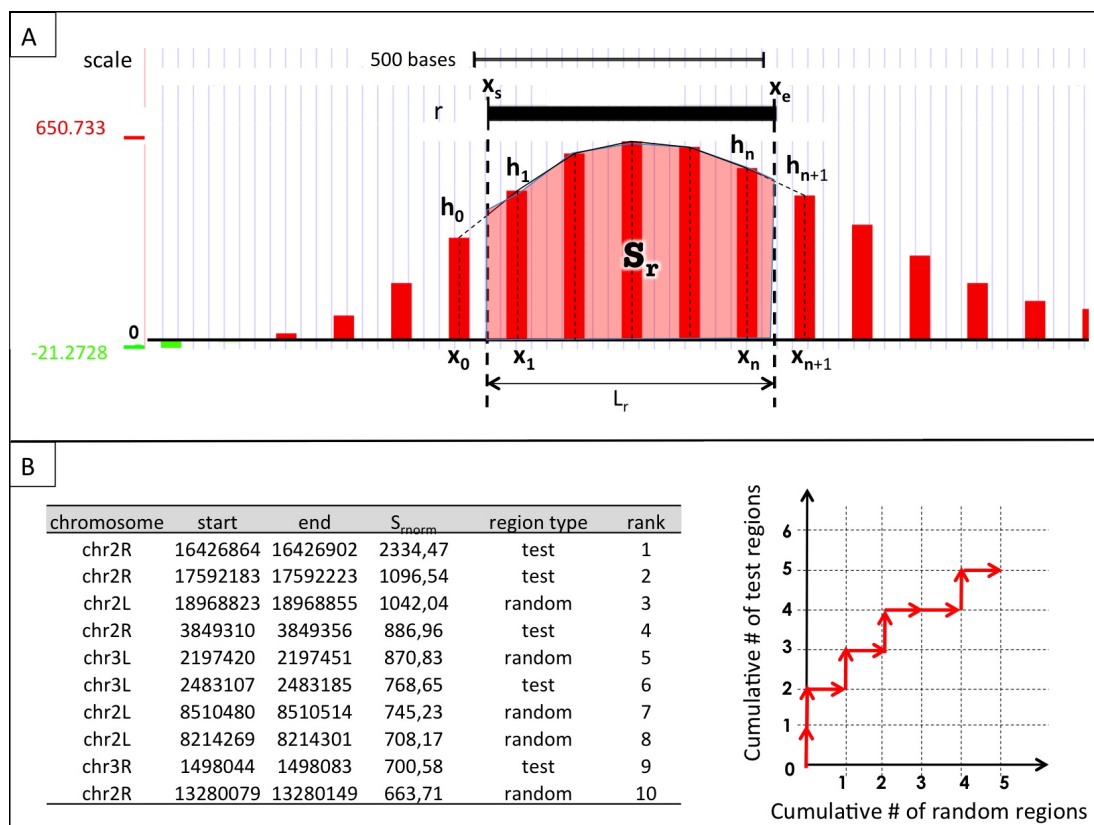


FIGURE 4.4: Représentation schématique du calcul de la densité en reads intégrée sous une région donnée (A) et obtention d'une courbe de ROC (B). - A. L'histogramme représente les densités en reads discrétisées en segment de 75 pb. La hauteur de chaque barre correspond à la densité en reads normalisée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

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

attendu si l'on cherche à détecter les CRM en utilisant le profil de liaison de Zelda. L'ordonnée représente la fraction de CRM présents dans le classement. La ligne diagonale représente donc une répartition homogène des CRM et régions aléatoires dans le classement. Dans le cas présenté ici, nous pouvons voir qu'un seuil de S_r minimale de liaison de Zelda de 9.23 (les valeurs d'intensité allant de 4344.79 à 0) permet de détecter 80% des CRM et 20% des FP. Ceci indique que le facteur Zelda est recruté spécifiquement dans les CRM. De manière générale, plus une courbe s'éloigne de la diagonale en suivant l'axe des ordonnées, plus le facteur (ou marque chromatinienne) testé est spécifique des CRM. Inversement, si la courbe suit la diagonale ou l'axe des abscisses, cela indique qu'il n'y a pas d'enrichissement particulier ou que la liaison du facteur est évitée dans les régions définies par les CRM respectivement. Afin de mesurer plus précisément la spécificité des facteurs testés pour les CRM, j'ai calculé l'aire sous la courbe (AUC) pour un seuil de détection de 20% de FP. Similairement au comportement des courbes par rapport à la diagonale, des AUC positives, nulles ou négatives indiquent un enrichissement spécifique, aucun enrichissement ou un évitement des facteurs testés.

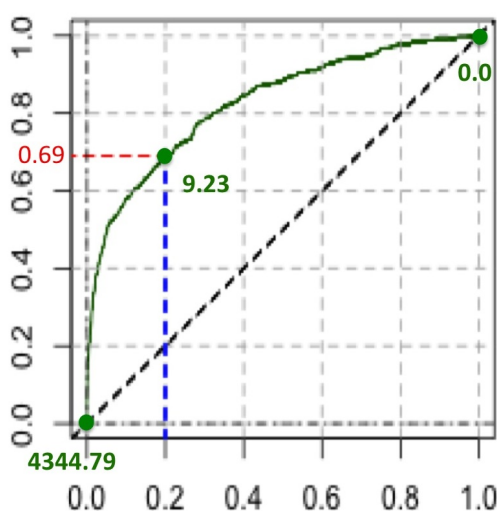


FIGURE 4.5: Courbe de ROC correspondant au classement des CRM AGZ (prédits dans les régions en amont du TSS) parmi des régions non codantes choisies au hasard en fonction des densité intégrée obtenues avec les données de ChIP-seq du facteur Zelda (3h). - L'abscisse et l'ordonnée représentent respectivement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

4.2.2.3 Combinaison d'expériences

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

$$\tilde{S}_r = \tilde{m}_{k_\omega} \quad (4.8)$$

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

4.2.3 Résultats

J'ai tout d'abord procédé à l'analyse séparée de chaque jeux de données, afin de confronter les résultats obtenus entre les CRM prédits dans les régions non codantes des gènes AGZ (que je nommerais par la suite CRM AGZ) avec ceux obtenus avec les divers contrôles, et ainsi détecter les associations spécifiques de certaines marques avec les CRM AGZ. J'ai ensuite analysé le comportement de la liaison, de l'occupation et de l'accessibilité de la chromatine au cours du temps dans les CRM AGZ et les CRM présent dans RedFLy, afin d'évaluer la spécificité des enrichissements découverts pendant la période couvrant l'AGZ. Enfin, j'ai calculé l'enrichissement pour des combinaisons de marques, afin de détecter des associations particulières. Je me limiterai ici à la présentation des résultats obtenus pour les CRM prédits dans les régions en amont des TSS. En effet, les CRM prédits dans les 5'UTR et dans les premiers introns montrent de faibles enrichissements. Par ailleurs, l'analyse de tous les CRM AGZ montre un enrichissement plus faible que les CRM prédits dans les régions en amont du TSS. La permutation des matrices permet de produire des motifs aléatoires ayant le même contenu informationnel que les matrices de départ. Ce contrôle va donc permettre de tester si l'enchaînement Les cinq contrôles consistant à scanner des groupes de gènes choisis au hasard avec les matrices AGZ ne retournent que très peu de CRER et ne révèlent aucun enrichissement particulier (figure 4.6) ; un seule contrôle est montré). Ceci suggère que les matrices AGZ sélectionnées sont spécifiques du groupe de gènes AGZ et prédisent des CRER dans des régions pourvues de signatures chromatiniennes particulières. En revanche, les CRER prédits en scannant les régions non codantes des gènes AGZ avec les matrices AGZ permutées sont des régions montrant les mêmes signatures chromatiniennes que les CRM AGZ (figure 4.6). Environ 380 CRER ont été prédits avec les matrices permutées, parmi ces CRER, 260 sont chevauchant avec au moins

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

un CRM AGZ. Ceci s'explique par la sélection de plusieurs matrices AGZ riches en GA (Trl-like, motif2, motif5 et motif6) dont les permutations produisent des motifs peu différents des matrices AGZ.

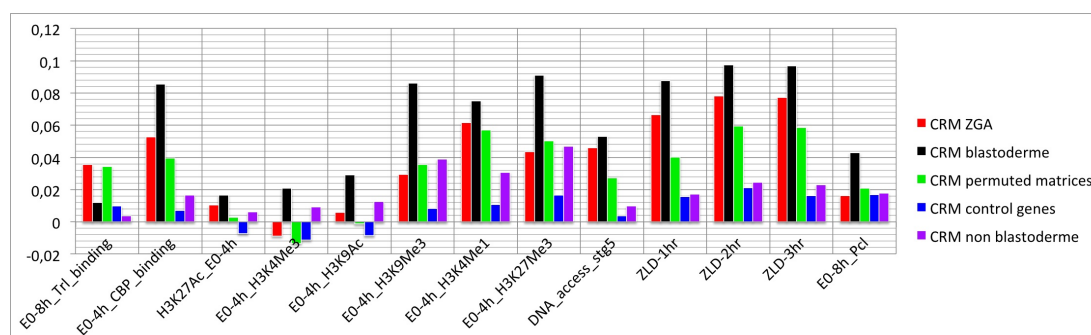


FIGURE 4.6: Distribution des valeurs d'AUC calculées en fonction des intensités de liaison, d'occupation ou d'accessibilité de la chromatine sous les CRM AGZ et contrôles. - L'abscisse indique les noms des facteurs (CBP, Trl et Pcl), modifications d'histone H3 et accessibilité de la chromatine (DNA_access)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

4.2.3.1 La chromatine dans tous ses états

Marques liées à l'activité transcriptionnelle

Plusieurs expériences permettent d'analyser l'activité transcriptionnelle au niveau de la chromatine. Tout d'abord l'accessibilité de la chromatine peut être évaluée grâce aux expériences de DNase1. Les CRM AGZ et blastoderme sont associés spécifiquement aux régions d'ouverture de la chromatine au stade 5 (figure 4.7A), par ailleurs, un faible enrichissement est obtenu dans les CRM contrôles négatifs. L'association des CRM AGZ et des CRM blastoderme avec cette marque diminue au cours du temps (figure 4.7 B-C), ce qui suggère que ces correspondent à des régions accessibles de la chromatine. L'activité transcriptionnelle peut également être révélée par la liaison de l'ARN polymérase II (polII). Cependant, je n'ai pas trouvé de données de CHIP-seq pour la polII dans un intervalle de temps englobant l'AGZ, les données les plus précoces disponibles correspondant à 4-8h après la fécondation (figure 4.7D). Durant cette période aucun enrichissement n'est détecté dans aucun des tests. L'analyse des profils d'occupation de la polII disponibles (figure 4.7 E-F), indique que les CRM AGZ et les CRM blastoderme présentent les mêmes profils d'enrichissement, avec un fort entre 8 et 12h.

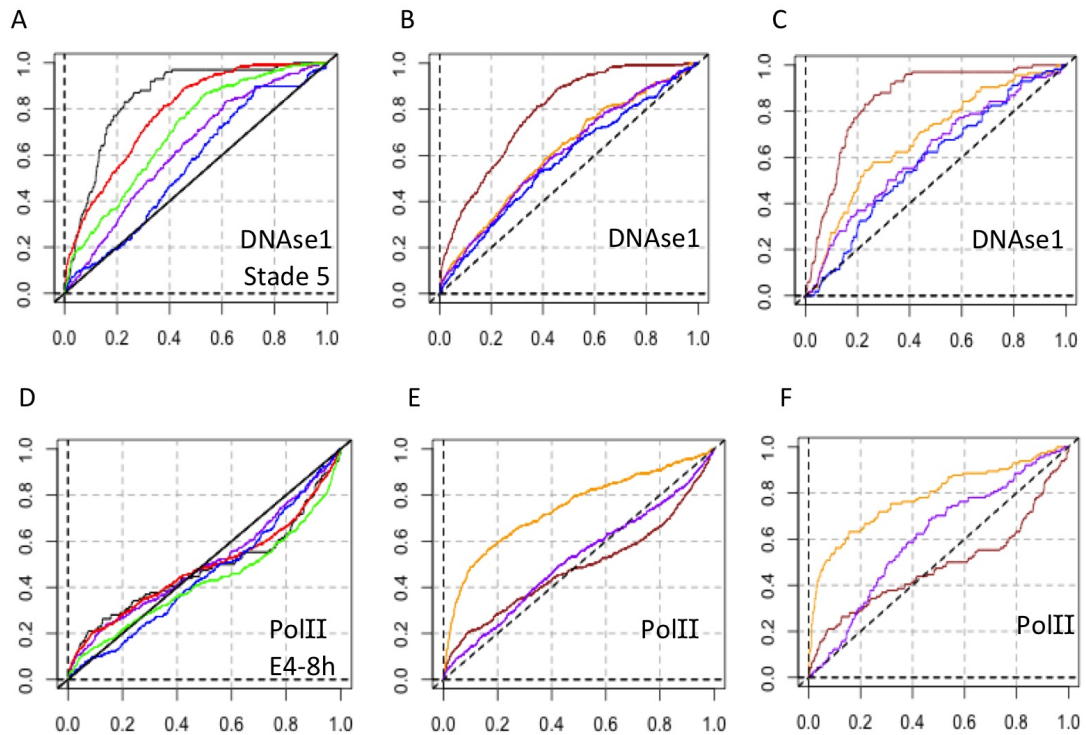


FIGURE 4.7: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques d'ouverture de la chromatine (DNase1) et de liaison de l'ARN Polymérase II (polII).
 - A et D. La protéine analysée ainsi que l'intervalle temporel de chaque expérience est indiquée dans le coin bas à droite de chaque graphe. La courbe rouge correspond au classement des CRM AGZ ; les courbes noires et violettes correspondent au classement des CRM blastoderme et non blastoderme respectivement ; les courbes bleues et vertes correspondent aux classements des CRM contrôles obtenus avec un jeu de gènes aléatoires et avec les matrices permutées respectivement.
 B-C, E-F. Résultats d'analyses d'ouverture de la chromatine (B-C) et l'occupation de la PolII (E-F) au cours du temps pour les CRM AGZ (B, E) et les CRM blastoderme (C, F). Les stades développementaux des expériences de DNase1 sont représentés en rouge (stade 5), orange (stade 9), violet (stade 11) et bleu (stade 14). Les intervalles temporels des expériences d'occupation de la PolII sont représentés en rouge (4-8h), orange (8-12h) et violet (12-16h)

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

Les régions activatrices de la transcription sont associées à différentes marques (116). En effet, Nègre et collaborateurs ont mis en évidence l'enrichissement en CBP et H3K4me1 dans les CRM blastoderme. Les résultats présentés dans la figure 4.8 sont en accord avec cette étude. L'enrichissement observé pour ces deux marques dans les CRM AGZ indique que ceux-ci correspondent à dans des régions activatrices de la transcription. L'analyse des marques CBP au cours du temps montre que les CRM AGZ et blastoderme ne restent pas associés à CBP. Ceci suggère que les CRM AGZ sont spécifiquement actifs dans les embryons âgés de 0 à 4h.

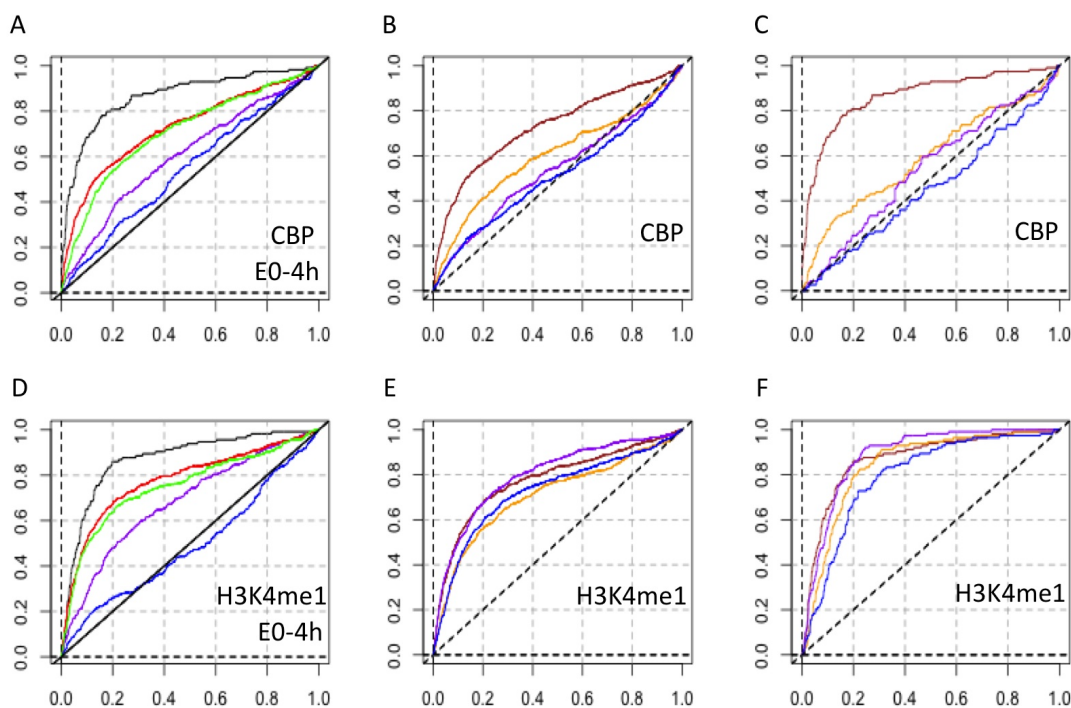


FIGURE 4.8: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques associées aux régions activatrices. - A et D. La protéine analysée ainsi que l'intervalle temporel de chaque expérience est indiquée dans le coin bas à droite de chaque graphe. Le code couleur des courbes est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

Nègre et collaborateurs ont également identifié plusieurs marques d'histones associées aux promoteurs actifs (H3K4me3, H3K9ac et H3K27ac). Un enrichissement apparaît dans les courbes ROC construite à partir des trois marques dans les CRM blastoderme (figure 4.9 A-C). La valeur des AUC (figure 4.6) montre cependant une très faible spécificité. Les CRM AGZ ne

montrent pas d'enrichissement particulier. En effet, la marque H3K4me3 semble même évitée (figure 4.9A). En ce qui concerne les autres marques, les courbes sont similaires aux courbes obtenues avec les CRM non blastoderme. Le contrôle construit à partir d'un groupe de gènes choisis au hasard montre clairement un évitement de ces marques. Ceci suggère que les gènes AGZ sont actifs, mais que les CRM AGZ, à l'inverse des CRM blastoderme, ne correspondent pas à des promoteurs actifs. Ceci suggère que les CRM correspond plutôt à des régions enhanceurs.

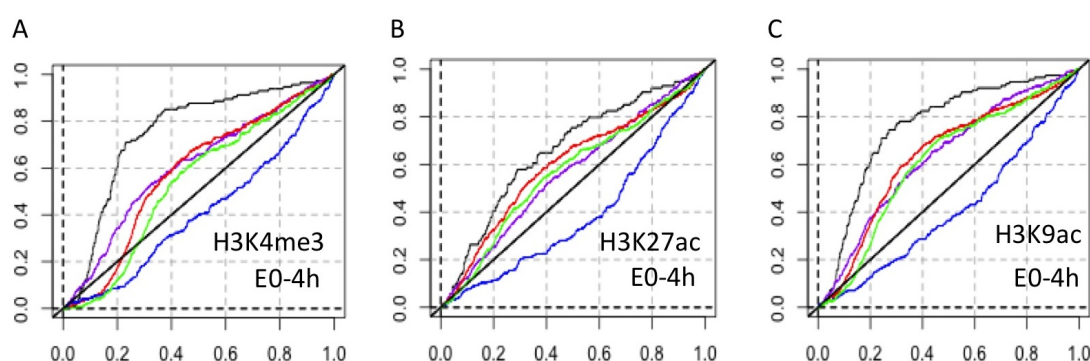


FIGURE 4.9: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques associées au promoteurs actifs. - A-C. La protéine analysée ainsi que l'intervalle temporel de chaque expérience est indiquée dans le coin bas à droite de chaque graphe. Le code couleur des courbes est identique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

L'analyse des régions non codantes des gènes ZGA a mis en évidence les motifs de liaison de Trl. L'enrichissement n'est pas aussi important que celui observé pour d'autres marques (CBP ou H3K4me1), mais l'enrichissement est clairement spécifique aux les CRM AGZ (figure 4.10A), ce qui confirme la pertinence du motif de liaison de Trl découvert plus haut. Les intervalles temporels utilisés pour l'étude de la location génomique de Trl sont relativement longs (8h). Cette faible résolution temporelle pourrait diluer le signal, puisque les embryons sont récoltés sur l'ensemble de la période. De plus, notre analyse révèle une diminution de l'enrichissement de la liaison de Trl (Figure 34B) dans les régions définies par les CRM ZGA au cours du temps. Sur la base des données disponibles, nous ne pouvons pas déterminer à partir de quel moment Trl pourrait cesser de se lier aux CRM AGZ.

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

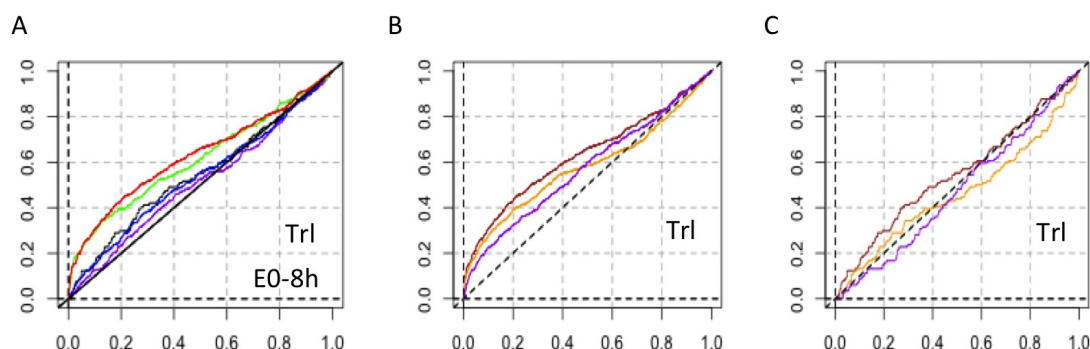


FIGURE 4.10: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles de la liaison de Trl. - Le code couleur des courbes est identique à celui des figures 4.7 A et D. B-C. Résultats d'analyses de Trl au cours du temps pour les CRM AGZ (B) et les CRM blastoderme (C). Les intervalles temporels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

Marques liées à l'état répressif de la chromatine

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

Combinaison de marques associées spécifiquement au CRM AGZ

Comme indiqué plus haut, les densités intégrées de reads provenant des expériences de liaison de Trl, du recrutement de CBP, d'occupation d'H3K4me1, et d'accessibilité de chromatine permettent de discriminer clairement les CRM AGZ des régions non codantes choisies

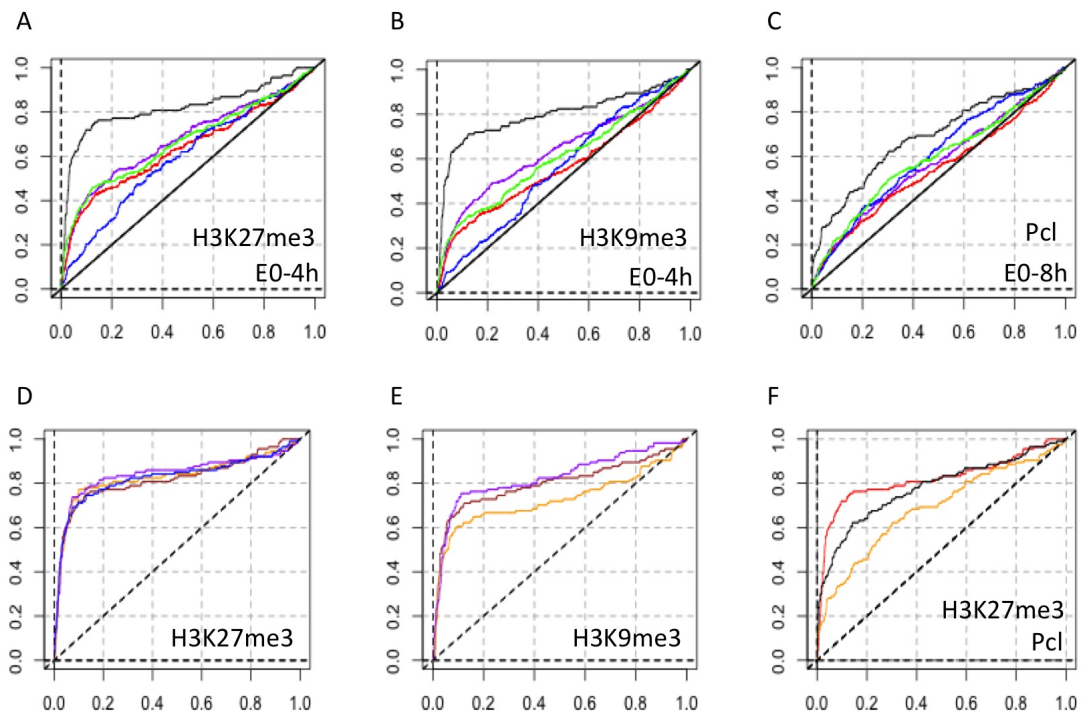


FIGURE 4.11: Courbes de ROC représentant l'enrichissement des CRM AGZ et contrôles en marques répressives. - A-C. La protéine analysée ainsi que l'intervalle temporel de chaque expérience est indiquée dans le coin bas à droite de chaque graphe. Le code couleur des courbes est identique à celui des figures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

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

au hasard. Pour affiner cette constatation, j'ai analysé séquentiellement la combinaison de ces différentes marques. Une combinaison est considérée comme spécifique dans la mesure où elle améliore le classement de ces CRM par rapport au meilleur classement individuel (cf. 4.2.2.3).

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

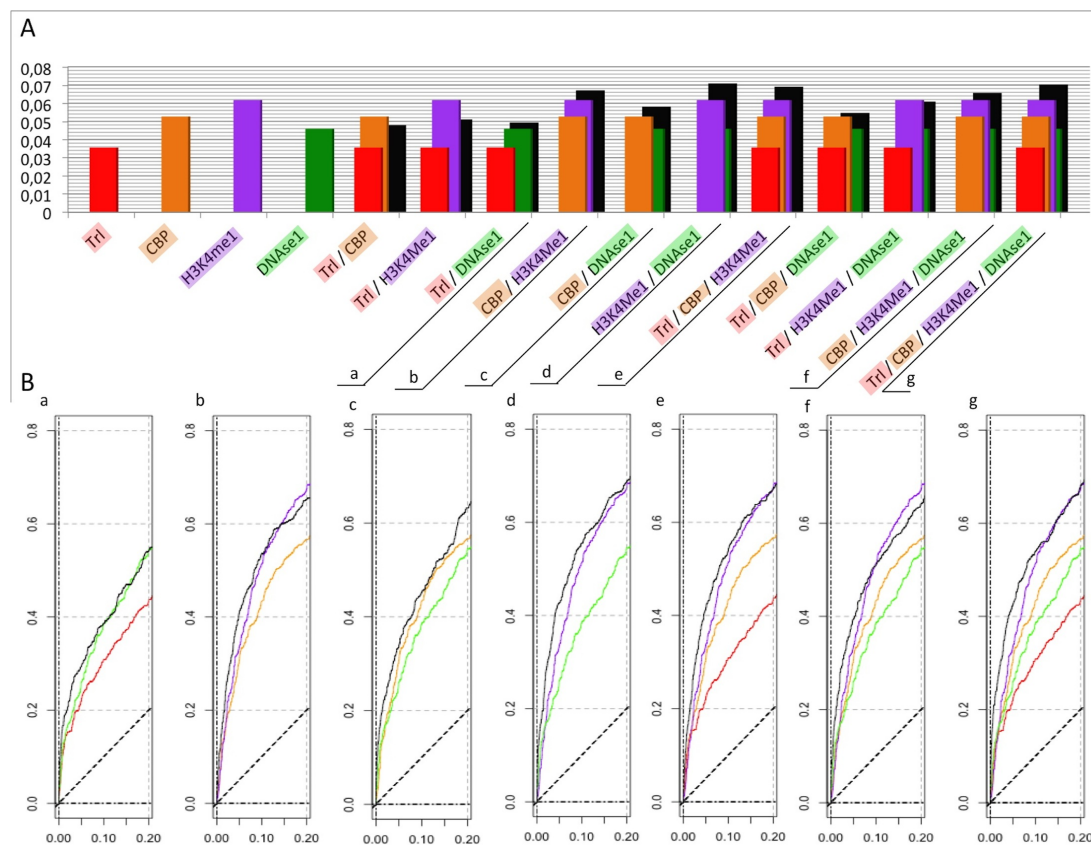


FIGURE 4.12: Résultats de l'analyse de la combinaison de différentes marques. - A. Distribution des AUC. L'abscisse indique les combinaisons analysées. B. Agrandissement de courbes de ROC pour l'analyse du classement contenant 20% des régions aléatoires. Dans les deux panels, les couleurs rouge, orange,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

4.2.3.2 Localisation de Zelda

L'implication du FT Zelda dans l'AGZ a été largement étudiée. Les résultats obtenus à partir des profils de liaison de Zelda (figure 4.13B) confirme sa liaison spécifique dans les CRM blastoderme. La présence d'un motif ressemblant au motif de liaison de Zelda permet de prédire des sites effectivement reconnus par Zelda. Les densités de reads intégrées calculées à partir des données d'expérience de localisation de Zelda apparaissent être le meilleur critère de discrimination de CRM AGZ (figure 4.13A). La combinaison de Zelda et H3K4me1 améliore légèrement ce classement (figure 4.13D).

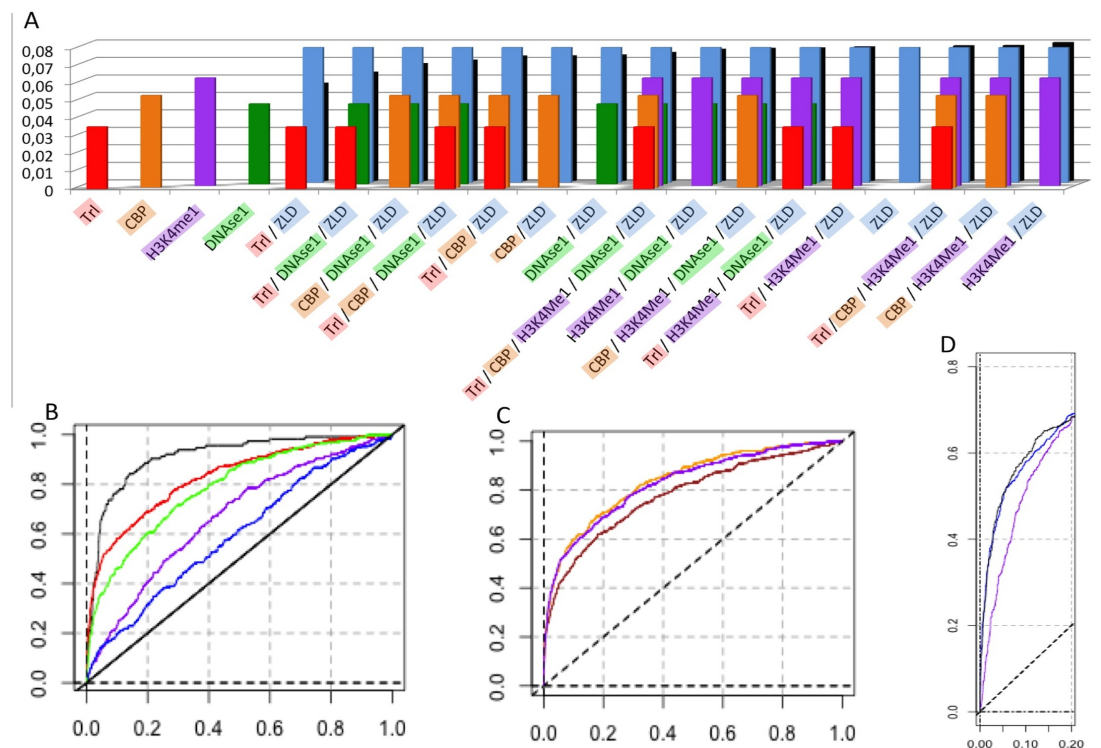


FIGURE 4.13: Analyse de l'enrichissement des CRM AGZ et contrôles de la liaison de Zelda.

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

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

4.3 Conclusion du chapitre

L'analyse des données venant d'expériences de localisation à grande échelle nous a permis de mettre en évidence que les CRM AGZ sont majoritairement localisés dans des régions accessibles de la chromatine. Trl s'est également révélé enrichi dans les CRM AGZ même si sa spécificité est relativement faible. La mauvaise résolution des intervalles temporels pourrait diluer le signal de liaison de Trl au moment de l'AGZ (figure 4.14). De plus, l'association de Trl avec différentes marques améliore le classement des CRM AGZ parmi les régions aléatoires. Des associations spécifiques entre les CRM AGZ et la présence d'histone H3K4me1 et de CBP, qui sont deux marqueurs associés aux régions activatrices de la transcription, ont été révélées et la présence prédominante de Zelda entre 1h et 3h conforte la pertinence des CRM prédits. La découverte de motifs dans les pics de liaison de Trl et Zelda ainsi que d'ouverture de la chromatine et d'occupation de H3K4me1 a mis chevauchant les régions non-codantes des gènes AGZ en évidence les motifs de liaison de Zelda, de Trl et du motif découvert "CAGATACA". Ceci suggère que ces motifs sont fortement liés aux régions activatrices accessibles de la chromatine durant l'AGZ.

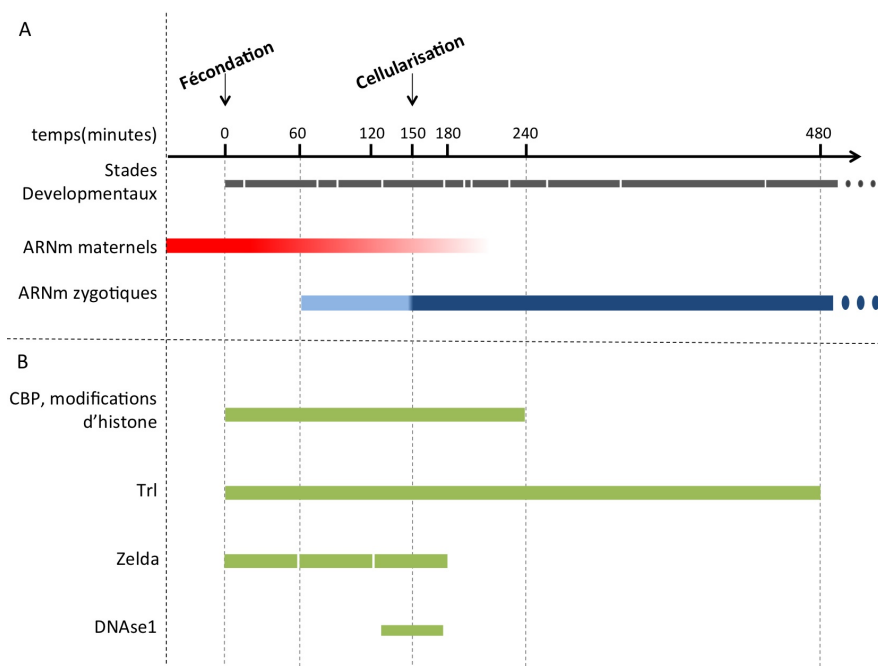


FIGURE 4.14: Intervalles temporels de d'expériences de localisation génomique de facteurs, d'histones et d'accessibilité de la chromatine. - A. L'axe horizontal représente le temps en minutes après la fécondation. Les stades développementaux sont indiqués en gris sous l'axe temporel. Le gradient rouge représente la variation d'abondance des ARNm maternels au cours du temps. L'axe bleu correspond aux ARNm zygotiques synthétisés durant la première (bleu clair) et seconde (bleu foncé) vagues d'activation du génome zygotique.

4. ANALYSE GÉNOMIQUE DES PROFILS D'OCCUPATION ET D'ACCESSIBILITÉ DE LA CHROMATINE

5

Conclusion

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

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

5.1.1 Avant l'AGZ

Durant la première heure de développement, la transcription du génome zygotique est silencieuse. Nous avons vu dans l'introduction que ceci peut être dû à différents facteurs : cycles mitotiques rapides, répression par des facteurs maternel, compaction de la chromatine. Pagans et collaborateurs (2002) (94) ont montré que TTK peut inhiber l'action de Trl via une interaction protéine-protéine directe (figure 5.1 Ab) dans le cadre de la régulation de l'expression du

5. CONCLUSION

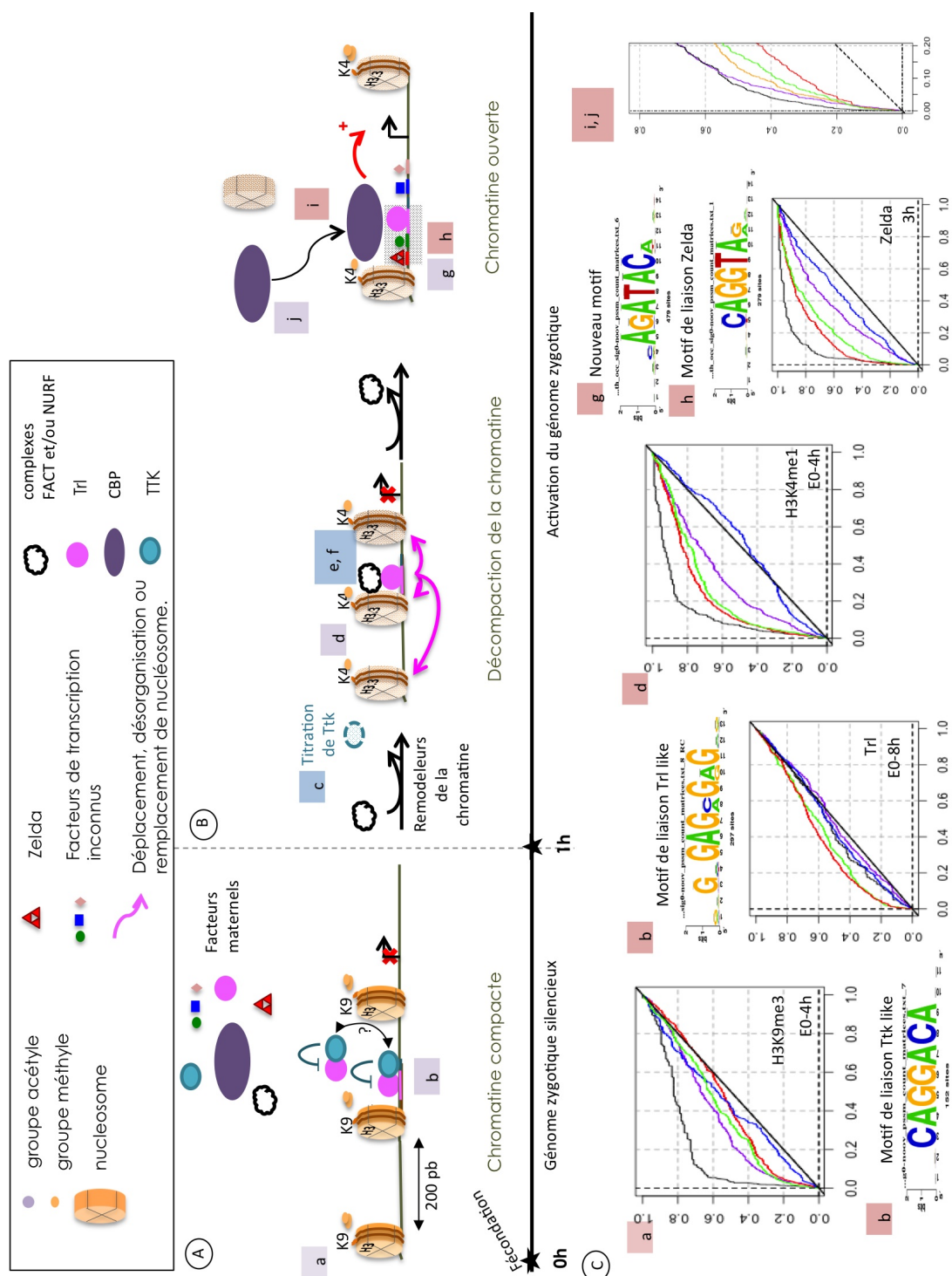


FIGURE 5.1: Présentation d'un modèle spéculatif pour la régulation de l'activation transcriptionnelle du génome zygotique (AGZ). - A. Modèle de l'état de la chromatine et du comportement de différents acteurs avant l'AGZ. B. Modèle d'ouverture de la chromatine et de l'activation de la transcription pendant l'AGZ. C. Indications apportées par mes analyses. Les lettres indiquent les événements à discuter (cf. texte) qui ont été apportés par la recherche bibliographique (bleu), par mes analyses (rouge) ou par les deux (violet). Les détails de la légende sont affichés dans la figure.

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

gène *even-skipped*. La présence de motifs ressemblant aux motifs de liaison de Trl et TTK dans les régions non-codantes AGZ (figure 5.1 Cb) suggèrent que ce mécanisme pourrait être plus général. La cooccurrence de sites prédits pour Trl et TTK est seulement présente dans 22% des CRM comportant Trl et/ou TTK. Cependant Pagans et al. (2002) n'ont pas réussi à déterminer si TTK piègeait Trl en dehors de la chromatine ou non.

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

5.1.2 Activation de la transcription

5.1.2.1 Fin de la répression

La répression de la transcription jusqu'à l'AGZ peut-être due au moins aux deux phénomènes discutés précédemment. Tout d'abord, si TTK est effectivement impliqué dans la répression des gènes AGZ inactivant Trl, nous pouvons imaginer que lorsque TTK est titré par le ratio NC augmentant au fil des cycles mitotiques (69), il libère Trl qui peut alors agir (figure 5.1 Bc). De plus, les données de RNA-seq de Gelbart et Emmert (2010) (93) montrent que l'abondance de Trl augmente entre 2 et 4h. TTK pourrait alors réprimer Trl tant que sa concentration est relativement faible. Il pourrait y avoir une complémentarité entre la titration de TTK et l'augmentation de Trl pour la régulation de l'activation des gènes AGZ soumise au ratio NC ou non, respectivement. Ensuite, l'activation est probablement en partie due à la décompaction de la chromatine. Trl est impliqué dans le recrutement de divers complexes de remodelage de la chromatine (figure 5.1 Be,f). En effet, Nakayama et al. (2007) (124) ont montré que Trl

5. CONCLUSION

interagissait avec le complexe FACT pour le remplacement ciblé des histones H3K9me3 par un variant H3.3 associé aux régions actives de la chromatine et est largement distribué dans les promoteurs des gènes dans l'embryon précoce (125). Les variants H3.3 sont préférentiellement méthylés sur leur lysine 4 (K4). Nous avons identifié un enrichissement en H3K4me1 dans les CRM AGZ (figure 5.1 Bd) qui pourrait être relié à la présence de ces variants. D'autre part, Trl est connu pour remodeler les nucléosomes de façon dépendante à l'ATP en collaborant avec le complexe NURF (126) produisant ainsi des régions chromatiniennes accessibles. Enfin, l'ensemble des protéines formant ces complexes est présent dans l'embryon entre 2 et 4h (93).

Une fois la chromatine accessible, les différents facteurs activateurs pourraient venir se lier à leur séquences cibles (figure 5.1 Bg, h). La présence du motif de Zelda, de Trl et du motif découvert "aAGATACa" sont apparus non seulement sur-représentés dans les régions non codantes des gènes AGZ mais aussi dans les pics chevauchant ces mêmes régions. Les autres motifs sont peut-être facultatifs. Le modèle proposé implique (au moins) ces trois motifs dans le recrutement de Zelda, Trl est d'un facteur inconnu. Les CRM ont montré un enrichissement spécifique pour la combinaison de Trl, CBP, H3K4me1 et de chromatine ouverte (figure 5.1 Bi, Ci,j). Ainsi, la combinaison de ces marques pourrait s'expliquer par la liaison des différents facteurs cités ci-dessus dans une région accessible de la chromatine. Ces facteurs pourraient recruter CBP. Le rôle de CBP a déjà été discuté précédemment dans le chapitre 4. Brièvement, CBP est recruté par des centaines de facteurs différents, probablement pas tous identifiés. CBP peut avoir un rôle acétyl-transférase d'histone, mais les CRM ne sont pas enrichis en histone H3 acétylés (K9 et K27). Dans ce cas, CBP pourrait jouer un rôle de pont entre les facteurs de transcription et la machinerie basale de transcription. Les derniers travaux concernant la localisation de Zelda (88) montrent clairement que ce facteur joue un rôle prépondérant et général dans l'AGZ. Le modèle proposé pourrait ainsi être applicable à l'activation de la transcription en plusieurs vagues.

5.1.3 Rôles de Trl et modèles alternatifs

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

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

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

Trl possède un domaine riche en glutamine (Q-rich), qui, dans d'autres protéines telles que CREB ou SP1, est connu pour stimuler la transcription en interagissant avec TAF3, un composant du complexe TFIID (127). Vaquero et al. (2008) (89) ont montré que Trl augmente la transcription en stabilisant le PIC et en promouvant la réinitiation de la transcription via son domaine Q-rich. Chez la souris, un homologue de Trl, nommé facteur GAGA, est impliqué, en collaboration avec SP1, dans l'activation de *hsp70.1* au début de l'AGZ (128).

5.1.3.2 Trl pourrait recruter CBP

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

5.1.3.3 Trl est impliqué dans la pause des ARN polymérase II

Il a été montré que Trl est impliqué dans la pause de l'ARNpolIII en coopérant avec NELF au promoteur de *hsp70* et d'un grand nombre d'autres gènes (129). Les ARNpolIII en pause ont été détectées dans environ 1500 gènes des embryons de drosophile, incluant un grand nombre de gènes contrôlant le développement (130). Hendrix et collaborateurs ont suggéré que Trl induirait une conformation ouverte de la chromatine, permissive pour le recrutement et la pause d'activité de la ARNpolIII en recrutant NURF par exemple (126). En corrélation avec ce rôle hypothétique de Trl, une forte association entre les CRM et les régions de chromatine liées à Trl a été révélée (figure 4.10). De plus, en analysant le signal de liaison de Trl dans le CRM par rapport à leur position relative au TSS, un pic de signal est observé 200 pb autour du TSS (figure 5.2A). Comme le montre la figure 5.2B, les sites prédits avec la matrice ressemblant au motif de liaison de Trl sont concentrés entre 50 et 250 pb en amont du TSS.

5.1.4 Utilisation du modèle pour la sélection de CRM

Afin de sélectionner les CRM les plus pertinents en vue d'une validation expérimentale, j'ai utilisé le classement des CRM obtenus combinant les classements des marques individuelles.

5. CONCLUSION

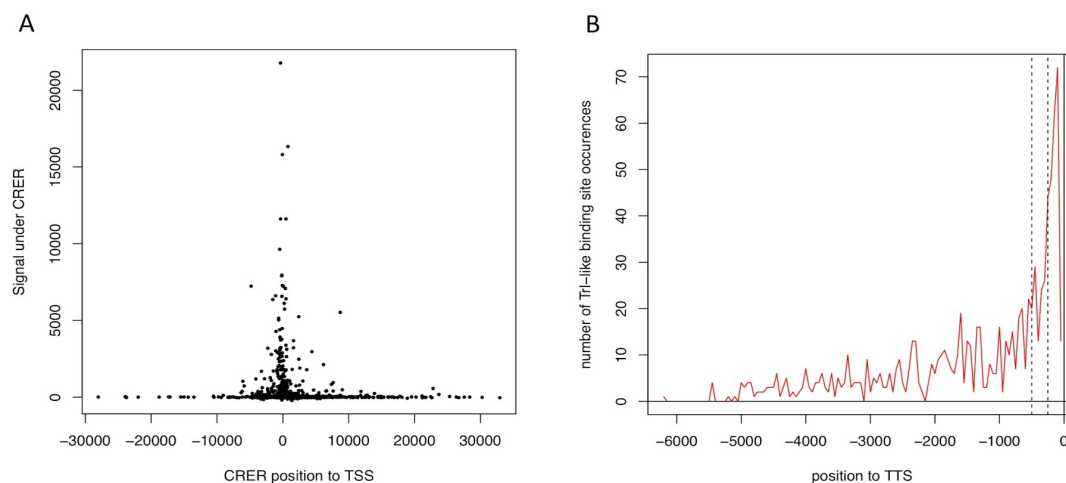


FIGURE 5.2: Analyse positionnelle de l'enrichissement des CRM en signaux de liaison de Trl et sites Trl par rapport au TSS. - A. Chaque point représente un CRM ; l'abscisse indique sa position relative au TSS et l'ordonnée la densité intégrée de reads ChIP-seq Trl sous ce CRM. B. L'abscisse indique la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

J'ai ensuite procédé à la visualisation de ces régions grâce au navigateur de génome du UCSC (<http://genome.ucsc.edu/>) pour avoir une vue d'ensemble des régions.

La figure 5.3 montre la région d'environ 6Kb en amont du gène *crocodile*, impliqué dans la spécification du segment le plus antérieur des trois segments à l'origine de la tête. Il s'agit d'un gène purement zygotique dont l'activation est soumise au ratio NC. Nous pouvons noter qu'aucun site de liaison de Zelda n'a été prédit dans le premier CRM (à gauche) alors que deux pics de liaison ont été identifiés par ChIP-seq.

5.2 Conclusion générale

Les travaux menés durant cette thèse m'ont permis de proposer un mécanisme de régulation de l'activation du génome zygotique chez la Drosophile. À partir de données transcriptomiques récupérées de publications qui visaient à répondre à différentes questions, j'ai proposé une méthode simple et adaptée à l'analyse de la variation de l'expression des gènes au cours du temps. L'analyse fonctionnelle (profils d'expression, analyse des régions non-codantes, enrichissement en classes fonctionnelles) des différents groupes de gènes co-exprimés m'a permis de sélectionner un groupe étendu de gènes activés durant l'AGZ. Les différents motifs détec-

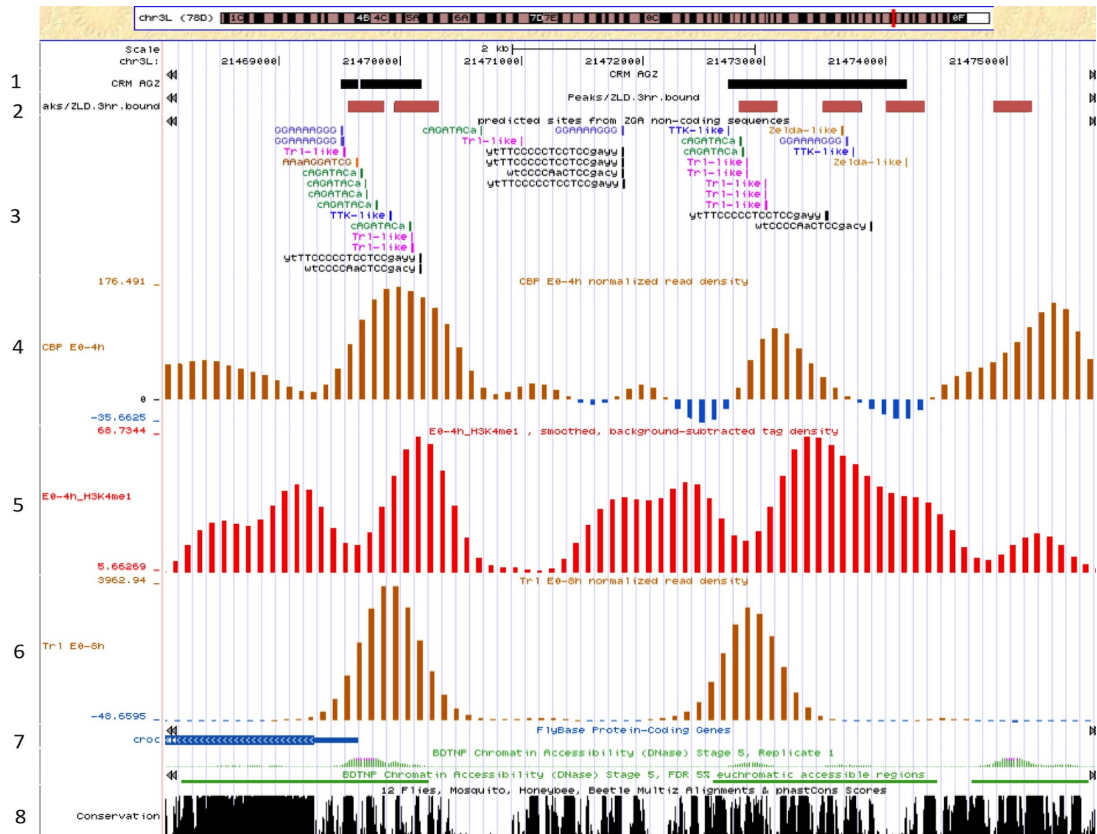


FIGURE 5.3: Visualisation de la région en amont du gène *crocodile*. - 1. CRM AGZ. 2. Pics de liaison de Zelda 3h provenant d'expérience de ChIP-seq. 3. Sites prédits en scannant les séquences avec le motif découvert dans le cluster AZG. 4. Densité normalisée des reads provenant de l'expérience de ChIP-seq de CBF. 5. Densité normalisée des reads provenant de l'expérience de ChIP-seq de H3K4me1. 6. Densité normalisée des reads provenant de l'expérience de ChIP-seq de Tr1. 7. Représentation de la partie 5' du gène (gène sur le brin anti-sens). 7. Ouverture de la chromatine. 8. Conservation entre les espèces de drosophiles.

5. CONCLUSION

tés m'ont permis de proposer des facteurs et co-facteurs agissant potentiellement en trans, et de prédire des régions potentiellement régulatrices. L'analyse de l'association spécifique de ces régions avec des marques chromatiniennes m'a amenée à proposer un modèle combinant l'action de différents facteurs (Zelda, Trl, CBP et inconnus) par leur liaison dans des régions accessibles et actives. L'apport de jeux de données de localisation génomique dans des intervalles temporels plus précis permettrait probablement d'affiner le modèle proposé ici. Enfin, une validation expérimentale permettrait de d'apprécier la pertinence de l'ensemble des prédictions et hypothèses générées durant cette thèse.

References

- [1] N WOYCHIK AND MICHAEL HAMPSEY. **The RNA Polymerase II Machinery Structure Illuminates Function.** *Cell*, **108**(4):453–463, February 2002. 2
- [2] J. E.F. BUTLER AND JAMES T. KADONAGA. **Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs.** *Genes & Development*, **15**:2515–2519, October 2001. 2
- [3] G D STORMO. **DNA binding sites : representation and discovery.** *Bioinformatics (Oxford, England)*, **16**(1):16–23, January 2000. 3, 6
- [4] D J GALAS AND A SCHMITZ. **DNase footprinting : a simple method for the detection of protein-DNA binding specificity.** *Nucleic Acids Research*, **5**(9):3157–3170, September 1978. 3
- [5] A M MAXAM AND W GILBERT. **A new method for sequencing DNA.** *Proceedings of the National Academy of Sciences of the United States of America*, **74**(2):560–564, February 1977. 3
- [6] R STOLTENBURG, C REINEMANN, AND B STREHLITZ. **SELEX, A (r)evolutionary method to generate high-affinity nucleic acid ligands.** *Biomolecular Engineering*, **24**(4):381–403, October 2007. 3
- [7] J. WANG, J. LU, G. GU, AND Y. LIU. **In vitro DNA-binding profile of transcription factors : methods and new insights.** *Journal of Endocrinology*, **210**(1):15–27, March 2011. 5
- [8] XIANGDONG MENG, MICHAEL H BRODSKY, AND SCOT A WOLFE. **A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors.** *Nature Biotechnology*, **23**(8):988–994, August 2005. 5
- [9] XIANGDONG MENG AND SCOT A WOLFE. **Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system.** *Nature Protocols*, **1**(1):30–45, June 2006. 5
- [10] MARTHA L BULYK. **Discovering DNA regulatory elements with bacteria.** *Nature Biotechnology*, **23**(8):942–944, August 2005. 5
- [11] MICHAEL J BUCK AND JASON D LIEB. **ChIP-chip : considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.** *Genomics*, **83**(3):349–360, March 2004. 5
- [12] RAJA JOTHI, SURESH CUDDAPAH, ARTEM BARSKI, KAIRONG CUI, AND KEJI ZHAO. **Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data.** *Nucleic acids research*, **36**(16):5221, 2008. 5
- [13] DUSAN ŠTANOJEVIĆ, TIMOTHY HOEY, AND MICHAEL LEVINE. **Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Krüppel in Drosophila.** *Nature*, **341**(6240):331–335, September 1989. 6
- [14] W H DAY AND F R MCMORRIS. **Critical comparison of consensus methods for molecular sequences.** *Nucleic Acids Research*, **20**(5):1093–1099, March 1992. 6, 12
- [15] GAVIN E CROOKS, GARY HON, JOHN-MARC CHANDONIA, AND STEVEN E BRENNER. **WebLogo : a sequence logo generator.** *Genome Research*, **14**(6):1188–1190, June 2004. 6
- [16] T D SCHNEIDER AND R M STEPHENS. **Sequence logos : a new way to display consensus sequences.** *Nucleic Acids Research*, **18**(20):6097–6100, October 1990. 7
- [17] G Z HERTZ AND G D STORMO. **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics (Oxford, England)*, **15**(7-8):563–577, August 1999. 7, 14
- [18] T.L. BAILEY, MIKAEL BODEN, F.A. BUSKE, MARTIN FRITH, C.E. GRANT, LUCA CLEMENTI, JINGYUAN REN, W.W. LI, AND W.S. NOBLE. **MEME SUITE : tools for motif discovery and searching.** *Nucleic Acids Research*, **37**(suppl 2):W202–W208, 2009. 9, 71
- [19] G Z HERTZ, G W 3RD HARTZELL, AND G D STORMO. **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Computer Applications in the Biosciences : CABIOS*, **6**(2):81–92, April 1990. 9
- [20] MATTHIEU DEFRANCE AND J. VAN HELDEN. **info-gibbs : a motif discovery algorithm that directly optimizes information content during sampling.** *Bioinformatics*, **25**(20):2715–2722, 2009. 9
- [21] GERT THUIS, KATHLEEN MARCHAL, MAGALI LESCOT, STEPHANE ROMBAUTS, BART DE MOOR, PIERRE ROUZÉ, AND YVES MOREAU. **A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes.** *Journal of Computational Molecular Cell Biology*, **9**(2):447–464, 2002. 14
- [22] JEAN-VALÉRY TURATSINZE, MORGANE THOMAS-CHOLLIER, MATTHIEU DEFRANCE, AND JACQUES VAN HELDEN. **Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules.** *Nature protocols*, **3**(10):1578–88, January 2008. 14, 17, 74, 79
- [23] M.C. FRITH, M.C. LI, AND ZHIPING WENG. **Cluster-Buster : Finding dense clusters of motifs in DNA sequences.** *Nucleic acids research*, **31**(13):3666–3668, 2003. 14, 17
- [24] BENJAMIN P BERMAN, BARRET D PFEIFFER, TODD R LAVERTY, STEVEN L SALZBERG, GERALD M RUBIN, MICHAEL B EISEN, AND SUSAN E CELNIKER. **Computational identification of developmental enhancers : conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.** *Genome biology*, **5**(9):R61, January 2004. 14, 17, 31
- [25] ANTHONY A PHILIPPAKIS, FANGXUE SHERRY HE, AND MARTHA L BULYK. **Modulefinder : a tool for computational discovery of cis regulatory modules.** *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 519–30, January 2005. 17
- [26] A. SOSINSKY. **Target Explorer : an automated tool for the identification of new target genes for a specified set of transcription factors.** *Nucleic Acids Research*, **31**(13):3589–3592, July 2003. 17
- [27] IAN JOHN DONALDSON, MICHAEL CHAPMAN, AND BERTHOLD GÖTTGENS. **TFBScluster : a resource for the characterization of transcriptional regulatory networks.** *Bioinformatics (Oxford, England)*, **21**(13):3058–9, July 2005. 17

REFERENCES

- [28] BENJAMIN P. BERMAN, YUTAKA NIBU, BARRET D PFEIFFER, PAVEL TOMANCAK, SUSAN E CELNIKER, MICHAEL LEVINE, GERALD M RUBIN, AND MICHAEL B EISEN. **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome.** *Proceedings of the National Academy of Sciences of the United States of America*, **99**(2):757–62, January 2002. 17, 31
- [29] YUTAO FU AND ZHIPING WENG. **Improvement of TRANSFAC matrices using multiple local alignment of transcription factor binding site sequences.** *Genome informatics. International Conference on Genome Informatics*, **16**(1):68–72, January 2005. 17
- [30] ELODIE PORTALES-CASAMAR, SUPAT THONGJUEA, ANDREW T KWON, DAVID ARENILLAS, XIAOBEI ZHAO, EIVIND VALEN, DIMAS YUSUF, BORIS LENHARD, WYETH W WASSERMAN, AND ALBIN SANDELIN. **JASPAR 2010 : the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic acids research*, **38**(Database issue):D105–10, January 2010. 17, 64, 80
- [31] LIHUA JULIE ZHU, RYAN G CHRISTENSEN, MAJID KAZEMIAN, CHRISTOPHER J HULL, METEWO SELASE ENUAMEH, MATTHEW D BASCIOTTA, JESSIE A BRASEFIELD, CONG ZHU, YUNA ASRIYAN, DAVID S LAPOINTE, SAURABH SINHA, SCOT A WOLFE, AND MICHAEL H BRODSKY. **FlyFactorSurvey : a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system.** *Nucleic acids research*, **39**(Database issue):D111–7, January 2011. 17, 64, 79
- [32] IVAN V KULAKOVSKIY, ALEXANDER V FAVOROV, AND VSEVOLOD J MAKEEV. **Motif discovery and motif finding from genome-mapped DNase footprint data.** *Bioinformatics (Oxford, England)*, **25**(18):2318–25, September 2009. 17
- [33] SOCORRO GAMA-CASTRO, HELADIA SALGADO, MARTIN PERALTA-GIL, ALBERTO SANTOS-ZAVALA, LUIS MUÑOZ RASCADO, HILDA SOLANO-LIRA, VERÓNICA JIMENEZ-JACINTO, VERENA WEISS, JAIR S GARCÍA-SOTELO, ALEJANDRA LÓPEZ-FUENTES, LILIANA PORRÓN-SOTELO, SHIRLEY ALQUICIRA-HERNÁNDEZ, ALEJANDRA MEDINA-RIVERA, IRMA MARTÍNEZ-FLORES, KEVIN ALQUICIRA-HERNÁNDEZ, RUTH MARTÍNEZ-ADAME, CÉSAR BONAVIDES-MARTÍNEZ, JUAN MIRANDA-RÍOS, ARACELI M HUERTA, ALFREDO MENDOZA-VARGAS, LEONARDO COLLADO-TORRES, BLANCA TABOADA, LETICIA VEGA-ALVARADO, MARICELA OLVERA, LETICIA OLVERA, RICARDO GRANDE, ENRIQUE MORETT, AND JULIO COLLADO-VIDES. **RegulonDB version 7.0 : transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units).** *Nucleic acids research*, **39**(Database issue):D98–105, January 2011. 18
- [34] ELODIE PORTALES-CASAMAR, STEFAN KIROV, JONATHAN LIM, STUART LITHWICK, MAGDALENA I SWANSON, AMY TICOLL, JAY SNODDY, AND WYETH W WASSERMAN. **PAZAR : a framework for collection and dissemination of cis-regulatory sequence annotation.** *Genome biology*, **8**(10):R207, January 2007. 19
- [35] ELODIE PORTALES-CASAMAR, DAVID ARENILLAS, JONATHAN LIM, MAGDALENA I SWANSON, STEVEN JIANG, ANTHONY MCCALLUM, STEFAN KIROV, AND WYETH W WASSERMAN. **The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences.** *Nucleic acids research*, **37**(Database issue):D54–60, January 2009. 19
- [36] CASEY M BERGMAN, JOSEPH W CARLSON, AND SUSAN E CELNIKER. ***Drosophila* DNase I footprint database : a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*.** *Bioinformatics*, **21**(8):1747–1749, November 2004. 19, 80
- [37] THOMAS A DOWN, CASEY M BERGMAN, JING SU, AND TIM J P HUBBARD. **Large-scale discovery of promoter motifs in *Drosophila melanogaster*.** *PLoS computational biology*, **3**(1):e7, January 2007. 19, 80
- [38] MARC S HALFON, STEVEN M GALLO, AND CASEY M BERGMAN. **REDfly 2.0 : an integrated database of cis-regulatory modules and transcription factor binding sites in *Drosophila*.** *Nucleic acids research*, **36**(Database issue):D594–8, January 2008. 19
- [39] MARCUS B NOYES, RYAN G CHRISTENSEN, ATSUYA WAKABAYASHI, GARY D STORMO, MICHAEL H BRODSKY, AND SCOT A WOLFE. **Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites.** *Cell*, **133**(7):1277–89, June 2008. 19, 79
- [40] MARCUS B NOYES, XIANGDONG MENG, ATSUYA WAKABAYASHI, SAURABH SINHA, MICHAEL H BRODSKY, AND SCOT A WOLFE. **A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system.** *Nucleic acids research*, **36**(8):2547–60, May 2008. 19, 79
- [41] ALAN M MOSES, DANIEL A POLLARD, DAVID A NIX, VENKY N IYER, XIAO-YONG LI, MARK D BIGGIN, AND MICHAEL B EISEN. **Large-scale turnover of functional transcription factor binding sites in *Drosophila*.** *PLoS computational biology*, **2**(10):e130, October 2006. 19
- [42] XIAO-YONG LI, STEWART MACARTHUR, RICHARD BOURGON, DAVID NIX, DANIEL A POLLARD, VENKY N IYER, AARON HECHMER, LISA SIMIRENKO, MARK STAPLETON, CRIS L LUENGO HENDRIKS, HOU CHENG CHU, NOBUO OGAWA, WILLIAM INWOOD, VICTOR SEMENTCHENKO, AMY BEATON, RICHARD WEISZMANN, SUSAN E CELNIKER, DAVID W KNOWLES, TOM GINGERAS, TERENCE P SPEED, MICHAEL B EISEN, AND MARK D BIGGIN. **Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm.** *PLoS biology*, **6**(2):e27, February 2008. 19, 31
- [43] R J KELLEHER, P M FLANAGAN, AND R D KORNBERG. **A novel mediator between activator proteins and the RNA polymerase II transcription apparatus.** *Cell*, **61**(7):1209–15, June 1990. 18
- [44] AMELIA CASAMASSIMI AND CLAUDIO NAPOLI. **Mediator complexes and eukaryotic transcription regulation : an overview.** *Biochimie*, **89**(12):1439–46, December 2007. 18
- [45] FLORIS BOSVELD, SJOERD VAN HOEK, AND ODY C M SIBON. **Establishment of cell fate during early *Drosophila* embryogenesis requires transcriptional Mediator subunit dMED31.** *Developmental biology*, **313**(2):802–13, January 2008. 18, 27, 31
- [46] KAROLIN LUGER, A W MÄDER, R K RICHMOND, D F SARGENT, AND T J RICHMOND. **Crystal structure of the nucleosome core particle at 2.8 Å resolution.** *Nature*, **389**(6648):251–60, September 1997. 18
- [47] SUSUMU HIROSE. **Crucial roles for chromatin dynamics in cellular memory.** *Journal of biochemistry*, **141**(5):615–9, May 2007. 20
- [48] E. B. LEWIS. **A gene complex controlling segmentation in *Drosophila*.** *Nature*, **276**(5688):565–570, December 1978. 20
- [49] CHRISTIANE NÜSSLEIN-VOLHARD AND ERIC WIESCHAUS. **Mutations affecting segment number and polarity in *Drosophila*.** *Nature*, **287**(5785):795–801, October 1980. 20
- [50] WAEL TADROS AND HOWARD D LIPSHITZ. **The maternal-to-zygotic transition : a play in two acts.** *Development (Cambridge, England)*, **136**(18):3033–42, September 2009. 22, 31
- [51] JUSTIN CREST, NATHAN OXNARD, JUN-YUAN JI, AND GEROLD SCHUBIGER. **Onset of the DNA replication checkpoint in the early *Drosophila* embryo.** *Genetics*, **175**(2):567–84, February 2007. 23, 31

REFERENCES

- [52] BEATRICE BENOIT, CHUN HUA HE, FAN ZHANG, SARAH M VOTRUBA, WAEL TADROS, J TIMOTHY WESTWOOD, CRAIG A SMIBERT, HOWARD D LIPSHITZ, AND WILLIAM E THEURKAUF. **An essential role for the RNA-binding protein Smaug during the Drosophila maternal-to-zygotic transition.** *Development (Cambridge, England)*, **136**(6) :923–32, March 2009. 23, 28, 30
- [53] C A SMIBERT, J E WILSON, K KERR, AND P M MACDONALD. **smaug protein represses translation of unlocalized nanos mRNA in the Drosophila embryo.** *Genes & development*, **10**(20) :2600–9, October 1996. 25
- [54] RONGWEN XI, JENNIFER R MCGREGOR, AND DOUGLAS A HARRISON. **A gradient of JAK pathway activity patterns the anterior-posterior axis of the follicular epithelium.** *Developmental cell*, **4**(2) :167–77, February 2003. 26
- [55] JENS JANUSCHKE, LOUIS GERVAIS, LAURENT GILLET, GUY KERYER, MICHEL BORNENS, AND ANTOINE GUICHET. **The centrosome-nucleus complex and microtubule organization in the Drosophila oocyte.** *Development (Cambridge, England)*, **133**(1) :129–39, January 2006. 26
- [56] NATALIE DENEFF AND TRUDI SCHÜPBACH. **Patterning : JAK-STAT signalling in the Drosophila follicular epithelium.** *Current biology : CB*, **13**(10) :R388–90, May 2003. 26
- [57] DIERK NIESSING, STEPHEN BLANKE, AND HERBERT JÄCKLE. **Bicoid associates with the 5'-cap-bound complex of caudal mRNA and represses translation.** *Genes & development*, **16**(19) :2576–82, October 2002. 25
- [58] FANNY PILOT, JEAN-MARC PHILIPPE, CÉLINE LEMMERS, JEAN-PAUL CHAUVIN, AND THOMAS LECUIT. **Developmental control of nuclear morphogenesis and anchoring by charleston, identified in a functional genomic screen of Drosophila cellularisation.** *Development (Cambridge, England)*, **133**(4) :711–23, February 2006. 28, 32, 35, 38, 41, 122
- [59] STEFANO DE RENZIS, OLIVIER ELEMENTO, SAEED TAVAZOIE, AND ERIC F WIESCHAUS. **Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo.** *PLoS biology*, **5**(5) :e117, May 2007. 28, 29, 30, 31, 32, 36, 45, 49, 64, 65
- [60] SEAN D HOOPER, STEPHANIE BOUÉ, ROLAND KRAUSE, LARS J JENSEN, CHRISTOPHER E MASON, MURAD GHANIM, KEVIN P WHITE, EILEEN E M FURLONG, AND PEER BORK. **Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis.** *Molecular systems biology*, **3**(3) :72, January 2007. 28
- [61] ERIC LÉCUYER, HIDEKI YOSHIDA, NEELA PARTHASARATHY, CHRISTINA ALM, TOMAS BABAK, TANJA CEROVINA, TIMOTHY R HUGHES, PAVEL TOMANCAK, AND HENRY M KRAUSE. **Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function.** *Cell*, **131**(1) :174–187, 2007. 28
- [62] XUEMIN LU, JENNIFER M LI, OLIVIER ELEMENTO, SAEED TAVAZOIE, AND ERIC F WIESCHAUS. **Coupling of zygotic transcription to mitotic control at the Drosophila mid-blastula transition.** *Development (Cambridge, England)*, **136**(12) :2101–10, June 2009. 28, 32, 35, 45
- [63] WAEL TADROS, AARON L GOLDMAN, TOMAS BABAK, FIONA MENZIES, LEAH VARDY, TERRY ORR-WEAVER, TIMOTHY R HUGHES, J TIMOTHY WESTWOOD, CRAIG A SMIBERT, AND HOWARD D LIPSHITZ. **SMAUG Is a Major Regulator of Maternal mRNA Destabilization in Drosophila and Its Translation Is Activated by the PAN GU Kinase.** *Developmental Cell*, **12**(1) :143–155, 2007. 28, 29, 30
- [64] STEFAN THOMSEN, SIMON ANDERS, SARATH CHANDRA JANGA, WOLFGANG HUBER, AND CLAUDIO R ALONSO. **Genome-wide analysis of mRNA decay patterns during early Drosophila development.** *Genome biology*, **11**(9) :R93, January 2010. 28, 29
- [65] NATASCHA BUSHATI, ALEXANDER STARK, JULIUS BRENNENCKE, AND STEPHEN M COHEN. **Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in Drosophila.** *Current biology : CB*, **18**(7) :501–6, April 2008. 28
- [66] ANDRÉ P GERBER, STEFAN LUSCHNIG, MARK A KRASNOW, PATRICK O BROWN, AND DANIEL HERSCHLAG. **Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster.** *Proceedings of the National Academy of Sciences of the United States of America*, **103**(12) :4487–92, March 2006. 29
- [67] B A EDGAR AND S A DATAR. **Zygotic degradation of two maternal Cdc25 mRNAs terminates Drosophila's early cell cycle program.** *Genes & development*, **10**(15) :1966–77, August 1996. 29
- [68] B A EDGAR AND P H O'FARRELL. **Genetic control of cell division patterns in the Drosophila embryo.** *Cell*, **57**(1) :177–87, April 1989. 30, 31
- [69] D K PRITCHARD AND G SCHUBIGER. **Activation of transcription in Drosophila embryos is a gradual process mediated by the nucleocytoplasmic ratio.** *Genes & development*, **10**(9) :1131–42, May 1996. 30, 65, 109
- [70] O C SIBON, V A STEVENSON, AND W E THEURKAUF. **DNA-replication checkpoint control at the Drosophila midblastula transition.** *Nature*, **388**(6637) :93–7, July 1997. 31
- [71] D READ, T NISHIGAKI, AND J L MANLEY. **The Drosophila even-skipped promoter is transcribed in a stage-specific manner in vitro and contains multiple, overlapping factor-binding sites.** *Molecular and cellular biology*, **10**(8) :4334–44, August 1990. 31
- [72] GREGORY T REEVES AND ANGELIKE STATHOPOULOS. **Graded dorsal and differential gene regulation in the Drosophila embryo.** *Cold Spring Harbor perspectives in biology*, **1**(4) :a000836, October 2009. 31
- [73] JOHN R TEN BOSCH, JOSEPH A BENAVIDES, AND THOMAS W CLINE. **The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription.** *Development (Cambridge, England)*, **133**(10) :1967–77, May 2006. 31, 64
- [74] HSIAO-LAN LIANG, CHUNG-YI NIEN, HSIAO-YUN LIU, MARK M METZSTEIN, NIKOLAI KIROV, AND CHRISTINE RUSHLOW. **The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila.** *Nature*, **456**(7220) :400–3, December 2008. 31, 64
- [75] AMY TSURUMI, FAN XIA, JINGHONG LI, KIMBERLY LARSON, RUSSELL LAFRANCE, AND WILLIS X LI. **STAT Is an Essential Activator of the Zygotic Genome in the Early Drosophila Embryo.** *PLoS genetics*, **7**(5) :e1002086, May 2011. 31
- [76] ROBERT K BRADLEY, XIAO-YONG LI, COLE TRAPNELL, STUART DAVIDSON, LIOR PACTHER, HOU CHENG CHU, LEATH A TONKIN, MARK D BIGGIN, AND MICHAEL B EISEN. **Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species.** *PLoS biology*, **8**(3) :e1000343, March 2010. 31
- [77] E. HUBBELL, W.-M. LIU, AND R. MEI. **Robust estimators for expression analysis.** *Bioinformatics*, **18**(12) :1585–1592, December 2002. 36
- [78] RAFAEL A IRIZARRY, BENJAMIN M BOLSTAD, FRANCOIS COLLIN, LESLIE M COPE, BRIDGET HOBBS, AND TERENCE P SPEED. **Summaries of Affymetrix GeneChip probe level data.** *Nucleic acids research*, **31**(4) :e15, February 2003. 36, 38

REFERENCES

- [79] W.-M LIU, R. MEI, X. DI, T. B. RYDER, E. HUBBELL, S. DEE, T. A. WEBSTER, C. A. HARRINGTON, M.-H. HO, J. BAID, AND S. P. SMEEKENS. **Analysis of high density expression microarrays with signed-rank call algorithms.** *Bioinformatics*, **18**(12) :1593–1599, December 2002. 36
- [80] STUART D PEPPER, EMMA K SAUNDERS, LAURA E EDWARDS, CLAIRE L WILSON, AND CRISPIN J MILLER. **The utility of MAS5 expression summary and detection call algorithms.** *BMC bioinformatics*, **8**:273, January 2007. 36, 38
- [81] RAFAEL A IRIZARRY, BRIDGET HOBBS, FRANCOIS COLLIN, YASMIN D BEAZER-BARCLAY, KRISTEN J ANTONELLIS, UWE SCHERF, AND TERENCE P SPEED. **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics (Oxford, England)*, **4**(2) :249–64, April 2003. 38
- [82] ZHIJIN WU AND RAFAEL A. IRIZARRY. **Stochastic models inspired by hybridization theory for short oligonucleotide arrays.** *Journal of computational biology : a journal of computational molecular cell biology*, **12**(6) :882–93, July 2005. 38
- [83] BETTINA HARR AND CHRISTIAN SCHLÖTTERER. **Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons.** *Nucleic acids research*, **34**(2) :e8, January 2006. 38
- [84] R DEVELOPMENT CORE TEAM. **R : A language and environment for statistical computing.** pages ISBN 3–900051–07–0, 2008. 38
- [85] ELEANOR HOWE, KRISTINA HOLTON, SARITA NAIR, DANIEL SCHLAUCH, RAKTIM SINHA, AND JOHN QUACKENBUSH. **MeV : MultiExperiment Viewer.** *Biomedical Informatics for Cancer Research*, pages 267–277, 2010. 40
- [86] SYLVAIN BROHÉE, KAROLINE FAUST, GIPSI LIMA-MENDEZ, GILLES VANDERSTOCKEN, AND JACQUES VAN HELDEN. **Network Analysis Tools : from biological networks to clusters and pathways.** *Nature protocols*, **3**(10) :1616–29, January 2008. 57, 67
- [87] MORGANE THOMAS-CHOLLIER, OLIVIER SAND, JEAN-VALÉRY TURATSINZE, REKIN’S JANKY, MATTHIEU DEFRANCE, ERIC VERVISCH, SYLVAIN BROHÉE, AND JACQUES VAN HELDEN. **RSAT : regulatory sequence analysis tools.** *Nucleic acids research*, **36**(Web Server issue) :W119–27, July 2008. 58, 121
- [88] MELISSA M. HARRISON, XIAO-YONG LI, TOMMY KAPLAN, MICHAEL R. BOTCHAN, AND MICHAEL B. EISEN. **Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition.** *PLoS Genetics*, **7**(10) :e1002266, October 2011. 64, 72, 84, 110
- [89] ALEJANDRO VAQUERO, MARTA BLANCH, M LLUÏSA ESPINÁS, AND JORDI BERNUÉS. **Activation properties of GAGA transcription factor.** *Biochimica et biophysica acta*, **1779**(5) :312–7, May 2008. 64, 111
- [90] H GRANOK, B A LEIBOVITCH, C D SHAFFER, AND S C ELGIN. **Chromatin. Ga-ga over GAGA factor.** *Current biology : CB*, **5**(3) :238–41, March 1995. 64
- [91] TSUKASA SHIMOJIMA, MASAHIRO OKADA, TAKAHIRO NAKAYAMA, HITOSHI UEDA, KATSUYA OKAWA, AKIHIRO IWAMATSU, HIROSHI HANDA, AND SUSUMU HIROSE. **Drosophila FACT contributes to Hox gene expression through physical and functional interactions with GAGA factor.** *Genes & development*, **17**(13) :1605–16, July 2003. 64
- [92] MICHAEL LEHMANN. **Anything else but GAGA : a nonhistone protein complex reshapes chromatin structure.** *Trends in genetics : TIG*, **20**(1) :15–22, January 2004. 64
- [93] W.M. GELBART AND D.B. EMMERT. **FlyBase High Throughput Expression Pattern Data Beta Version.**, 2010. 64, 65, 109, 110
- [94] SARA PAGANS, MIGUEL ORTIZ-LOMBARDÍA, MA LLUÏSA ESPINÁS, JORDI BERNUÉS, AND FERNANDO AZORÍN. **The Drosophila transcription factor tramtrack (TTK) interacts with Trithorax-like (GAGA) and represses GAGA-mediated activation.** *Nucleic acids research*, **30**(20) :4406–13, October 2002. 65, 72, 107
- [95] F HIROSE, M YAMAGUCHI, K KURODA, A OMORI, T HACHIYA, M IKEDA, Y NISHIMOTO, AND A MATSUKAGE. **Isolation and characterization of cDNA for DREF, a promoter-activating factor for Drosophila DNA replication-related genes.** *The Journal of biological chemistry*, **271**(7) :3930–7, February 1996. 66
- [96] CRAIG M. HART, OLIVIER CUVIER, AND ULRICH K. LAEMMLI. **Evidence for an antagonistic relationship between the boundary element-associated factor BEAF and the transcription factor DREF.** *Chromosoma*, **108**(6) :375–383, November 1999. 66
- [97] DA WEI HUANG, BRAD T SHERMAN, QINA TAN, JACK R COLLINS, W GREGORY ALVORD, JEAN ROAYAEI, ROBERT STEPHENS, MICHAEL W BASELER, H CLIFFORD LANE, AND RICHARD A LEMPICKI. **The DAVID Gene Functional Classification Tool : a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome biology*, **8**(9) :R183, January 2007. 67
- [98] DA WEI HUANG, BRAD T SHERMAN, AND RICHARD A LEMPICKI. **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nature protocols*, **4**(1) :44–57, January 2009. 67
- [99] DAVID MARTIN, CHRISTINE BRUN, ELISABETH REMY, PIERRE MOURREN, DENIS THIEFFRY, AND BERNARD JACQ. **GOToolBox : functional analysis of gene datasets based on Gene Ontology.** *Genome biology*, **5**(12) :R101, January 2004. 67
- [100] DOUGLAS A HOSACK, GLYNN DENNIS, BRAD T SHERMAN, H CLIFFORD LANE, AND RICHARD A LEMPICKI. **Identifying biological themes within lists of genes with EASE.** *Genome biology*, **4**(10) :R70, January 2003. 68
- [101] STEIN AERTS, XIAO-JIANG QUAN, ANNELES CLAEYS, MARINA NAVAL SANCHEZ, PHILLIP TATE, JIEKUN YAN, AND BASSEM A HASSAN. **Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in Drosophila uncovers a regulatory basis for sensory specification.** *PLoS biology*, **8**(7) :e1000435, January 2010. 71
- [102] LONG LI, QIANQIAN ZHU, XIN HE, SAURABH SINHA, AND MARC S HALFON. **Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses.** *Genome biology*, **8**(6) :R101, January 2007. 71
- [103] V MOREL AND F SCHWEISGUTH. **Repression by suppressor of hairless and activation by Notch are required to define a single row of single-minded expressing cells in the Drosophila embryo.** *Genes & development*, **14**(3) :377–88, February 2000. 74, 131
- [104] OLIVIER SAND, MORGANE THOMAS-CHOLLIER, AND JACQUES VAN HELDEN. **Retrieve-ensembl-seq : user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl.** *Bioinformatics (Oxford, England)*, **25**(20) :2739–40, October 2009. 76
- [105] MATTHIEU DEFRANCE, REKIN’S JANKY, OLIVIER SAND, AND JACQUES VAN HELDEN. **Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences.** *Nature protocols*, **3**(10) :1589–603, January 2008. 78

REFERENCES

- [106] SHERYL T SMITH, SVETLANA PETRUK, YURIH SEDKOV, ELIZABETH CHO, SERGEI TILLIB, ELI CANAANI, AND ALEXANDER MAZO. **Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex.** *Nature cell biology*, **6**(2) :162–7, February 2004. 83
- [107] M. MANNERVIK. **Transcriptional Coregulators in Development.** *Science*, **284**(5414) :606–609, April 1999. 83
- [108] TOBIAS LILJA, HITOSHI AIHARA, MARIANNE STABELL, YUTAKA NIBU, AND MATTIAS MANNERVIK. **The acetyltransferase activity of Drosophila CBP is dispensable for regulation of the Dpp pathway in the early embryo.** *Developmental biology*, **305**(2) :650–8, May 2007. 83
- [109] HIROSHI AKIMARU, YANG CHEN, PING DAI, D X HOU, MAKI NONAKA, SARAH M. SMOLIK, STEVE ARMSTRONG, RICHARD H. GOODMAN, AND SHUNSUKE ISHII. **Drosophila CBP is a co-activator of cubitus interruptus in hedgehog signalling.** *Nature*, **386**(6626) :735–8, April 1997. 83
- [110] H AKIMARU, D X HOU, AND S ISHII. **Drosophila CBP is required for dorsal-dependent twist gene expression.** *Nature genetics*, **17**(2) :211–4, October 1997. 83
- [111] L WALTZER AND M BIENZ. **Drosophila CBP represses the transcription factor TCF to antagonize Wingless signalling.** *Nature*, **395**(6701) :521–5, October 1998. 83
- [112] HYE-KYUNG LEE, UI-HYUN PARK, EUN-JOO KIM, AND SOO-JONG UM. **MED25 is distinct from TRAP220/MED1 in cooperating with CBP for retinoid receptor activation.** *The EMBO journal*, **26**(15) :3545–57, August 2007. 83
- [113] N VO AND R H GOODMAN. **CREB-binding protein and p300 in transcriptional regulation.** *The Journal of biological chemistry*, **276**(17) :13505–8, April 2001. 83
- [114] FENG TIE, RAKHEE BANERJEE, CARL A. STRATTON, JAYASHREE PRASAD-SINHA, VINCENT STEPANIK, ANDREI ZLOBIN, MANUEL O. DIAZ, PETER C. SCACHERI, AND PETER J. HARTE. **CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing.** *Development (Cambridge, England)*, **136**(18) :3131–41, September 2009. 84
- [115] NICHOLAS T CRUMP, CATHERINE A HAZZALIN, ERIN M BOWERS, RHODA M ALANI, PHILIP A COLE, AND LOUIS C MAHADEVAN. **Dynamic acetylation of all lysine-4 trimethylated histone H3 is evolutionarily conserved and mediated by p300/CBP.** *Proceedings of the National Academy of Sciences of the United States of America*, **108**(19) :7814–9, May 2011. 84
- [116] NICOLAS NÈGRE, CHRISTOPHER D BROWN, LIJIA MA, CHRISTOPHER AARON BRISTOW, STEVEN W MILLER, ULRICH WAGNER, POUYA KHERADPOUR, MATTHEW L EATON, PAUL LORIAUX, RACHEL SEALFON, ZIRONG LI, HARUHIKO ISHII, REBECCA F SPOKONY, JIA CHEN, LINDSAY HWANG, CHAO CHENG, RICHARD P AUBURN, MELISSA B DAVIS, MARC DOMANUS, PARANTU K SHAH, CAROLYN A MORRISON, JENNIFER ZIEBA, SARAH SUCHY, LIONEL SENDEROVICZ, ALEC VICTORSEN, NICHOLAS A BILD, A JASON GRUNDTAD, DAVID HANLEY, DAVID M MACALPINE, MATTIAS MANNERVIK, KOEN VENKEN, HUGO BELLEN, ROBERT WHITE, MARK GERSTEIN, STEVEN RUSSELL, ROBERT L GROSSMAN, BING REN, JAMES W POSAKONY, MANOLIS KELLIS, AND KEVIN P WHITE. **A cis-regulatory map of the Drosophila genome.** *Nature*, **471**(7339) :527–31, March 2011. 84, 98
- [117] MAXIM NEKRASOV, TETYANA KLYMENKO, SVEN FRATERMAN, BERNADETT PAPP, KATARZYNA OKTABA, THOMAS KÖCHER, ADRIAN COHEN, HENDRIK G STUNNENBERG, MATTHIAS WILM, AND JÜRIG MÜLLER. **Pc1-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes.** *The EMBO journal*, **26**(18) :4078–88, September 2007. 84
- [118] SEAN THOMAS, XIAO-YONG LI, PETER J SABO, RICHARD SANDSTROM, ROBERT E THURMAN, THERESA K CANFIELD, ERIKA GISTE, WILLIAM FISHER, ANN HAMMONDS, SUSAN E CELNIKER, MARK D BIGGIN, AND JOHN A STAMATOYANNOPOULOS. **Dynamic reprogramming of chromatin accessibility during Drosophila embryo development.** *Genome biology*, **12**(5) :R43, May 2011. 84, 85
- [119] XIAO-YONG LI, SEAN THOMAS, PETER J SABO, MICHAEL B EISEN, JOHN A STAMATOYANNOPOULOS, AND MARK D BIGGIN. **The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding.** *Genome biology*, **12**(4) :R34, April 2011. 84
- [120] YONG ZHANG, TAO LIU, CLIFFORD A MEYER, JÉRÔME ECKHOUTE, DAVID S JOHNSON, BRADLEY E BERNSTEIN, CHAD NUSBAUM, RICHARD M MYERS, MYLES BROWN, WEI LI, AND X SHIRLEY LIU. **Model-based analysis of ChIP-Seq (MACS).** *Genome biology*, **9**(9) :R137, January 2008. 85
- [121] AARON R. QUINLAN AND IRA M. HALL. **BEDTools : a flexible suite of utilities for comparing genomic features.** *Bioinformatics (Oxford, England)*, **26**(6) :841–2, March 2010. 85
- [122] PAULINE A FUJITA, BROOKE RHEAD, ANN S ZWEIG, ANGIE S HINRICHS, DONNA KAROLCHIK, MELISSA S CLINE, MARY GOLDMAN, GALT P BARBER, HIRAM CLAWSON, ANTONIO COELHO, MARK DIEKHANS, TIMOTHY R DRESZER, BELINDA M GIARDINE, RACHEL A HARTE, JENNIFER HILLMAN-JACKSON, FAN HSU, VANESSA KIRKUP, ROBERT M KUHN, KATRINA LEARNED, CHIN H LI, LAURENCE R MEYER, ANDY POHL, BRIAN J RANEY, KATE R ROSENBLUM, KAYLA E SMITH, DAVID HAUSSLER, AND W JAMES KENT. **The UCSC Genome Browser database : update 2011.** *Nucleic acids research*, **39**(Database issue) :D876–82, January 2011. 85, 121
- [123] STEVEN M GALLO, DAVE T GERRARD, DAVID MINER, MICHAEL SIMICH, BENJAMIN DES SOYE, CASEY M BERGMAN, AND MARC S HALFON. **REDfly v3.0 : toward a comprehensive database of transcriptional regulatory elements in Drosophila.** *Nucleic acids research*, **39**(Database issue) :D118–23, January 2011. 91
- [124] TAKAHIRO NAKAYAMA, KENICHI NISHIOKA, YI-XIN DONG, TSUKASA SHIMOJIMA, AND SUSUMU HIROSE. **Drosophila GAGA factor directs histone H3.3 replacement that prevents the heterochromatin spreading.** *Genes & development*, **21**(5) :552–61, March 2007. 109
- [125] EMMANUELLE SZENKER, DOMINIQUE RAY-GALLET, AND GENEVIÈVE ALMOUZNI. **The double face of the histone variant H3.3.** *Cell research*, **21**(3) :421–34, March 2011. 110
- [126] T TSUKIYAMA AND C WU. **Purification and properties of an ATP-dependent nucleosome remodeling factor.** *Cell*, **83**(6) :1011–20, December 1995. 110, 111
- [127] VIVEK SAROJKUMAR CHOPRA, ARUMUGAM SRINIVASAN, RAM PARIKSHAN KUMAR, KRISHNAVENI MISHRA, DENIS BASQUIN, MYLÈNE DOCQUIER, CAROLE SEUM, DANIEL PAULI, AND RAKESH KUMAR MISHRA. **Transcriptional activation by GAGA factor is through its direct interaction with dmTAF3.** *Developmental biology*, **317**(2) :660–70, May 2008. 111
- [128] A BEVILACQUA, M T FIORENZA, AND F MANGIA. **A developmentally regulated GAGA box-binding factor and Sp1 are required for transcription of the hsp70.1 gene at the onset of mouse zygotic genome activation.** *Development (Cambridge, England)*, **127**(7) :1541–51, April 2000. 111

REFERENCES

- [129] CHANHYO LEE, XIAOYONG LI, AARON HECHMER, MICHAEL EISEN, MARK D BIGGIN, BRYAN J VENTERS, CIZHONG JIANG, JIAN LI, B FRANKLIN PUGH, AND DAVID S GILMOUR. **NELF and GAGA factor are linked to promoter-proximal pausing at many genes in *Drosophila*.** *Molecular and cellular biology*, **28**(10):3290–300, May 2008. 111
- [130] DAVID A HENDRIX, JOUNG-WOO HONG, JULIA ZEITLINGER, DANIEL S ROKHSAR, AND MICHAEL S LEVINE. **Promoter elements associated with RNA Pol II stalling in the *Drosophila* embryo.** *Proceedings of the National Academy of Sciences of the United States of America*, **105**(22):7762–7, June 2008. 111

6

Annexes

6.1 Résultats supplémentaires

6.2 Articles

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

Ce protocole explique la façon d'utiliser *peak-motifs* via son interface web (<http://rsat.ulb.ac.be/rsat/>). *peak-motifs* est un programme dédié à la découverte de motifs sur-représentés dans des jeux de séquences complets provenant d'expériences de ChIP-seq et ChIP-chip. Ce programme intègre de nombreux outils de la suite RSAT (87). En effet, quatre algorithmes de découverte de motifs sont disponibles afin d'identifier des motifs significativement sur-représentés dans les jeux de pics. Ces motifs sont ensuite comparés à des motifs connus présents dans des bases de données. Les séquences sont ensuite scannées afin de prédire des sites de liaison, d'analyser leur enrichissement et leur position par rapport au centre des pics. Les pics analysés et les sites prédits à partir des motifs découverts peuvent être exportés vers le navigateur de génome de UCSC (122) afin d'être visualiser dans leur contexte génomique.

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

6. ANNEXES

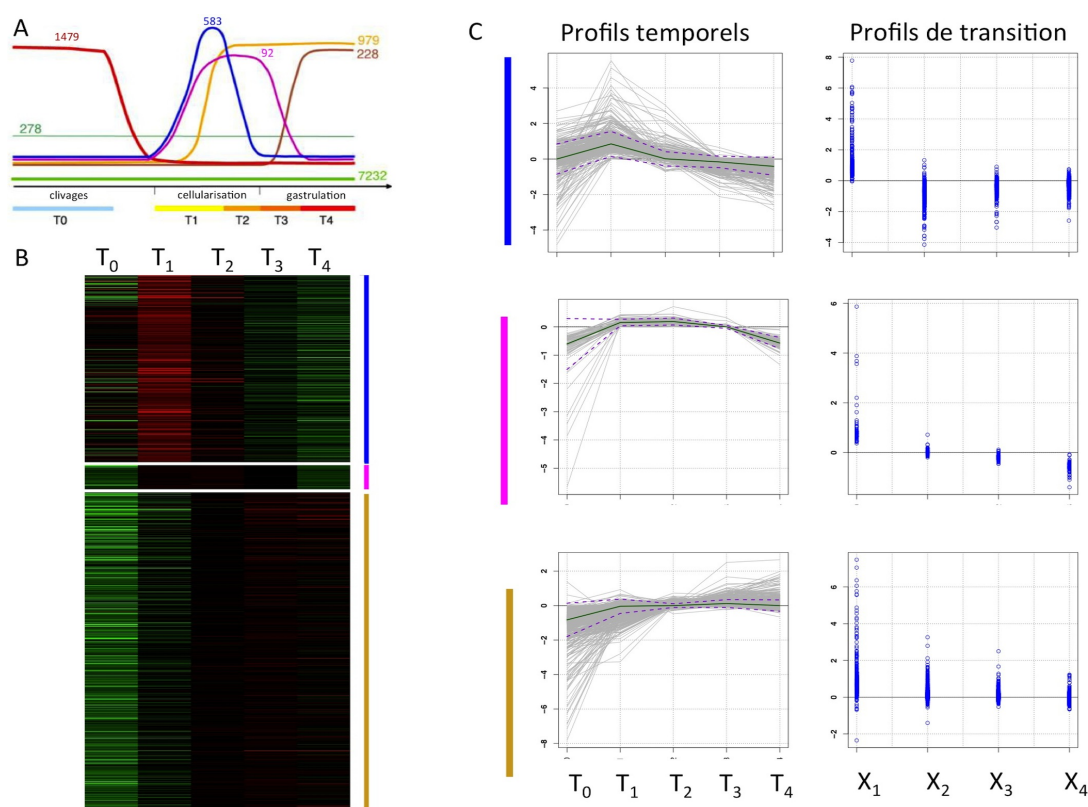


FIGURE 6.1: Exploration des clusters groupant les gènes potentiellement activés pendant l'AGZ selon Pilot et collaborateurs obtenus par la méthode de détection MAS5.0. - A. Représentation des différents groupes de co-expressions (D'après Pilot et al. (58)). B. Heatmap des profils temporels entre T0 et T4, vert, noir et rouge : intensité inférieure, égale ou supérieure à par l'intensité médiane du gène au travers de tous les points temporels, respectivement des trois groupes de gènes activés pendant la cellularisation (profils bleu, rose et jaune du panneau A). C. Profils temporels correspondant aux clusters d'intérêts où l'ordonnée indique la valeur d'intensité (log2) pour chaque gène (chaque courbe grise correspond à un gène) standardisée par l'intensité médiane du gène au travers de tous les points temporels. La ligne verte représente le profil médian et les lignes pointillées représentent la 1 déviation standart par rapport à la médiane. À droite : Profils de transition (X1 à X4) entre points temporels consécutifs. L'ordonnée indique la valeur du log ratio et chaque point représente un gène.

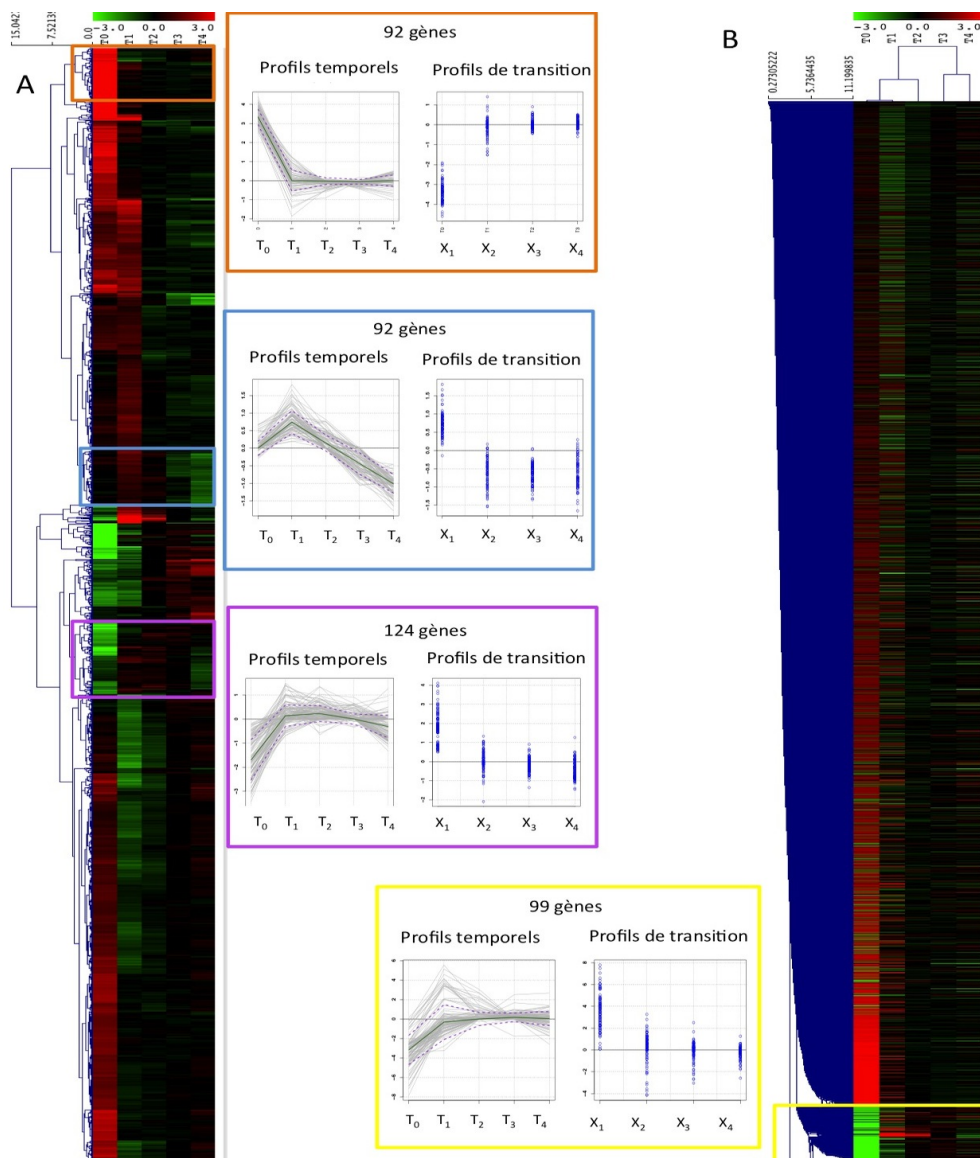


FIGURE 6.2: Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression. - Pour chaque panel A et B, les cadres de couleur indiquent les mêmes clusters représentés sous différentes formes : à droite, heatmap des profils temporels entre T₀ et T₄, vert, noir et rouge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

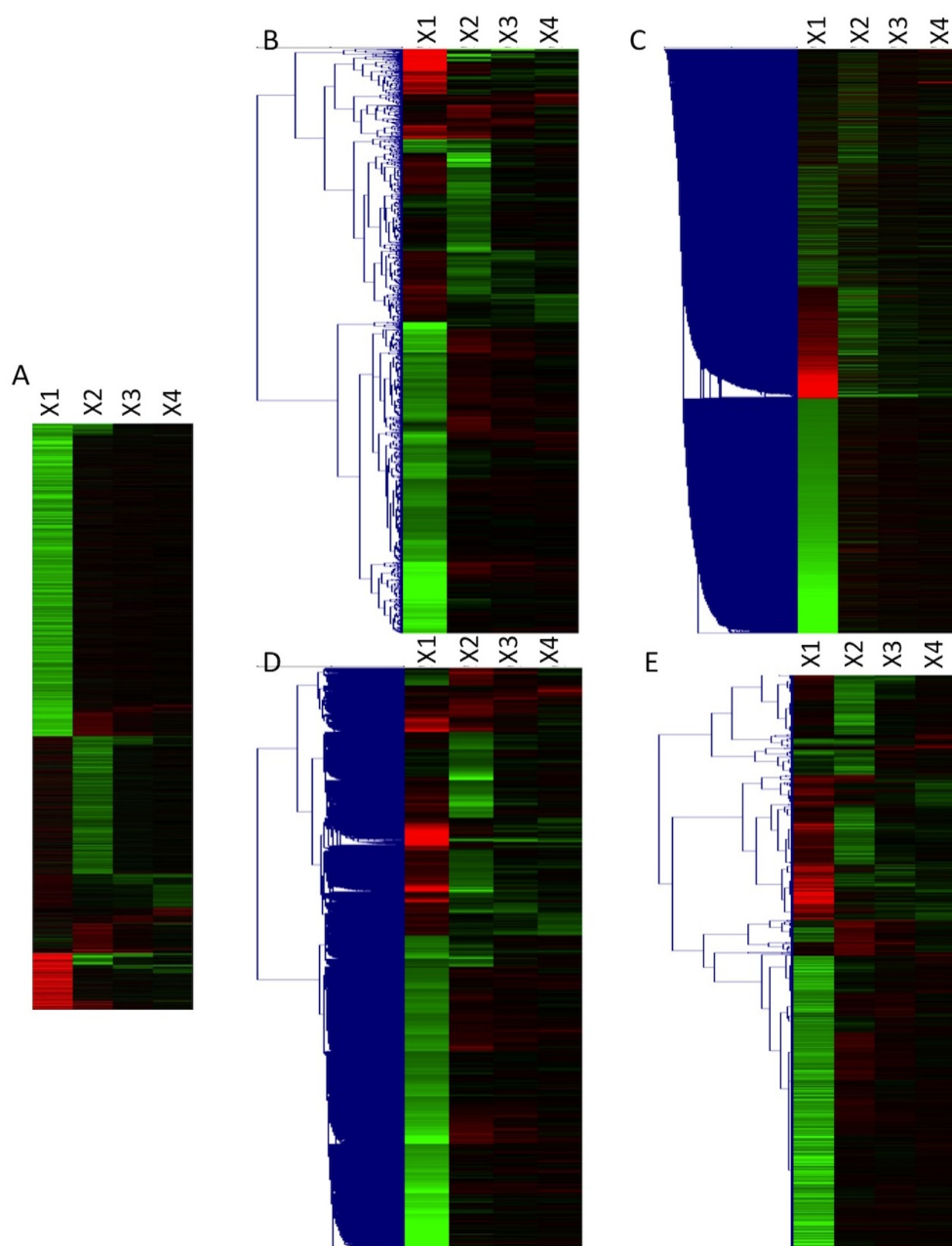


FIGURE 6.3: Impact de la méthode et des paramètres de clustering sur la cohérence des groupes de coexpression à partir des profils de transitions. - Heatmap représentant les clusters obtenus à partir des profils de transition (X_{i1} à X_{i4}) entre points temporels consécutifs avec différentes méthodes.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

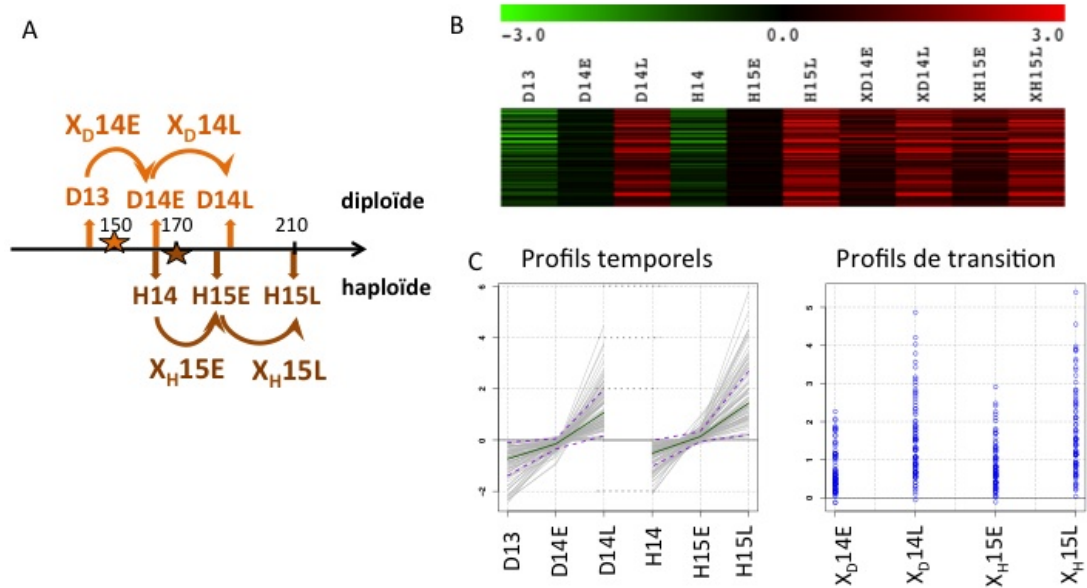


FIGURE 6.4: Exploration des clusters groupant les gènes potentiellement activés dépendamment du ratio NC pendant l'AGZ selon Lu et collaborateurs. - A. Echelle temporelle des données de puces à partir d'embryons sauvages diploïdes (haut), et mutants haploïdes (bas). L'axe horizontal représente le temps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

6. ANNEXES

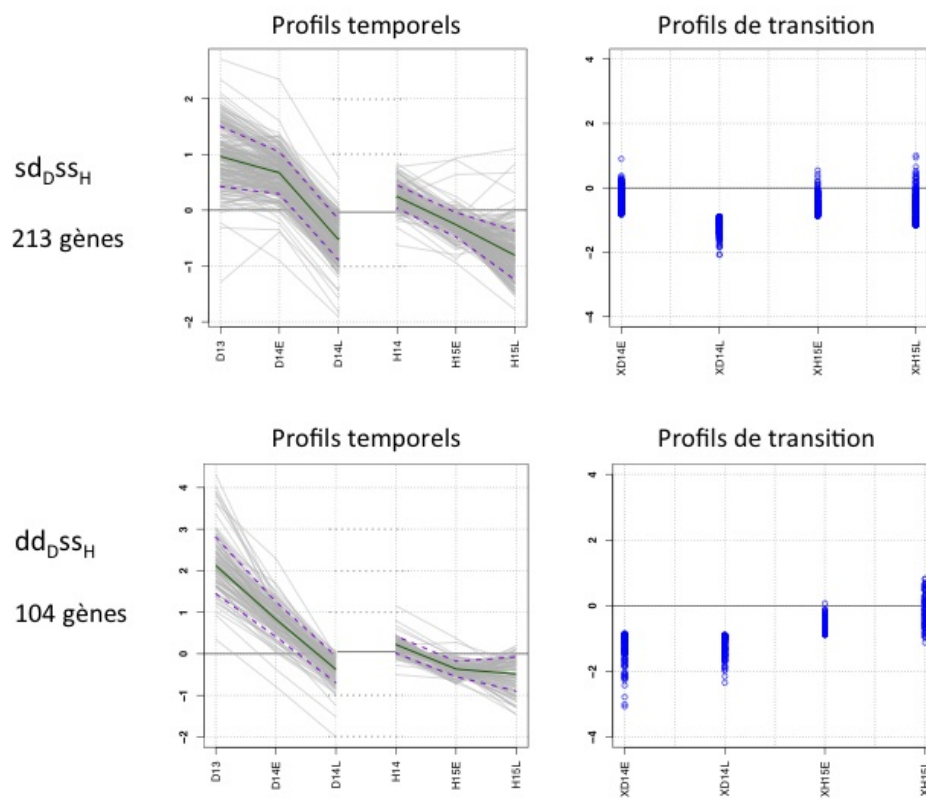


FIGURE 6.5: Profils temporels et de transition des clusters les plus importants ne correspondant à aucun modèle de régulation connu obtenus à partir des données de Lu et collaborateurs -

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

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

6. ANNEXES

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

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

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

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

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

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

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

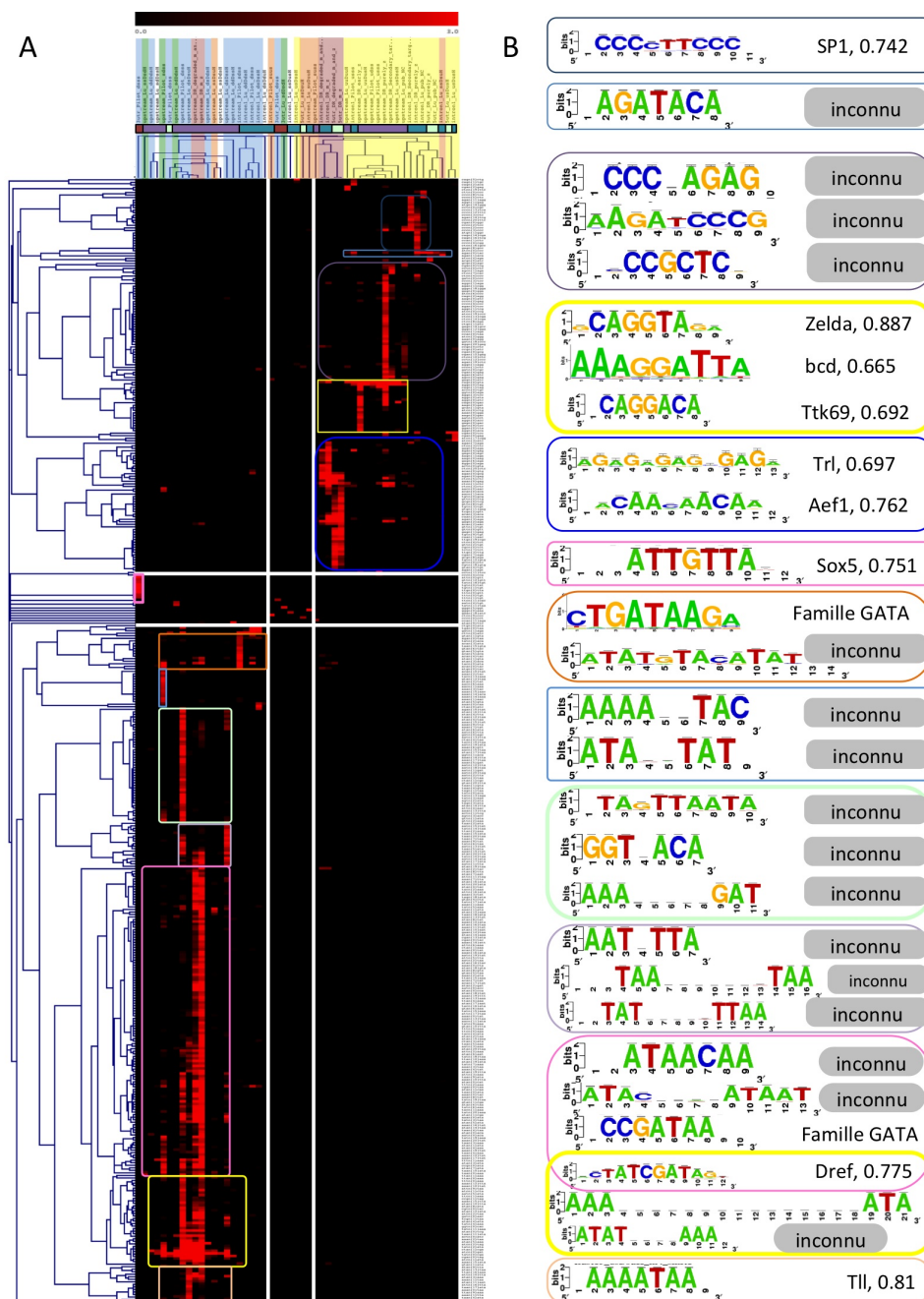


FIGURE 6.6: Clustering des dyades et des groupes de gènes en fonction de la significativité de la surreprésentation des dyades. -

6. ANNEXES

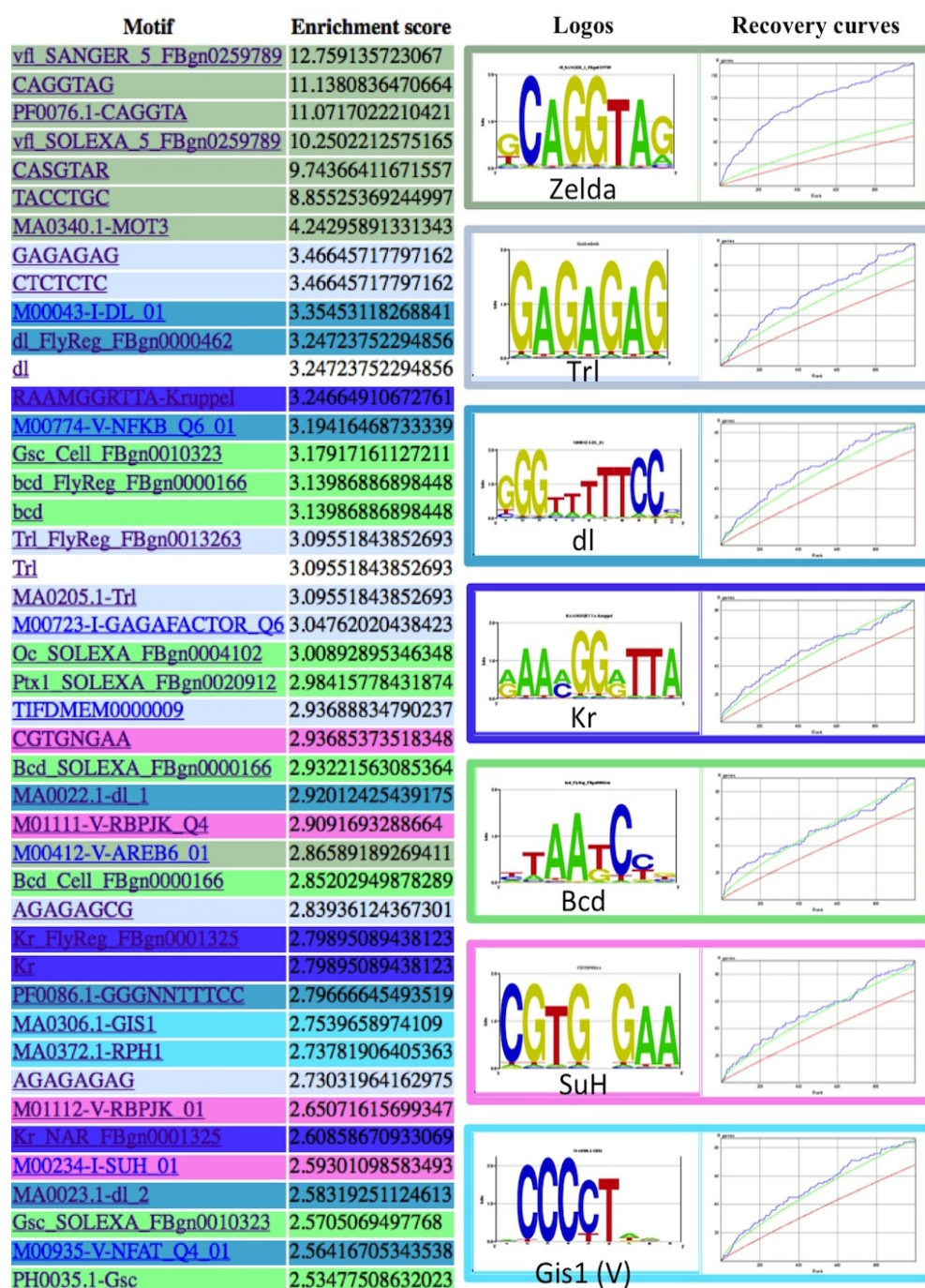


FIGURE 6.7: Résumé des résultats CisTargetX pour le cluster AGZ. - Les lignes marquées de la même couleur de fond indiquent des motifs similaires. Le logo et la courbe de recouvrement représentent le motif ayant le plus haut score d'enrichissement dans chaque groupe et le nom indique le facteur de transcription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

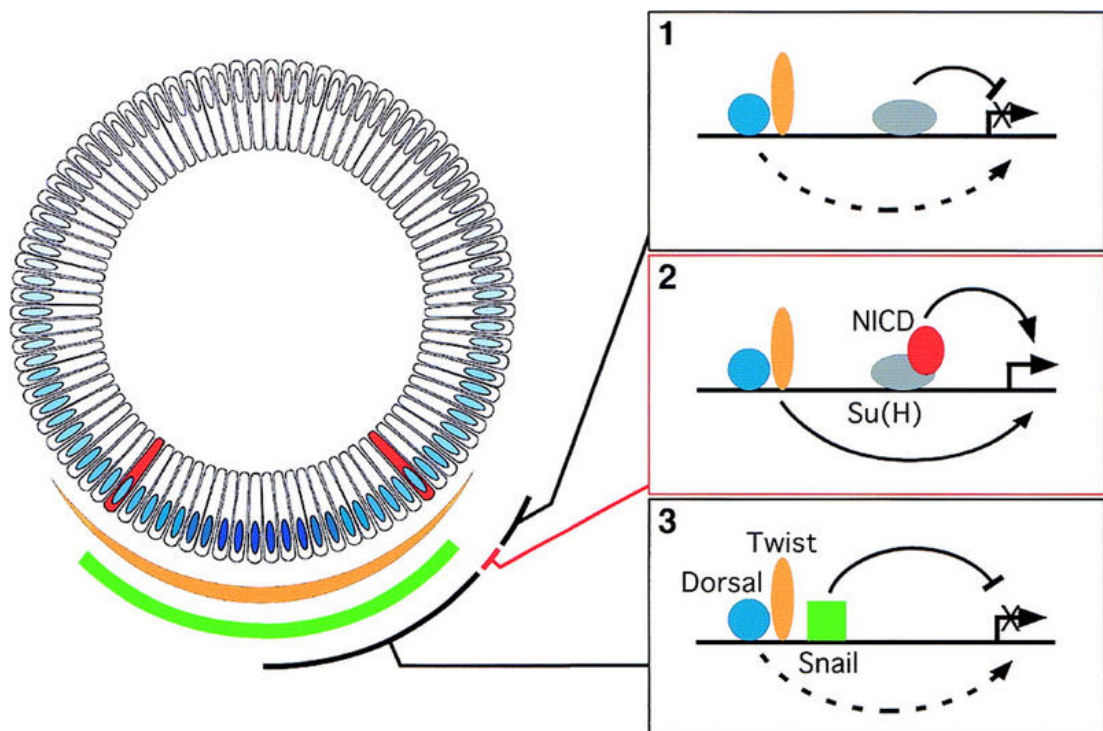


FIGURE 6.8: Action de Su(H) dans la définition du mésoderm. - En haut : partie dorsale ; en bas : partie ventrale. Gradient orange : Twist, Gradient bleu : Dorsal, vert : Snail. Les cellules rouges représentent le mésoderm. (D'après Morel et Schweisguth, 2000 (103))

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

**Morgane Thomas-Chollier¹, Elodie Darbo², Carl Herrmann², Matthieu Defrance³, Denis Thieffry^{2,4}
and Jacques van Helden^{5,2}**

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

ABSTRACT

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

ABBREVIATIONS

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

INTRODUCTION

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

density of reads. The typical result is a list of several thousands peak regions of varying sizes (from a few tens base pairs to several kilobases). There is a crucial need for efficient and user-accessible tools to extract relevant information from high-throughput sequencing data^{3,4}. While various programs have been developed to perform read mapping and peak calling⁵, the subsequent steps have not yet reached proper maturation: identifying relevant transcription factor binding motifs and the precise location of their binding sites remains a bottleneck. Most existing tools present limitations on sequence size, and typically restrict motif discovery to a few hundreds peaks^{6,7}, or to the central-most part of the peaks⁸.

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

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

1. **Sequence purging.** Sequences are automatically purged to discard redundant fragments (peak overlaps, duplications), which would bias the detection of over-represented motifs.
2. **Peak length distribution.** The distribution of sequence lengths provides a useful way to detect outlier peaks (i.e. exceptionally long peaks that may "dilute" the motif signal) or irregular lengths distributions resulting from problems during the peak calling procedure. Such indications may lead to redo the pre-processing in order to refine the peaks (e.g. with PeakSplitter⁹) before using *peak-motifs*.
3. **Sequence composition.** Nucleotide and dinucleotide compositions are computed and displayed in the form of heat maps and positional profiles (**Box 1**).
4. **Motif discovery.** The workflow combines four word-counting algorithms relying on two complementary criteria (over-representation and positional bias) to detect exceptional words (oligonucleotides) and spaced pairs of words (dyads) (**Box 2**). Significant words are used as seeds to build probabilistic description of motifs (position-specific scoring matrices), indicating residue variability at each position of the motif.
5. **Comparisons with motif databases.** Discovered motifs are compared to one or several public databases of annotated motifs, to predict associated transcription factors. Comparison results are displayed as multiple motif alignments to highlight matches with several annotated motifs (e.g. factors belonging to the same family, composite motifs bound by protein complexes).
6. **Reference motif(s).** In some cases, ChIP-seq experiments are done with transcription factors for which some binding motifs have already been characterized and annotated in specialized databases¹⁰⁻¹². Even in such cases, it is always interesting to discover motifs in peak sets, for

several reasons.

- a. Motifs discovered from large peak collections are generally more robust than those annotated from a handful of binding sites, leading to a significant refinement of their predictive power¹³.
- b. The discovery can result in multiple motifs corresponding to transcription factors interacting with the targeted factors¹⁴.

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

7. **Motif similarity networks.** All discovered motifs and their matches with reference or database motifs are displayed in the form of networks, where nodes represent motifs and edges their similarity. This visualization enables to grasp the groups of similar motifs returned by the different algorithms.
8. **Binding site predictions.** Sequences are scanned with the discovered motifs to predict binding sites. Peaks and predicted sites can be uploaded in the UCSC genome browser¹⁵, in order to visualize them in their genomic context.

Main advantages of *peak-motifs*

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

Main limitations

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

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

2. The output presents motifs in a somewhat redundant form, since the same motif can be discovered by multiple algorithms. We however chose to maintain this partly redundant presentation because detecting a motif by several independent programs indicates the robustness of the result, and can show that this motif is both over-represented (*oligo-analysis*, *dyad-analysis*) and concentrated in the center of the peaks (*position-analysis*, *local-word-analysis*).
3. The motif discovery algorithms consider all peak regions as equivalent, and cannot take into consideration the actual peak shape. Such information provided as a coverage file can be taken into account by other programs such as ChIPMunk¹⁶.

Study case

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

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

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

MATERIALS

Equipment

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

SETUP).

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

Equipment setup

Peak collections

Peak sequences can be produced by custom experiments, or obtained from publicly available datasets. Such datasets can be obtained for example from UCSC¹⁵ (<http://genome.ucsc.edu/>), GEO¹⁸ (<http://www.ncbi.nlm.nih.gov/geo/>) or Galaxy¹⁹ (<http://main.g2.bx.psu.edu/>).

Download the peak sequences for testing this protocol

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

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

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

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

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

Reference matrices and custom motif databases (optional)

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

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

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

convert-matrix.

PROCEDURE

Access the *peak-motif* web form

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

Sequences

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

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

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

- a) Make sure that the sequence file contains **peak sequences** and **not the raw reads**. A peak file should have a size in the range of several Mbytes whereas a read file with millions of reads has a size of hundreds Mbytes to few Gbytes. It is crucial to run *peak-motifs* on peak sequences, since the reads generally correspond to short fragments (typically 30bp) *on the left and on the right sides* of the actual binding sites. In addition, files containing several million reads are too large for online treatment. Files containing read sequences should first be treated with a read mapping program (e.g. Bowtie²⁰) that will align the reads on the reference genome. The resulting mapped reads should be processed with a peak-calling program (e.g. MACS²¹) to obtain the peak coordinates, and the corresponding sequences can be retrieved from specialized Web resources (UCSC, Galaxy).
- b) The program expects peak sequences (in FASTA format) and not peak coordinates

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

- c) Depending on the peak-calling program used, peaks may span several hundreds to thousands base pairs. Long peaks often result from the merging of a series of neighboring peaks. In this case, *peak-motifs* will perform better if these peaks are refined into individual peaks, for example with PeakSplitter⁹. This will increase the performance of *position-analysis* and *local-word-analysis*, as both algorithms search for motifs with positional biases, which are diluted when the peak regions are too broad.

TROUBLESHOOTING (see Table 1)

4. *Peak-motifs* offers the possibility to easily reduce the analyzed dataset, by focusing on a given number of top sequences, and/or by trimming the sequences to a desired length around peak centers. Click on the triangle on the right of 'Reduce peak sequences' to expand the hidden panel (Figure 5, top panel). For the study case, the option `Number of top sequences to retain` is left blank, as we will use the full-size set of sequences.
5. The option `Cut peak sequences` restricts the analysis to the most central region of each peak (e.g. peak center +/- 200bp), supposed to be dense in binding sites. This assumption is nevertheless highly dependent on the peak-calling program, and on whether the peak centers actually correspond to their summits (where the binding site is supposed to be found). We thus advise to generally leave this option blank, or run both the restricted and complete analyses, and compare the results. However in the Krüppel dataset, peaks have variable sizes up to 13,205bp, which probably reflects a problem with the peak-calling procedure rather than the natural extension of Krüppel binding regions. For this particular study case, restrict the analysis to 1000bp around the center of each peak.

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

Motif discovery parameters

6. Click on the triangle on the right of 'Change motif discovery parameters' to expand the motif-discovery option panel (Figure 5, bottom panel). For the study case, keep the options *oligo-analysis* and *position-analysis* checked (the other algorithms may be checked for a full analysis, but this takes more time). Check the values 6 and 7 for '`Oligomer length`', in order to detect significant hexanucleotides and heptanucleotides.
7. The background model must be specified when analyzing a single set of peaks. In the case of differential analysis, the second set of peaks (control set) will serve as background to estimate the random expectation of each oligonucleotide. In single-dataset mode, the background model is built from the test sequences based on frequencies of smaller words. For the study case, select the most stringent background model (`oligo length -2`).

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

TROUBLESHOOTING (see Table 1)

Comparisons of discovered motifs with motif databases

8. Click on the triangle on the right of 'Compare discovered motifs with databases' to reveal the 'Compare motifs' panel. This section displays a list of public motif databases, such as JASPAR¹⁰, that are directly supported by *peak-motifs*. Each discovered motif will be compared to the selected collection(s) of motifs, in order to identify which transcription factors may correspond to these binding motifs, or to pinpoint the currently unknown motifs. Motif databases should be chosen according to the studied organism. For the study case, deselect the default database ('JASPAR core Vertebrates') and select all the databases related to drosophila ('JASPAR core Insects', 'Drosophila DMMPMM' and 'Drosophila IDMPMM'), as illustrated in Figure 6.

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

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

Search for binding sites and export as UCSC custom track

11. Click on the triangle on the right of 'Locate motifs and export as UCSC custom track' to expand the panel with the options for searching putative binding sites in the peak sequences (Figure 6, bottom). Check the box 'Search putative binding sites in the peak sequences'.
12. Optionally, the sites can be exported as a custom track (BED file) that can be uploaded in the UCSC genome browser¹⁵ or Ensembl²², to visualize these putative binding sites in their annotated genomic environment. By default, this very helpful way to interpret the results is disabled, as it requires information in addition to peak sequences (genome assembly version and coordinates of

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

Submit the form

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

Viewing the results

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

TIME TAKEN

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

TROUBLESHOOTING

See Table 1.

ANTICIPATED RESULTS

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

Sequence length distribution and composition

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

Reference motifs

For the study case, we use as reference two Krüppel binding motifs extracted from JASPAR¹⁰ and FlyReg²³, respectively. Logos are displayed in both direct and reverse complementary orientations (Figure 8). The colored consensus sequences are shown above the logos, and can be searched for in the html output using the browser search function.

Discovered motifs (by algorithm)

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

Discovered motifs (with motif comparison)

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

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

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

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

Motif comparisons with multiple logo alignments

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

Predicted sites and enrichment profiles

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

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

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

the annotation tracks (Figure 12).

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

REFERENCES

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

BOXES

Box 1. Sequence composition and background models.

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

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

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

Box 2. Motif discovery algorithms

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

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

Global over-representation of words (oligo-analysis)

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

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

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

Positional biases (position-analysis)

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

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

Local over-representation (local-words)

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

Statistical significance of exceptional words

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

probability for one particular word to show a given level of over-representation or positional bias by chance, according to the background model. Since the significance test is applied on several thousands of words, a multi-testing correction is applied by converting the P-value to an E-value ($E\text{-value} = P\text{-value} * \text{number tested words}$) representing the expected number of false positives. This E-value is in turn converted to a significance index $sig = -\log_{10}(E\text{-value})$, providing an intuitive feeling of the reliability of the result (the higher the better).

Building matrices from lists of words

Each of the word-based motif discovery algorithms described above returns a set of exceptional words (oligonucleotides or dyads) sorted by significance. This list generally includes groups of mutually overlapping words, which reveal shifted fragments of the same motif. Those words can be used as seeds to collect putative sites, which are then aligned to build a position-specific scoring matrix. The final result of the motif discovery is a set of such PSSM, which can be used to scan sequences for predicting binding sites. See our previous protocols for the principle of matrix building from words²⁸ and sequence scanning with matrices²⁹.

TABLES

Table 1. Troubleshooting

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

LEGENDS TO FIGURES

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

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

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

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

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

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

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

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

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

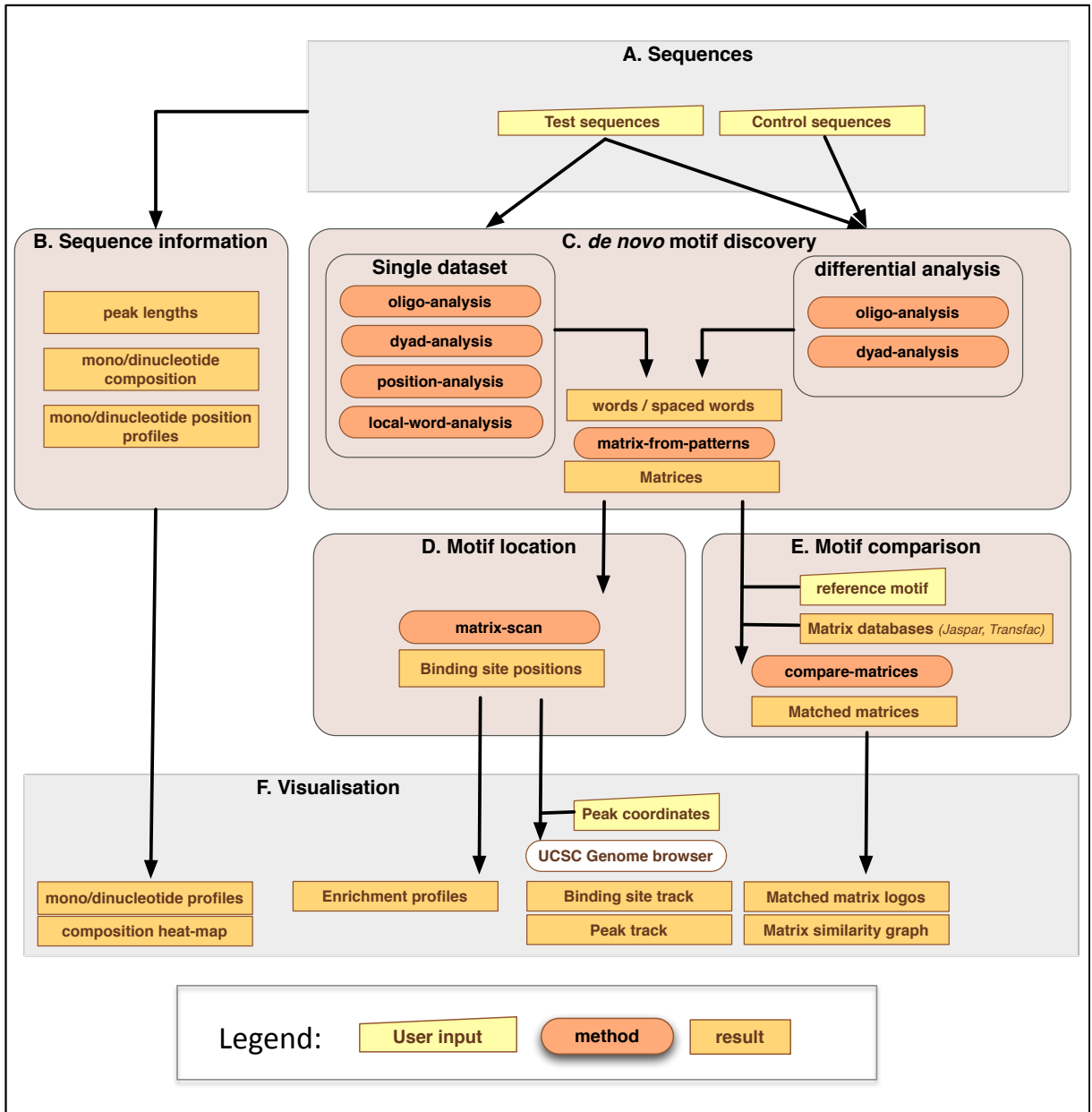
logos are displayed. The last column contains links toward intermediate results (over-represented words, assemblies of overlapping words and significance matrices) and matrices in TRANSFAC and tab formats.

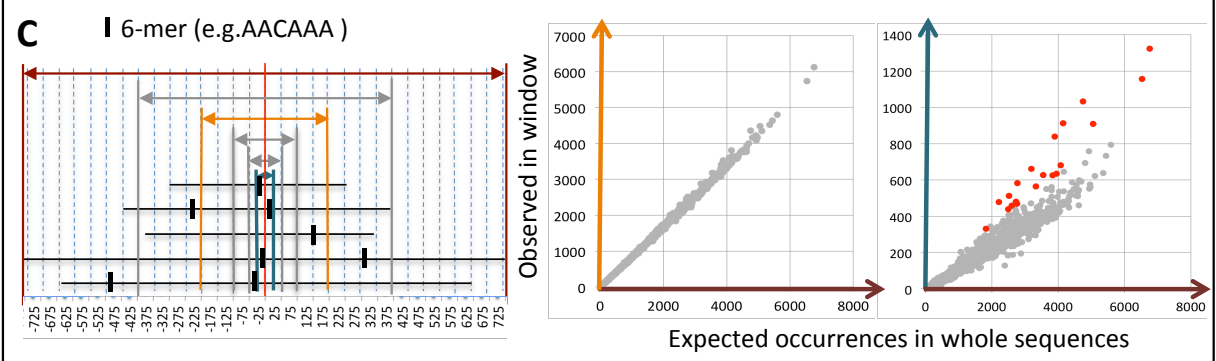
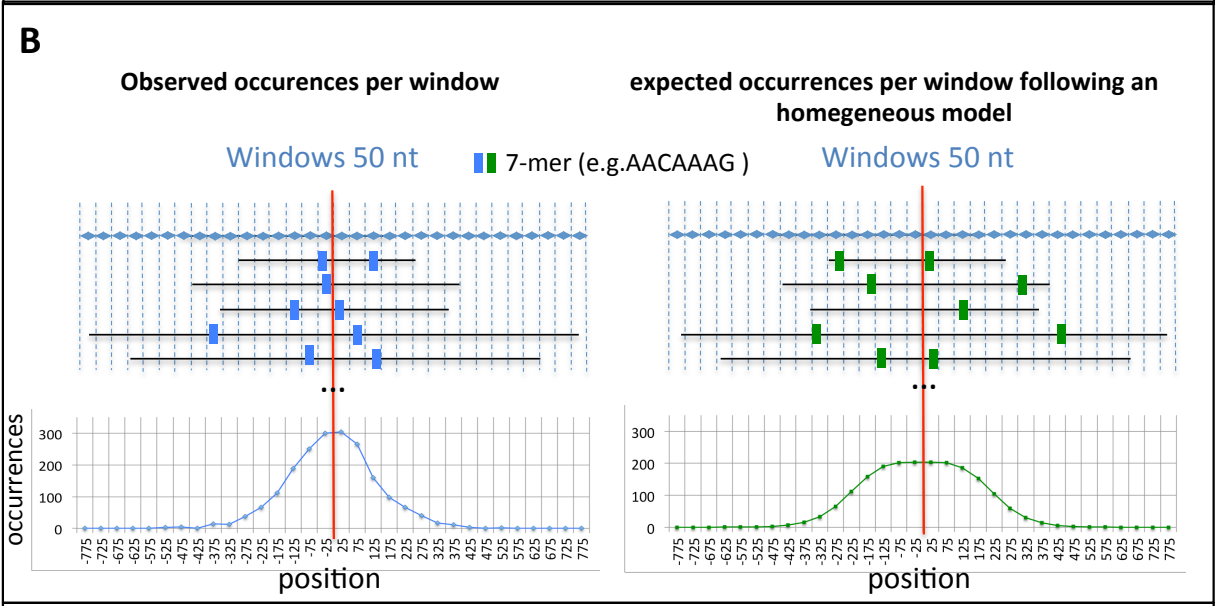
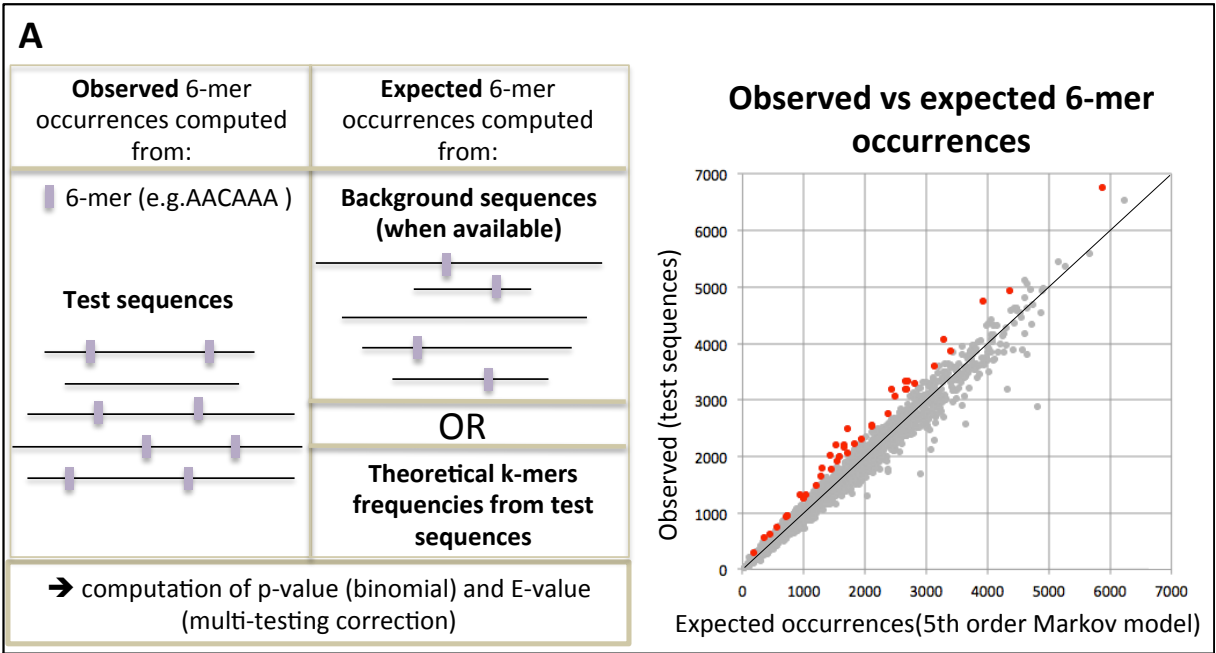
Figure 10. **Discovered motifs with motif comparisons.**

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

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

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



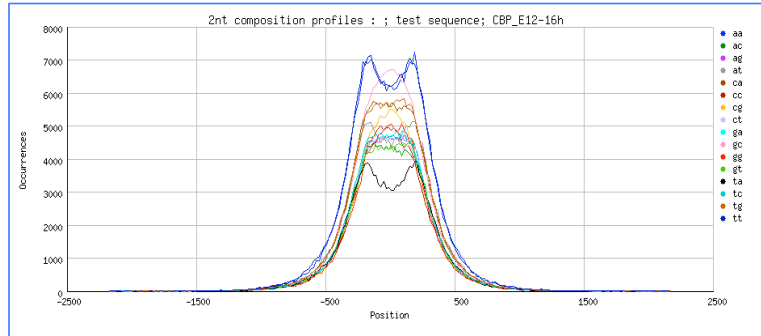


A. Drosophila transcriptional co-activator: CBP

Transition frequencies

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

Position profile

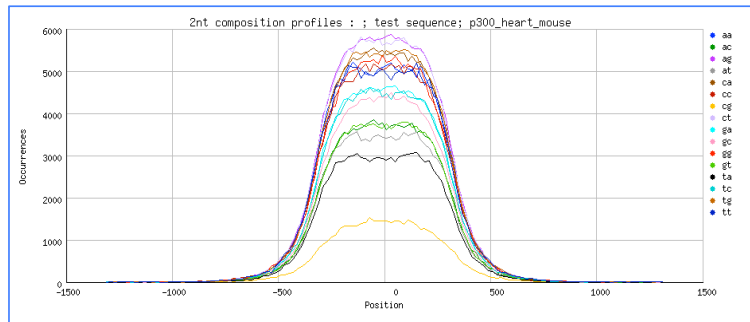


B. Mouse ortholog transcriptional co-activator: p300

Transition frequencies

pr\sufr	a	c	g	t	P_prefix
a	0.28516	0.20576	0.31378	0.19530	0.255
c	0.30680	0.28831	0.08174	0.32316	0.245
g	0.25648	0.24093	0.29096	0.21157	0.247
t	0.17107	0.24799	0.29850	0.28244	0.253

Position profile



RSA-tools - peak-motifs

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

Conception^c, implementationⁱ and testing^t: Jacques van Helden^{cd}, Morgane Thomas-Chollier^{cd}, Matthieu Defrance^{cd}, Olivier Sandiⁱ, Denis Thieffry^{ct} and Carl Herrmann^{ct}

► Information on the methods used in peak-motifs

Peak Sequences	
Title <input type="text" value="Kr D.mel 1-3h Markov -2"/>	
Peak sequences Paste your sequence in fasta format in the box below	Optional: control dataset for differential analysis (test vs control)
<input type="text"/>	Control sequences Paste your sequence in fasta format in the box below
<input type="text"/>	<input type="text"/>
Or select a file to upload (.gz compressed files supported)	Or select a file to upload (.gz compressed files supported)
<input type="button" value="Choisissez un fichier"/> Kr_D.mel...p1.fasta	<input type="button" value="Choisissez un fichier"/> Aucun fichier choisi
<small>(I only have coordinates in a BED file, how to get sequences ?)</small>	

► Reduce input peak sequences

► Change motif discovery parameters

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

► Locate motifs and export as UCSC custom track

Output display email

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

▼ Reduce input peak sequences

Restrict the input dataset

Number of top sequences to retain

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

▼ Change motif discovery parameters

Discover motifs

Continuous words

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

Oligomer length for the three programs above 6 7

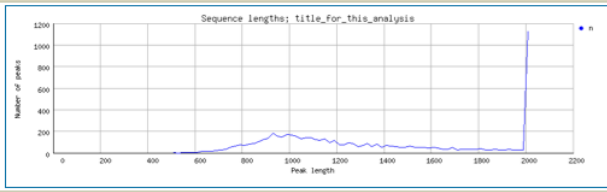

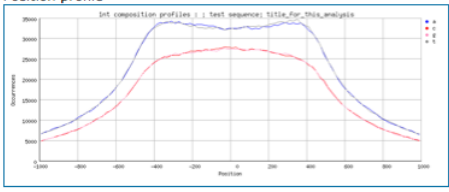
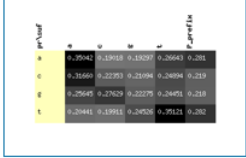
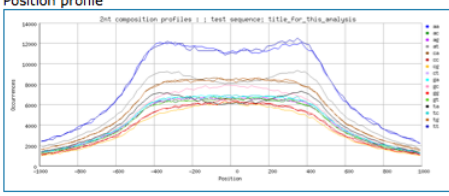
Note: motifs can be larger than word sizes (words are used as seed for building matrices)

Spaced words pairs

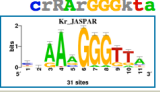


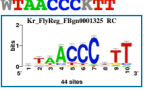
- Discover over-represented spaced word pairs [dyad-analysis]

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

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

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

▼ Reference motifs

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

▼ Compare discovered motifs with databases (e.g. against Jaspas) or custom reference motifs

Compare motifs

Compare discovered motifs with known motifs from databases

- JASPAR core Vertebrates
- JASPAR core Fungi
- JASPAR core Insects
- JASPAR core Nematodes
- JASPAR core Plants
- JASPAR core Urochordates
- JASPAR PBM (UNIPROBE) Mouse
- JASPAR PBM (UNIPROBE) Homeo Mouse
- JASPAR PBM (UNIPROBE) HLH Nematode
- Drosophila DMMPMM (<http://line.imb.ac.ru/DMMPMM/>)
- Drosophila IDMPMM (<http://line.imb.ac.ru/IDMPMM/>)
- RegulonDB prokaryotes

Add your own motif database:

Matrices should be in **Transfac format** (other formats can be converted with *convert-matrix*).

Add known reference motifs for this experiment:

Matrices should be in **Transfac format** (other formats can be converted with *convert-matrix*).

▼ Locate motifs and export as UCSC custom track

Locate motifs

- Search putative binding sites in the peak sequences [matrix-scan]

Visualize motifs in genome browser

- No
- Yes; sequences fetched from **Galaxy** (fasta headers should be in the form: >mm9_chr1_3473041_3473370_+)
- Yes; use the following BED file.

BED file with peak coordinates Aucun fichier choisi **Assembly version (UCSC)**


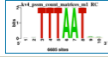










The 4th column of the BED file (feature name) correspond to the fasta headers of input sequences


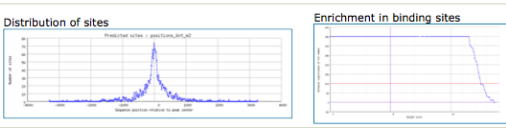

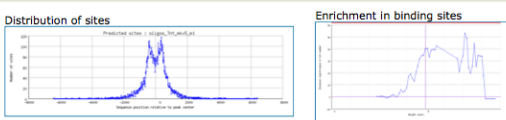
Output display email

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

[\[MANUAL\]](#) [\[TUTORIAL\]](#) [\[ASK A QUESTION\]](#)

▼ Discovered motifs (by algorithm)

motif discovery		
oligos_6nt_mkv4	asmb: attaaa (sig=169.67) RC: ttaa  	[discovered words: text] [assembly: text sig matrices] [matrices: tab format transfac format]
	asmb: cttatca (sig=112.46) RC: tgataag  	
	asmb: atgcaa (sig=108.46) RC: ttgcat  	
positions_6nt	asmb: aataatttattataaataat (sig=300) RC: atattttaataataattatt  	[discovered words: text] [assembly: text sig matrices] [matrices: tab format transfac format]
	asmb: gaaagggtta (sig=300) RC: taacccttc  	
	asmb: gagagag (sig=36.51) RC: ctctctc  	

Motif 5 positions_gnt_m2		[matrix: tab format transfac format]																																				
Reference motifs																																						
<table border="1"> <thead> <tr> <th>name</th> <th>id</th> <th>strand</th> <th>Nb overlap columns</th> <th>% aligned</th> <th>Pearson correlation</th> <th>Normalized cor</th> <th>aligned col. motif</th> <th>aligned col. match</th> </tr> </thead> <tbody> <tr> <td>Kr_JASPAR</td> <td>Kr_JASPAR</td> <td>D</td> <td>11</td> <td>0.7857</td> <td>0.945</td> <td>0.743</td> <td>I V r A A G G G T T r a r</td> <td>c r A A a G G C T T a</td> </tr> <tr> <td>Kr_FlyReg_FBgn0001325</td> <td>Kr_FlyReg_FBgn0001325</td> <td>D</td> <td>10</td> <td>0.7143</td> <td>0.959</td> <td>0.685</td> <td>. . . A A G G G T T r a .</td> <td>A A m G G G T t a w</td> </tr> </tbody> </table> <p>Total matches= 2</p>			name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match	Kr_JASPAR	Kr_JASPAR	D	11	0.7857	0.945	0.743	I V r A A G G G T T r a r	c r A A a G G C T T a	Kr_FlyReg_FBgn0001325	Kr_FlyReg_FBgn0001325	D	10	0.7143	0.959	0.685	. . . A A G G G T T r a .	A A m G G G T t a w									
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match																														
Kr_JASPAR	Kr_JASPAR	D	11	0.7857	0.945	0.743	I V r A A G G G T T r a r	c r A A a G G C T T a																														
Kr_FlyReg_FBgn0001325	Kr_FlyReg_FBgn0001325	D	10	0.7143	0.959	0.685	. . . A A G G G T T r a .	A A m G G G T t a w																														
jaspar_core_insects																																						
<table border="1"> <thead> <tr> <th>name</th> <th>id</th> <th>strand</th> <th>Nb overlap columns</th> <th>% aligned</th> <th>Pearson correlation</th> <th>Normalized cor</th> <th>aligned col. motif</th> <th>aligned col. match</th> </tr> </thead> <tbody> <tr> <td>Kr</td> <td>MA0452.1</td> <td>D</td> <td>11</td> <td>0.7857</td> <td>0.945</td> <td>0.743</td> <td>I V r A A G G G T T r a r</td> <td>c r A A a G G C T T a</td> </tr> </tbody> </table> <p>Total matches= 1</p>			name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match	Kr	MA0452.1	D	11	0.7857	0.945	0.743	I V r A A G G G T T r a r	c r A A a G G C T T a																		
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match																														
Kr	MA0452.1	D	11	0.7857	0.945	0.743	I V r A A G G G T T r a r	c r A A a G G C T T a																														
DMMPMM_drosophila																																						
<table border="1"> <thead> <tr> <th>name</th> <th>id</th> <th>strand</th> <th>Nb overlap columns</th> <th>% aligned</th> <th>Pearson correlation</th> <th>Normalized cor</th> <th>aligned col. motif</th> <th>aligned col. match</th> </tr> </thead> <tbody> <tr> <td>Kr</td> <td>Kr</td> <td>D</td> <td>10</td> <td>0.7143</td> <td>0.967</td> <td>0.691</td> <td>. . . r A A A G G G T T r . .</td> <td>a A A a G G G T t a</td> </tr> </tbody> </table> <p>Total matches= 1</p>			name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match	Kr	Kr	D	10	0.7143	0.967	0.691	. . . r A A A G G G T T r . .	a A A a G G G T t a																		
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match																														
Kr	Kr	D	10	0.7143	0.967	0.691	. . . r A A A G G G T T r . .	a A A a G G G T t a																														
IDMMPMM_drosophila																																						
<table border="1"> <thead> <tr> <th>name</th> <th>id</th> <th>strand</th> <th>Nb overlap columns</th> <th>% aligned</th> <th>Pearson correlation</th> <th>Normalized cor</th> <th>aligned col. motif</th> <th>aligned col. match</th> </tr> </thead> <tbody> <tr> <td>Kr</td> <td>Kr</td> <td>R</td> <td>11</td> <td>0.7857</td> <td>0.967</td> <td>0.760</td> <td>. V r A A A G G G T T r . .</td> <td>Y R A A A G G G T T A</td> </tr> </tbody> </table> <p>Total matches= 1</p>			name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match	Kr	Kr	R	11	0.7857	0.967	0.760	. V r A A A G G G T T r . .	Y R A A A G G G T T A																		
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match																														
Kr	Kr	R	11	0.7857	0.967	0.760	. V r A A A G G G T T r . .	Y R A A A G G G T T A																														
Predicted sites on input peaks		[view in genome browser : UCSC] [sites: text BED (UCSC track)] [distribution: text] [enrichment: text]																																				
Motif 7 oligos_7nt_mkv5_m1		[matrix: tab format transfac format]																																				
Reference motifs																																						
<table border="1"> <thead> <tr> <th>name</th> <th>id</th> <th>strand</th> <th>Nb overlap columns</th> <th>% aligned</th> <th>Pearson correlation</th> <th>Normalized cor</th> <th>aligned col. motif</th> <th>aligned col. match</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>W t A T T T k T A W W</td> <td></td> </tr> </tbody> </table> <p>Total matches= 0</p>			name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match								W t A T T T k T A W W																			
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match																														
							W t A T T T k T A W W																															
jaspar_core_insects																																						
<table border="1"> <thead> <tr> <th>name</th> <th>id</th> <th>strand</th> <th>Nb overlap columns</th> <th>% aligned</th> <th>Pearson correlation</th> <th>Normalized cor</th> <th>aligned col. motif</th> <th>aligned col. match</th> </tr> </thead> <tbody> <tr> <td>Abd-B</td> <td>MA0165.1</td> <td>D</td> <td>6</td> <td>0.5000</td> <td>0.875</td> <td>0.438</td> <td>. T K T A W W</td> <td>T T T A T k .</td> </tr> <tr> <td>cad</td> <td>MA0216.1</td> <td>D</td> <td>6</td> <td>0.5000</td> <td>0.846</td> <td>0.423</td> <td>. T K T A W W</td> <td>T T T A T T .</td> </tr> <tr> <td>hb</td> <td>MA0049.1</td> <td>R</td> <td>9</td> <td>0.7500</td> <td>0.782</td> <td>0.586</td> <td>. . . A T T T K T A W W</td> <td>T T T T T W Y K .</td> </tr> </tbody> </table> <p>Total matches= 3</p>			name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match	Abd-B	MA0165.1	D	6	0.5000	0.875	0.438 T K T A W W	T T T A T k .	cad	MA0216.1	D	6	0.5000	0.846	0.423 T K T A W W	T T T A T T .	hb	MA0049.1	R	9	0.7500	0.782	0.586	. . . A T T T K T A W W	T T T T T W Y K .
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match																														
Abd-B	MA0165.1	D	6	0.5000	0.875	0.438 T K T A W W	T T T A T k .																														
cad	MA0216.1	D	6	0.5000	0.846	0.423 T K T A W W	T T T A T T .																														
hb	MA0049.1	R	9	0.7500	0.782	0.586	. . . A T T T K T A W W	T T T T T W Y K .																														
DMMPMM_drosophila																																						
<table border="1"> <thead> <tr> <th>name</th> <th>id</th> <th>strand</th> <th>Nb overlap columns</th> <th>% aligned</th> <th>Pearson correlation</th> <th>Normalized cor</th> <th>aligned col. motif</th> <th>aligned col. match</th> </tr> </thead> <tbody> <tr> <td>zen</td> <td>zen</td> <td>R</td> <td>6</td> <td>0.5455</td> <td>0.817</td> <td>0.445</td> <td>. T K T A W W</td> <td>T T W A T T</td> </tr> <tr> <td>hb</td> <td>hb</td> <td>R</td> <td>9</td> <td>0.8182</td> <td>0.805</td> <td>0.659</td> <td>. . . A T T T K T A W W</td> <td>T T T T T R T K</td> </tr> <tr> <td>br-Z1</td> <td>br-Z1</td> <td>R</td> <td>8</td> <td>0.7273</td> <td>0.807</td> <td>0.587</td> <td>. . . T T T K T A W W</td> <td>T T W G T A T Y</td> </tr> </tbody> </table> <p>Total matches= 4 (1 more)</p>			name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match	zen	zen	R	6	0.5455	0.817	0.445 T K T A W W	T T W A T T	hb	hb	R	9	0.8182	0.805	0.659	. . . A T T T K T A W W	T T T T T R T K	br-Z1	br-Z1	R	8	0.7273	0.807	0.587	. . . T T T K T A W W	T T W G T A T Y
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match																														
zen	zen	R	6	0.5455	0.817	0.445 T K T A W W	T T W A T T																														
hb	hb	R	9	0.8182	0.805	0.659	. . . A T T T K T A W W	T T T T T R T K																														
br-Z1	br-Z1	R	8	0.7273	0.807	0.587	. . . T T T K T A W W	T T W G T A T Y																														
IDMMPMM_drosophila																																						
<table border="1"> <thead> <tr> <th>name</th> <th>id</th> <th>strand</th> <th>Nb overlap columns</th> <th>% aligned</th> <th>Pearson correlation</th> <th>Normalized cor</th> <th>aligned col. motif</th> <th>aligned col. match</th> </tr> </thead> <tbody> <tr> <td>cad</td> <td>cad</td> <td>D</td> <td>7</td> <td>0.5000</td> <td>0.805</td> <td>0.403</td> <td>. T K T A W W</td> <td>L T T T A T k</td> </tr> </tbody> </table> <p>Total matches= 1</p>			name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match	cad	cad	D	7	0.5000	0.805	0.403 T K T A W W	L T T T A T k																		
name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif	aligned col. match																														
cad	cad	D	7	0.5000	0.805	0.403 T K T A W W	L T T T A T k																														
Predicted sites on input peaks		[view in genome browser : UCSC] [sites: text BED (UCSC track)] [distribution: text] [enrichment: text]																																				

One-to-n alignments

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

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

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

UCSC Genome Browser on *D. melanogaster* Apr. 2006 (BDGP R5/dm3) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr3L:20,685,320-20,692,146 size 6,827 bp. [2011 ENCODE Usability Survey](#)



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

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

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

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

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

Received: 12 February 2008 / Accepted: 28 April 2008 / Published online: 17 June 2008
© Springer-Verlag 2008

Abstract Duplications are an important mechanism for the emergence of genetic novelties. Reports on duplicated genes are numerous, and mechanisms for polyploidization or local gene duplication are beginning to be understood. When a local duplication is studied, searches are usually done gene-by-gene, and the size of duplicated segments is not often investigated. Therefore, we do not know if the gene in question has duplicated alone or with other genes, implying that “en bloc” duplications are poorly studied. We propose a method for identification of “en bloc” duplication using mapping, phylogenetic and statistical analyses. We show that two segments present in the major histocompatibility complex (MHC) region of human chromosome 6 have resulted from an “en bloc” duplication that took place between divergence of amniotes and methaterian/eutherian

separation. These segments contain members of the same multigenic families, namely olfactory receptors genes, genes encoding proteins containing B30.2 domain, genes encoding proteins containing immunoglobulin V domain and MHC class I genes. We will discuss the fact that olfactory receptors and MHC genes have undergone positive selection, which could have helped in fixation of the surrounding genes.

Keywords Evolution · “En bloc” duplication · MHC · Phylogeny · Hitchhiking effect

Introduction

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

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

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

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

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

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

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

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

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

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

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

As we do not know the importance of “en bloc” duplication “a priori”, we decided to search for paralogous regions, which are not remnants of polyploidization events, and we considered essentially the composition of the segments (similar sets of paralogs in each region). We then verified if these paralogs had resulted from the same duplication event. For this, we used phylogenetic analyses to date the duplications and then used statistical analysis to test the significance of the clusters of paralogous genes. We propose this strategy using the major histocompatibility complex (MHC) region as a pilot region, as during previous investigations of the human MHC region, we observed two

segments (Horton et al. 2004) with similar genes composition and which contain members of the same multigenic families (olfactory receptors, MHC class I, proteins containing B30.2 domain and proteins containing IgV-MOG domain). We hypothesise that the gene distribution observed in these regions could be due to “en bloc” duplication rather than to a random process or a polyploidization event; indeed, this organisation could have occurred by chance. To test this hypothesis, we used a combination of mapping, phylogenetic and statistical analyses.

Materials, methods and strategy

Gene identification and description

Olfactory receptors family

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

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

Proteins with domain B30.2 family

The B30.2 protein family, observed only in vertebrate lineage, contains approximately 50 members. These proteins contain a globular domain, called B30.2, in the C-terminal extremity and different domains in N-terminal extremity. At least three groups are identified: RING finger, immunoglobulin-like and toxins (Henry et al. 1998). These proteins are more and more studied because they are involved in the recognition of some viruses (e.g. TRIM5 α /HIV; Yap et al. 2004).

Myelin oligodendrocyte glycoprotein

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

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

Butyrophilins

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

The class I MHC family

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

Other genes

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

Domains extraction

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

Phylogenetic analyses

We used the Figenix bioinformatics platform (Gouret et al. 2005) based on an expert system that allows automation of biological annotation pipelines. The phylogenetic pipeline CassiopePhylo+M was used here; it allows the detection of homologous sequences (orthologs and paralogs) for a given gene. It contains 52 steps (sequences retrieval, multiple alignments, tree building, filters, etc.). Consensus phylogenetic trees are obtained by fusion of trees built

from three methods: maximum likelihood, maximum parsimony and distance. If tree topologies are not congruent, fusion cannot be achieved or only partially. Orthologs/paralogs inference is made from comparison between this tree and a reference taxonomy tree, available at NCBI web site (<http://www.ncbi.nlm.nih.gov/>). Robustness of the obtained trees (for the three methods used) was tested with a bootstrap method.

ORs family

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

Proteins with domain B30.2 family

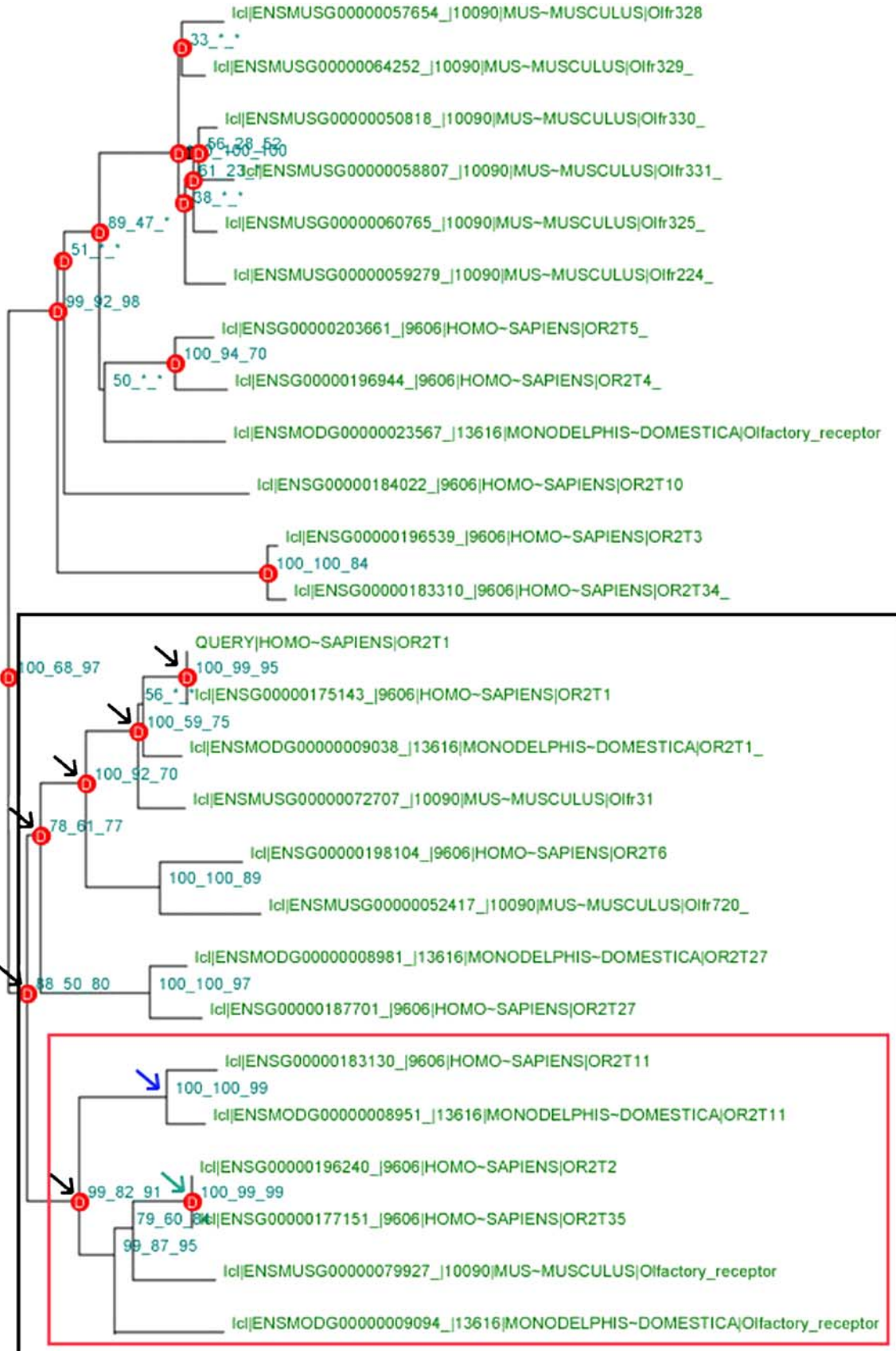
We searched all B30.2 proteins. We used BLASTp (<http://www.ncbi.nlm.nih.gov/BLAST/>; Altschul et al. 1990) to recover phylogenetically remote sequences so as to be sure we had retrieved all members of this family. We used the TRIM27 B30.2 domain as a query against the NCBI NR database. The B30.2 domain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

Myelin oligodendrocyte glycoprotein

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

Butyrophilins

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



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

Class I MHC family

Phylogenies were made with $\alpha 1/\alpha 2$ and IgC domains of HLA-A, MICA, MICB and HFE proteins. We submitted the sequences against the ENSEMBL database (Fig. 1).

Histones

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

Other genes

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

Gene mapping

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

Statistical significance

Implementation of the binomial test

We defined departure and arrival regions as the two regions supposed to have been engendered by duplication. The binomial test is used to determine the probability that the members of one region (departure) gave paralogs in the supposed paralogous region (arrival) by chance. It is extremely important to note here that we will use the duplication time information. It has to be taken into an account that the paralogs have to be issued from duplications that occurred in the same window of times. In some cases, duplicate issued from these duplications have duplicated later. The late duplications occurred in the same region “cis duplication” and remained there, or, alternatively, could have translocated elsewhere in the genome, in which case, we cannot know a priori the ancestral

localisation. Therefore, in such cases, a corrector coefficient should be applied. However, we have never witnessed the last case; see after). We applied the binomial test twice; we applied the test once with region 1 [human chromosome 6: 30.4Mb (TRIM39)..28.98Mb (TRIM27)] as the departure region and region 2 [human chromosome 6: 28.1Mb (OR1F12).. 26.07Mb (TRIM38)] (Fig. 2) as arrival region, and on one occasion with region 2 as departure region and region 1 as arrival region.

We tested two alternative hypotheses H0 and H1:

- H0: Gene distribution between “departure” region and “arrival” paralogous region results from a random process.
- H1: Gene distribution between “departure” region and “arrival” paralogous region does not result from a random process.

p : probability of each gene contained in “departure” region having a paralog (or a group of paralogs in the case of late duplication) in the “arrival” region.

q : probability that the paralogs localise elsewhere in the genome ($q=1-p$).

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

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

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

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

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

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

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

- $\alpha < 5\%$: results are significant
- $\alpha < 1\%$: results are very significant
- $\alpha < 0.1\%$: results are highly significant

Strategy

Boundaries of the regions and paralogs distribution analysis

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

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

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

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

(TRIM27); region 2 starts at 28.1 Mb (OR1F12) and ends at 26.07 Mb (TRIM38; Fig. 2 and Table 1).

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

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

Dating duplication events

Determination of the duplication time is deduced with regard to all phylogenies, and we determined in which period genes in regions of interest had duplicated. We started analyses using all vertebrate species available in ENSEMBL + NR. These first results were not analysable because of the excessively large size of the data; moreover, species too close phylogenetically (that diverge recently as apes and humans or mice and rats) did not provide information about the evolution of our regions. Therefore, we decided to use the more phylogenetic informative species as the species of reference: *Gallus gallus* (chicken, taxID: 9031), *Homo sapiens* (human, taxID: 9606), *Mus musculus* (mouse, taxID: 10090), *Monodelphis domestica* (opossum, taxID: 13616). We chose these species with regards to the first phylogenies, made from members of families present in the regions studied, using all available vertebrates’ genomes. Phylogenies from the genes contained in regions under investigation [region 1: human chromosome 6: 30.4Mb (TRIM39)..28.98Mb (TRIM27) and region 2: human chromosome 6: 28.1Mb (OR1F12)..26.07Mb (TRIM38)] (Fig. 2) indicate that the duplication events have engendered a duplicate in each region and occurred after bird/mammal separation. Indeed,

when orthologs were found in *G. gallus*’s genome, they systematically constituted an outgroup. Some genes appeared more recently (species specific genes); for example, TRIM48/TRIM49/TRIM51 have emerged in the human genome after rodents/primates separation. Others have an ortholog for at least one human or mouse duplicated gene in *Monodelphis domestica* genome. This indicates that duplications have probably occurred before placental mammal/marsupial separation.

Counting duplication events

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

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

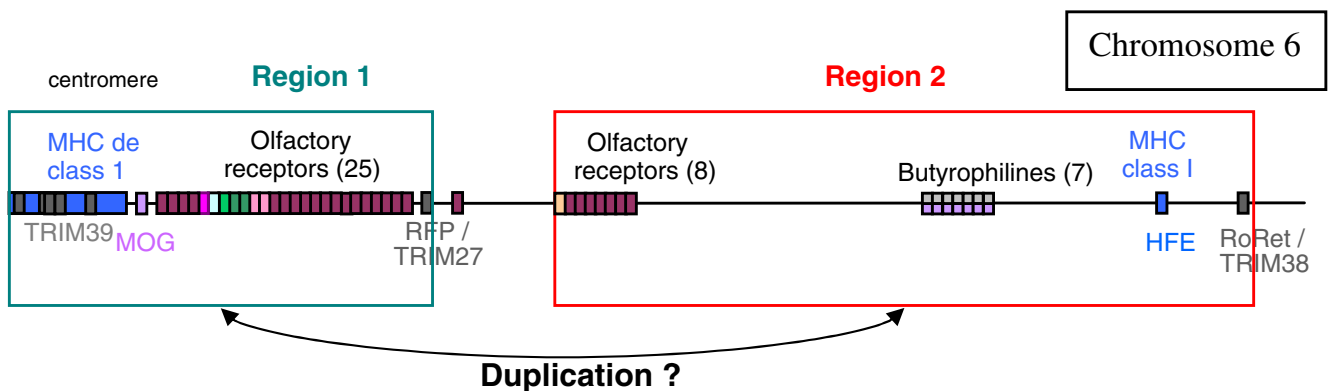


Fig. 2 Simple description of human MHC distal region of chromosome 6. Region 1 (green) and region 2 (red) contain members of the same multigenic families: olfactory receptors (subfamilies 1 in beige, 2 in

purple, 5 in light pink, 10 in white, 11 in pink, 12 in green), MHC class I in blue, MOG in light purple, B30.2 in grey and butyrophilines in light purple and grey

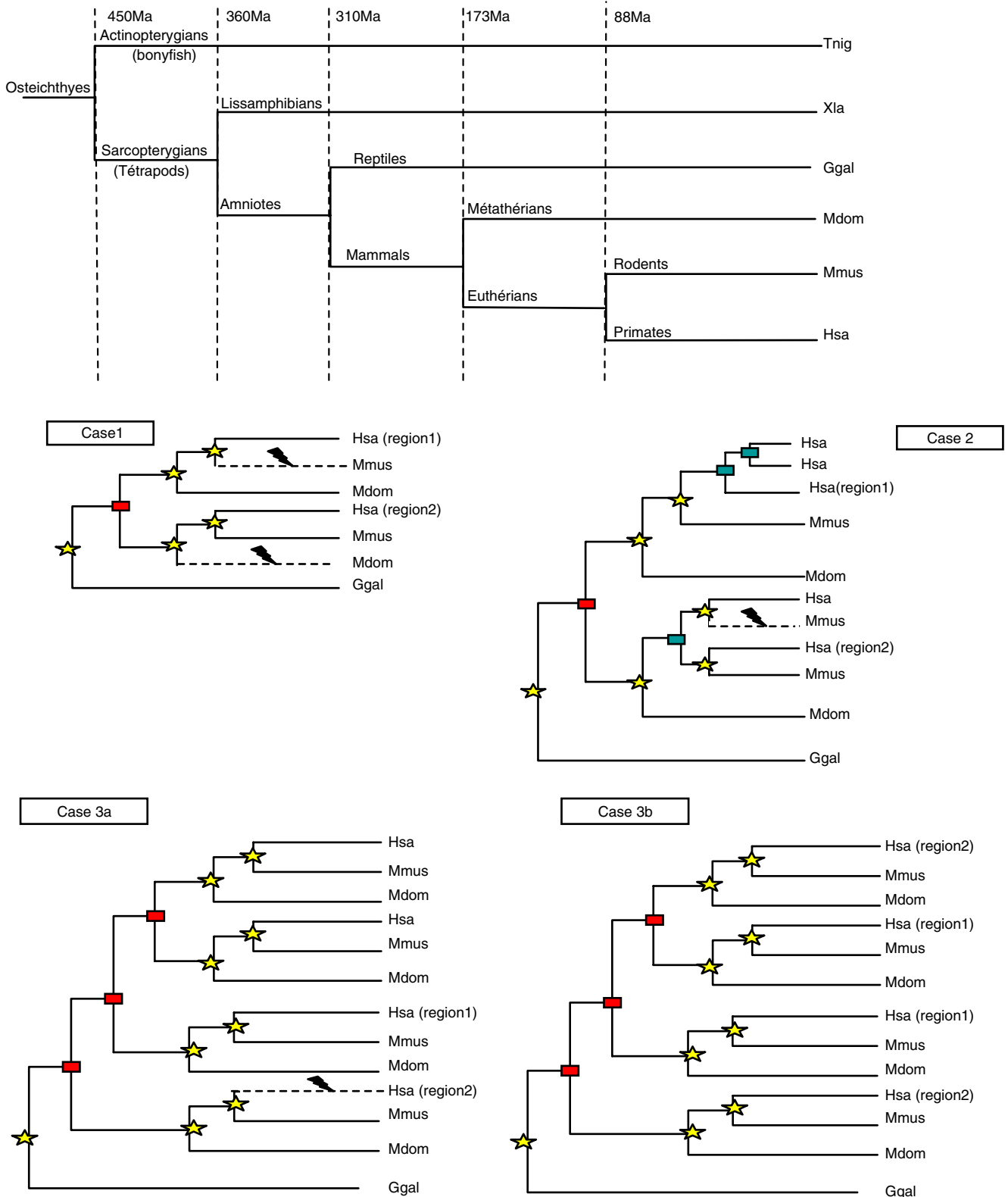


Fig. 3 Schematic representation of the different cases found in the phylogenetic trees as described in strategy (“Materials, methods and strategy”). At the top, we present the phylogenetic tree of vertebrate species. *Hsa* *H. sapiens*, *Mmus* *M. musculus*, *Mdom* *M. domestica*, *Ggal* *Gallus gallus*, *Xla* *Xenopus laevis*, *Tnig* *T. nigroviridis*. Following *Xla* and *Tnig* are noted “other vertebrates”: *yellow stars*

represent speciation events, *red squares* represent duplication events that occurred between amniotes and mammals radiations (310–173 Mya), *blue squares* represent duplication events that occurred out of the period of interest, *dotted lines and sparks* represent gene lost or rapid evolution

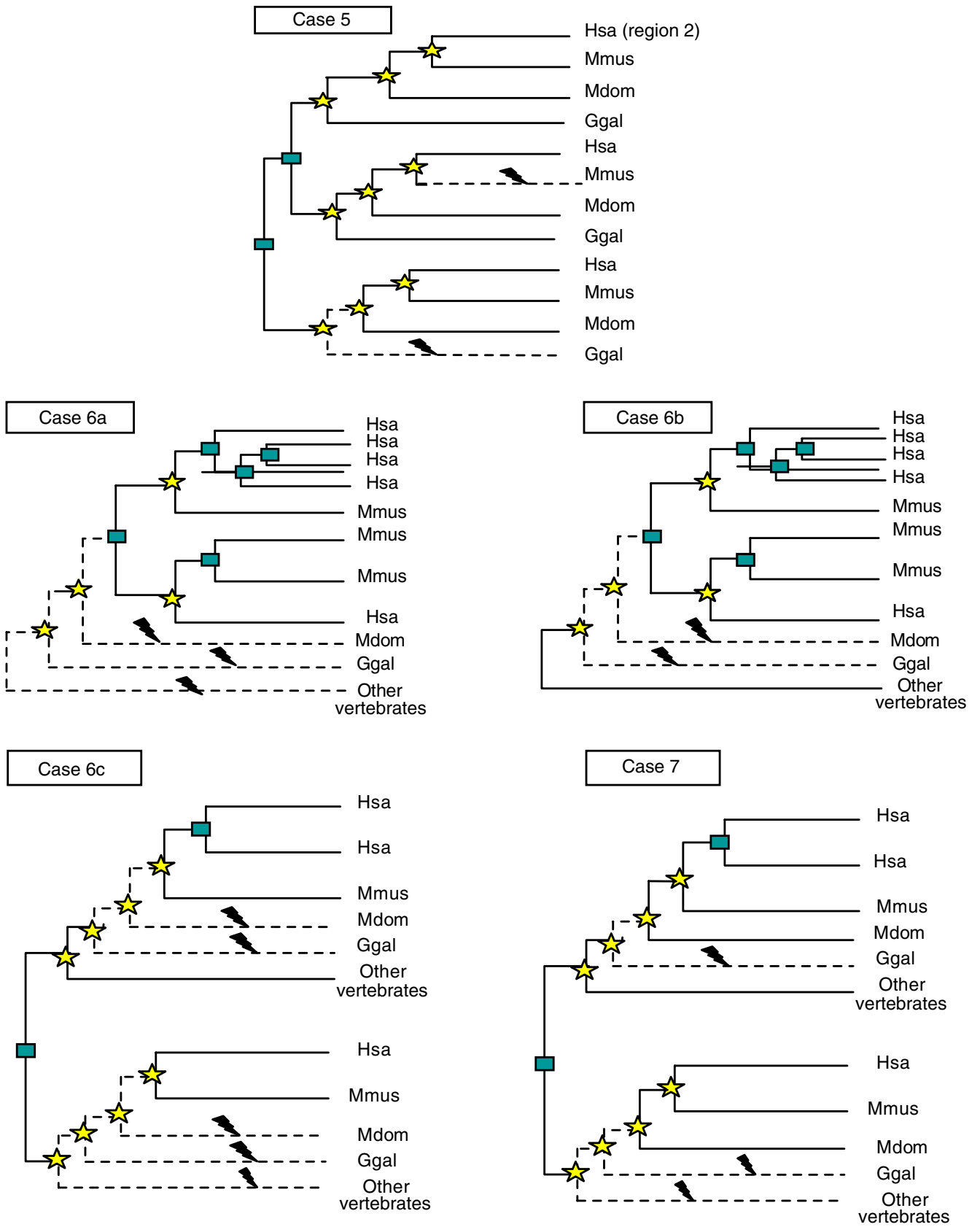


Fig. 3 (continued)

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

- Case 3 Criteria (2) and (3) are filled and phylogenetic links between members of a family are unclear, but an ortholog in *M. domestica* genome exists for each human paralog; we count a single duplication for each member of the family -1 (for example, three duplications engender four genes). Therefore, if m is the number of member engendered during the period of interest, $n=m-1$. For k parameter, two sub-cases exist: (3a) if one or more members localise in only one of the two regions even if a copy that localise in the other region could be lost, $k=0$; (3b) one or more members are localised in the two regions, $n \geq k \geq 1$, it would be determined for each case depending on the tree topology.
- Case 4 It is a particular case of case 3: phylogenies do not give us clear relations, but groups of paralogs can still be highlighted. For example, for the B30.2 protein family, six members belong to one sub-tree in all phylogenies, but links between them change. Therefore, for this group, we counted five duplications ($n=5$; k is determined in “Results” below; Table 1).
- Case 5 Criterion (3) is not filled; one or more orthologs are found in *G. gallus* genome for at least one human paralog. The duplication occurred before the amniotes radiation. Therefore, $n=k=0$.
- Case 6 No ortholog exists in *G. gallus* and *M. domestica* genomes: we used phylogenetically distant species from amniotes lineage as *Xenopus laevis*, *Tetraodon nigroviridis* or all available vertebrate genomes.
- 6a No more orthologs are found in the remote species (likely due to too rapid evolution): the duplication is suggested to have occurred after the mammalian radiation.
- 6b One ortholog is found in one or more distant species genomes located as out group(s) in the tree: we deduced that the duplication occurred after the mammalian radiation.
- 6c Orthologs are found for one or more mammalian paralogs in distant species genomes: this suggests that the duplication occurred before the amniotes radiation. In these three cases, n and k are not incremented.
- Case 7 Criterion (3) is not filled: no ortholog exists in *G. gallus*; we used phylogenetically distant species from amniotes lineage as *X. laevis*, *T. nigroviridis*

or all available vertebrate genomes. We should conclude that the duplication occurred in the period of interest but, as in the case of some members of the B30.2 proteins family, such as TRIM16/16L, TRIM62, TRIM46, at least one ortholog was found in the distant species genomes for one or more mammalian paralogs, which suggests that the duplication occurred before the amniotes radiation. In this case, $n=k=0$.

Results

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

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

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

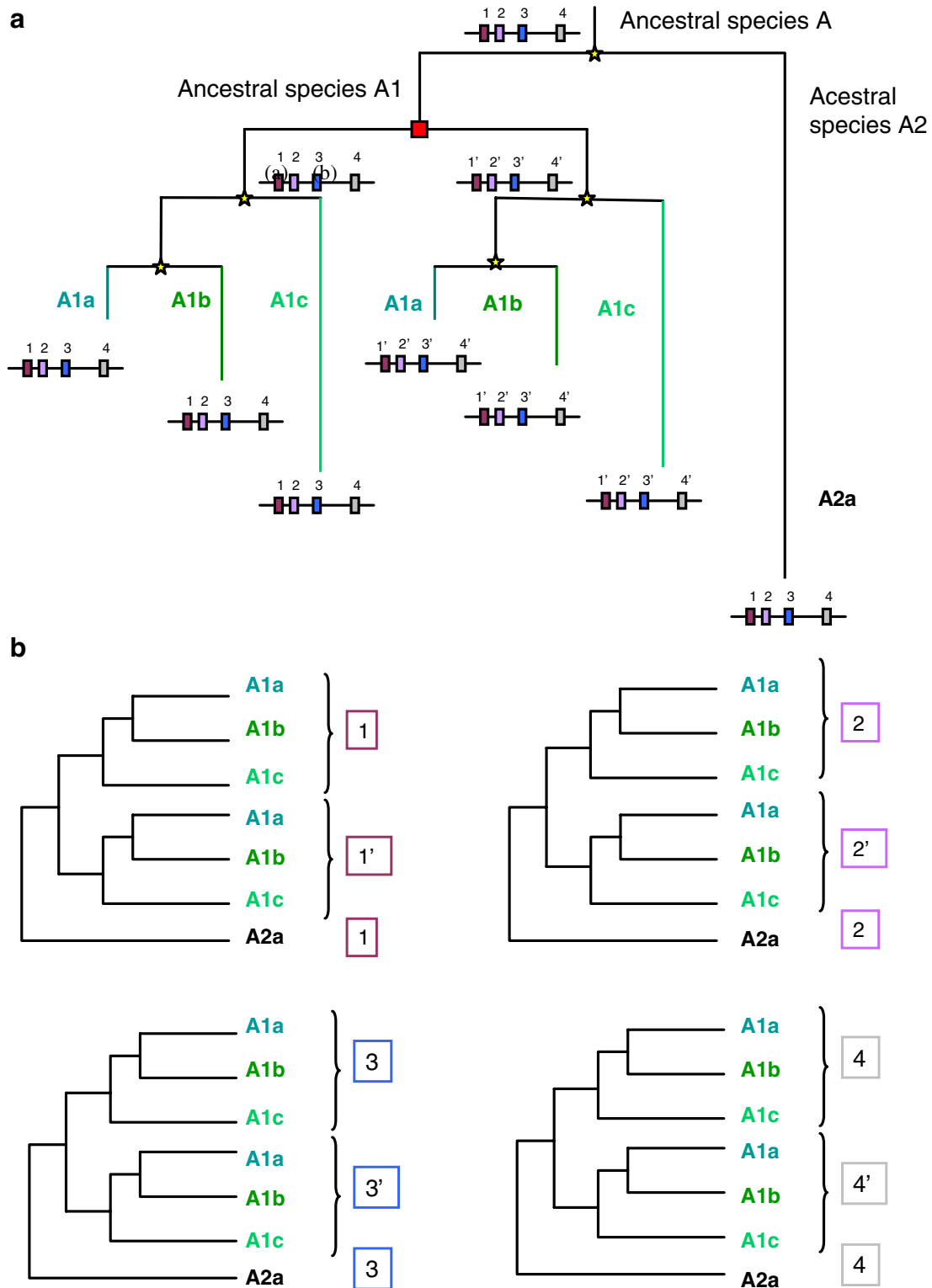


Fig. 4 “En bloc” duplication concerns a genomic segment containing several genes that has been duplicated. **A** Here, an ancestral species A contains in its genome a genomic segment with four genes (1, 2, 3 and 4). First, a speciation event (yellow stars) occurred, and species A diverged to species A1 and A2. In A1 lineage, an “en bloc” duplication occurred (red square); species A1 contains two paralogous segments (a) and (b). Genes contained in these segments are two–two paralogs (1–1'; 2–2'; 3–3', 4–4'). Two speciation events follow the duplication and species A1a, A1b and

A1c arose. These three species contains the two segments in their genomes. **B** Attempted phylogenetic results. Research of orthologs of the paralogs in given species allows dating the duplication; indeed, if genes contained in segment have duplicated in the same period, it is a first argument to prove “en bloc” duplication. Here, duplication has occurred after A1/A2 speciation and before A1a/A1b/A1c speciation. In species A2a, descending from species A2, we observed one copy of each gene. In phylogenies, these genes localise as outgroup with regard to other species



Fig. 5 OR2B4P phylogeny, consensus tree between parsimony and NJ tree. Proteins coding by genes contained in region 1 are *squared in red*, and those contained in region 2 are *squared in blue*. We can observe three duplications (1, 2 and 3) that occurred between eutherians/metatherians, separation and birds/mammals separation (presence of a *G. gallus* sequence as outgroup). Duplication 1

correspond to a *cis* duplication that occurred likely after the “en bloc” duplication, duplication 2 gave a member in the region 1, k is incremented by 1, and duplication 3 led to a paralog in region 1 but paralog in region 2 had been probably lost; so, as described in case 3a (“Materials, methods and strategy”), k is not incremented. Therefore, $n=3$, but $k=1$

Enumeration of duplication events by phylogenetic analyses

Olfactory receptors (ORs) family

ORs family is composed by several subfamilies as described previously

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

Phylogenetic studies performed on the ORs subfamily 5 highlighted 44 duplications between bird/mammal separation and placental mammal/marsupial separation. For ORs subfamily 11, 15 duplications occurred in this

period. For ORs subfamily 10, 32 duplications occurred, and for ORs subfamily 1, 17 duplications have been highlighted (Table 1).

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

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

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

BTN butyrophilins

Proteins with a B30.2 domain

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

Myelin oligodendrocyte glycoprotein

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

Butyrophilins

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

MHC class I family

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

Other genes

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

Statistical significance

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

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

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

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

From region 1 to region 2

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

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

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

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

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

α represents the risk to reject H_0 (random hypothesis materials and methods):

$$\alpha = 0.0003$$

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

From region 2 to region 1

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

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

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

$$n = 52 \quad \kappa = 5$$

α represents the risk to reject H0

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

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

Localisation of the orthologous genes in non-human mammals genomes

We have analyzed the localization of the orthologous genes from the human regions 1 and 2 in *M. musculus* and *M.*

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

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

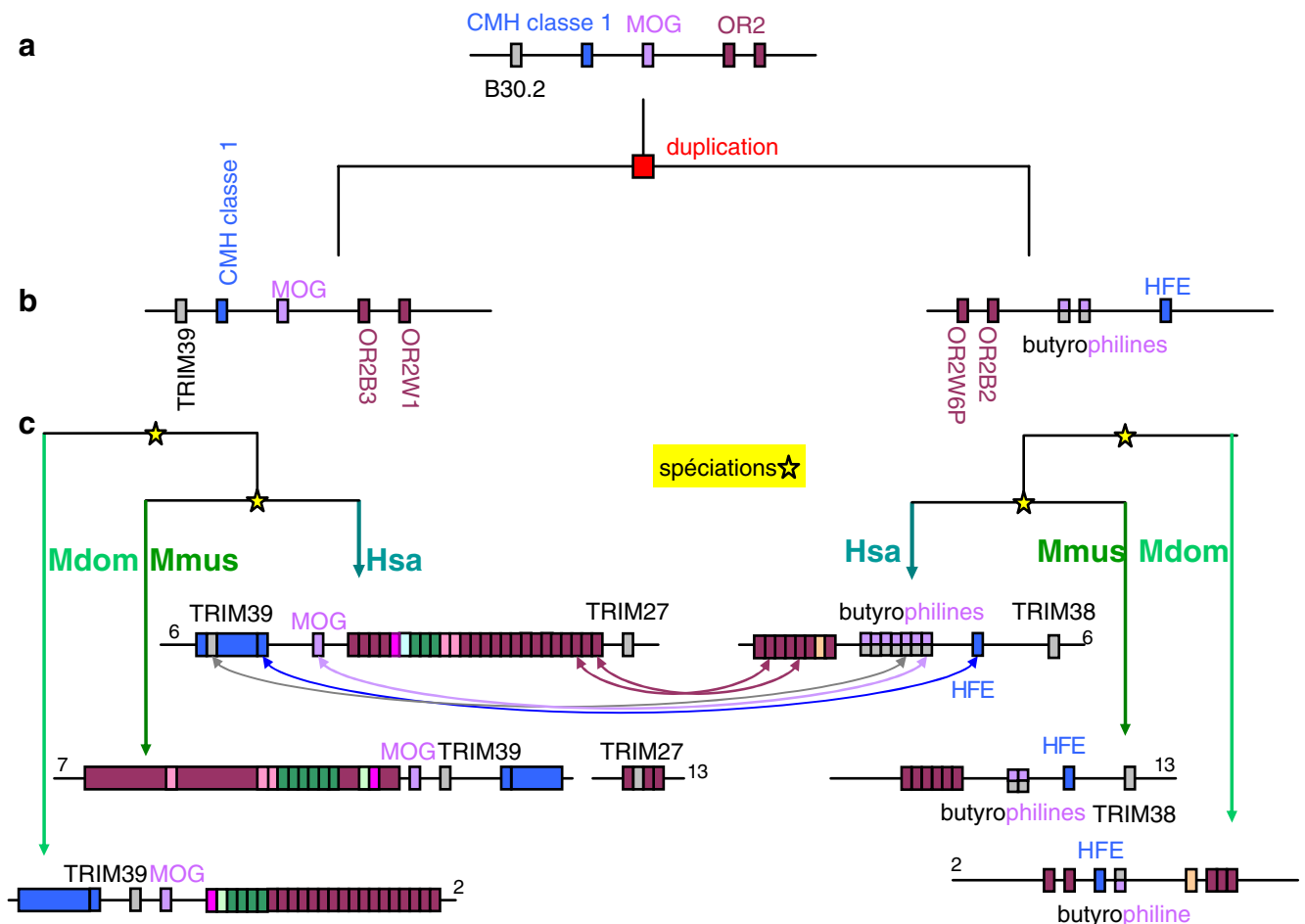


Fig. 6 Hypothesis of MHC distal region evolution in *H. sapiens* (*Hsa*), *M. musculus* (*Mmus*) and *M. domestica* (*Mdom*). **a** Unique hypothetical sequence in an ancestral species. **b** In the same ancestral species as in **a**, duplication event occurred and gave two hypothetical paralogous segments. **c** Sequences observed today in *Hsa* (number 6: chromosome 6), *Mmus* (numbers 7 and 13: chromosome 7 and 13) and *Mdom* (number 2: chromosome 2) obtained after speciation events. Colour code is the same as in previous figure. Presence of segments

with same colour in different species highlights presence of orthologs. Arrows between the two *Hsa* segments represent paralogous relationships; it is a summary of duplications that occurred between birds/mammals separation and eutherians/methaterians separation. Purple arrows represent ORs gene subfamily 2 paralogy links, grey arrows represent B30.2 genes family paralogy links, the light purple arrow represents a MOG gene family paralogy link, and the blue arrow represents the MHC class I gene family paralogy link

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

Discussion

As a result of phylogenetic and genes distribution studies about the two chromosomal regions [Chr6, 30.4 Mb (TRIM39)—28.98 Mb (TRIM27); Chr6, 28.1 Mb (ORIF12)—26,07Mb (TRIM38)], we can reject random distribution hypothesis of duplicated genes under investigation. We suggest that a co-duplication (location and period) or “en bloc” duplication happened, and therefore, we used genomic data from *M. domestica* (opossum) and *M. musculus* (mouse) genomes and we observed that orthologs of human genes are organised in clusters with a conserved synteny between these species. “En bloc” duplication probably occurred before mammals/marsupials separation. Another possibility is that independent duplication occurred from the same region in the ancestral mammalian genome and that each duplicate localised in the same region forming, a new cluster.

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

In our study, the duplication probably concerned a regional duplication because no polyploidization events are known to have occurred in mammalian lineage (Hillier et al. 2004). We showed “en bloc” duplication for one region, and it would be interesting to make this kind of analyses on the whole genome. Concerning this point, it should be noted that Dehal and Boore (2005) highlighted 80 blocks of duplications that occurred after the bony fish–

tetrapods split. Phylogenetic analyses did not provide precise dates because they used only four species, human/mouse, Fugu and Ciona that are remote species, so they do not allow dating precisely duplication events. It would be important to reanalyse these data using the protocol proposed in this article to show that these segments evolved indeed by regional duplication.

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

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

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

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
- Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* 7:R43
- Cui L, Wall PK, Leebens-Mack JH et al (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16:738–749
- Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3:e314
- Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, Martinez P, Kulski J (1999) Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol Rev* 167:275–304
- Force A, Lynch M, Bryan FB, Pickett M, Yan Y, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545

- Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* 16:805–814
- Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EG (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics* 6:198
- Henry J, Mather IH, McDermott MF, Pontarotti P (1998) B30.2-like domain proteins: update and new insights into a rapidly expanding family of proteins. *Mol Biol Evol* 15:1696–1705
- Hillier LW, Miller W, Birney E et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716
- Horton R, Wilming L, Rand V et al (2004) Gene map of the extended human MHC. *Nat Rev Genet* 5:889–899
- Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci* 86:958–962
- Hughes AL, Yeager M, Elshof AET, Chorney MJ (1999) A new taxonomy of mammalian MHC class I molecules. *Immunol* 20:22–26
- Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* 39:1361–1368
- Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428:617–624
- Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160:765–777
- Krautwurst D, Yau KW, Reed R (1998) Identification of ligands for olfactory receptors by functional expression of a receptor library. *Cell* 95:917–926
- Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature* 392:917–920
- Lundin LG (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* 16:1–19
- Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. *J Struct Funct Genomics* 3:35–44
- Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159:1789–1804
- Malik HS, Henikoff S (2003) Phylogenomics of the nucleosome. *Nat Struct Biol* 10:882–891
- McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet* 31:204–205
- Otto SP, Yong P (2002) The evolution of gene duplicates. *Adv Genet* 46:451–483
- Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197
- Reed RR (2000) Regulating olfactory receptor expression: controlling globally, acting locally. *Nat Neurosci* 7:638–639
- Rubin GM, Yandell MD, Wortman JR et al (2000) Comparative genomics of the eukaryotes. *Science* 287:2204–2215
- Tazi-Ahnini R, Henry J, Offer C, Bouissou-Bouchouata C, Mather IH, Pontarotti P (1997) Cloning, localization, and structure of new members of the butyrophilin gene family in the juxta-telomeric region of the major histocompatibility complex. *Immunogenetics* 47:55–63
- Yap MW, Nisole S, Lynch C, Stoye JP (2004) Trim5 protein restricts both HIV-1 and murine leukemia virus. *Proc Natl Acad Sci* 101:10786–10791

6.2.3 Integration of Evolutionary Biology Concepts for Functional Annotation and Automation of Complex Research in Evolution : The Multi-Agent Software System DAGOBAN

Le but de la thèse entreprise au sein du laboratoire LATP UMR 6632 CNRS Evolution biologique et Modélisation sous la direction de M. Pontarotti, était d'automatiser la détection des différents événements génétiques (mélange de domaines, duplication, perte/gain de gènes etc.) afin de pouvoir analyser l'évolution globale du protéome humain. Pour cela, en collaborant avec Philippe Gouret (ingénieur informaticien), je devais d'une part, formaliser les différents événements génétiques sous forme d'un ensemble de règles logiques simples et d'autre part, participer à la mise en place d'une ontologie décrivant les relations entre événements au niveau moléculaire et protéique.

Durant la première année, je me suis focaliser sur l'événement de mélange d'exons en analysant deux cas de mélanges d'exon décrit dans la littérature. J'ai proposé une stratégie de détection de ces événements combinant l'analyse phylogénétique des différents domaines composant les protéines d'intérêts et une comparaison au niveau génétique des séquences apomorphiques (dérivée) et plésiomorphiques (ancestrale) afin d'identifier le point de recombinaison entre les domaines. Cette stratégie a été traduite en règles logiques simples qui ont été implémentée par Philippe Gouret dans le système multi-agents DAGOBAN présenté dans le chapitre du livre suivant.

Chapter 5

Integration of Evolutionary Biology Concepts for Functional Annotation and Automation of Complex Research in Evolution: The Multi-Agent Software System DAGOBAN

Philippe Gouret, Julien Paganini, Jacques Dainat, Dorra Louati, Elodie Darbo, Pierre Pontarotti, and Anthony Levasseur

Abstract Various strategies have been proposed for predicting protein function. They are derived from the classical homology-based approaches and emerging alternative approaches taking into account gene history in the framework of phylogenetic comparative methods. The growing numbers of available genome sequences and data require bioinformatics tools, in which methodological approaches are set according to the biological issues to be addressed. Much effort has already been devoted to integrating evolutionary biology into bioinformatics tools; e.g., homology-based functional annotation has been successfully integrated in a pipeline-assisted method. In addition, new concepts based on correlation of evolutionary events are emerging. For example, two independent events (e.g., systematic loss of specific genes) that happen repetitively can therefore be functionally linked. However, correlated gene profiles, also called “contextual annotation,” makes use of different bioinformatics resources based on multi-agent development. In this chapter, we describe evolutionary concepts and bioinformatics approaches proposed for future functional inference.

P. Gouret • J. Paganini • J. Dainat • E. Darbo • P. Pontarotti
UMR6632, Evolutionary Biology and Modeling, Université de Provence, 3 place Victor Hugo,
13331 Marseille, France
e-mail: philippe.gouret@univ-provence.fr

D. Louati
UMR6632, Evolutionary Biology and Modeling, Université de Provence, 3 place Victor Hugo,
13331 Marseille, France

(LAMSIN-IRD) ENIT, Ecole Nationale d’Ingénieurs de Tunis BP 37, Le Belvédère 1002-Tunis,
Tunisia

A. Levasseur
INRA, UMR1163 de Biotechnologie des Champignons Filamenteux, IFR86-BAIM, Universités
de Provence et de la Méditerranée, ESIL, 163 avenue de Luminy, CP 925, 13288 Marseille Cedex
09, France

Universités Aix-Marseille 1 et 2, UMR1163, 163 avenue de Luminy, CP925, 13288 Marseille
Cedex 09, France

5.1 Functional Annotation Strategies: Current and Future Approaches

5.1.1 Homology-Based Functional Annotation

Eisen was the first to conceptually rationalize phylogenetic methods to improve the accuracy of functional predictions. In 1998, he proposed a phylogenetic prediction of gene function and compared it to similarity-based functional prediction methods (Eisen 1998). In this work, all known functions on a phylogenetic tree were overlaid. The prediction task could then be split into two steps. In the first step, the tree could be used to decipher orthology and paralogy relationships. Most of the reports based on evolutionary biology methods used ortholog information to transfer functional annotation (see Gouret et al. 2005 and Danchin et al. 2007). Functional assignment could be performed for uncharacterized proteins only if the function of an ortholog was known (and if a similar function was evidenced for all characterized orthologs). Ideally, functional inference should be carried out for experimentally validated orthologs. Bibliographic analysis indicates that orthologs are more likely to keep a similar function than paralogs (e.g., Collette et al. 2003). Theoretically, after duplication, one of the copies is lost, or both duplicates undergo subfunctionalization, or one of the duplicates evolves toward a new function (Force et al. 1999). However, Studer has challenged this assumption, as orthologs and paralogs could have comparable mechanisms of divergence (Studer and Robinson-Rechavi 2009). Different and more complex fates of duplicates could also be evidenced (for a review, see Levasseur and Pontarotti 2011).

In the second step, parsimony reconstruction or alternative reconstructive propagation methods could be used to assign functions of uncharacterized genes by identifying the evolutionary scenario that requires the fewest functional changes over time. Inference of ancestral state on phylogenetic tree requires that character mapping be accurate. Uncertainty about trees and mapping is therefore counterbalanced by introducing Bayesian statistical methods, taking into account this inherent error parameter (Ronquist 2004).

To the best of our knowledge, the first report using both approaches was integrated in the work of Engelhardt et al. (2005). The authors constructed a model of molecular function evolution to infer function in a phylogenetic tree. The model takes into account evidence of varying quality and computes a posterior probability for every possible molecular function for each protein in the phylogeny. Different hypotheses were included in the strategy, i.e., each molecular function may evolve from any other function, and a protein's function may evolve more rapidly after duplication events than after speciation events (Engelhardt et al. 2005). Branch length and duplication are integrated in the methodological approach. In brief, methods may be summarized as propagating functional information from leaves to the root of the phylogeny and then propagating back out to the leaves of the phylogeny, based on the probabilistic model of function evolution.

Homology-based functional annotation is summarized in Fig. 5.1.

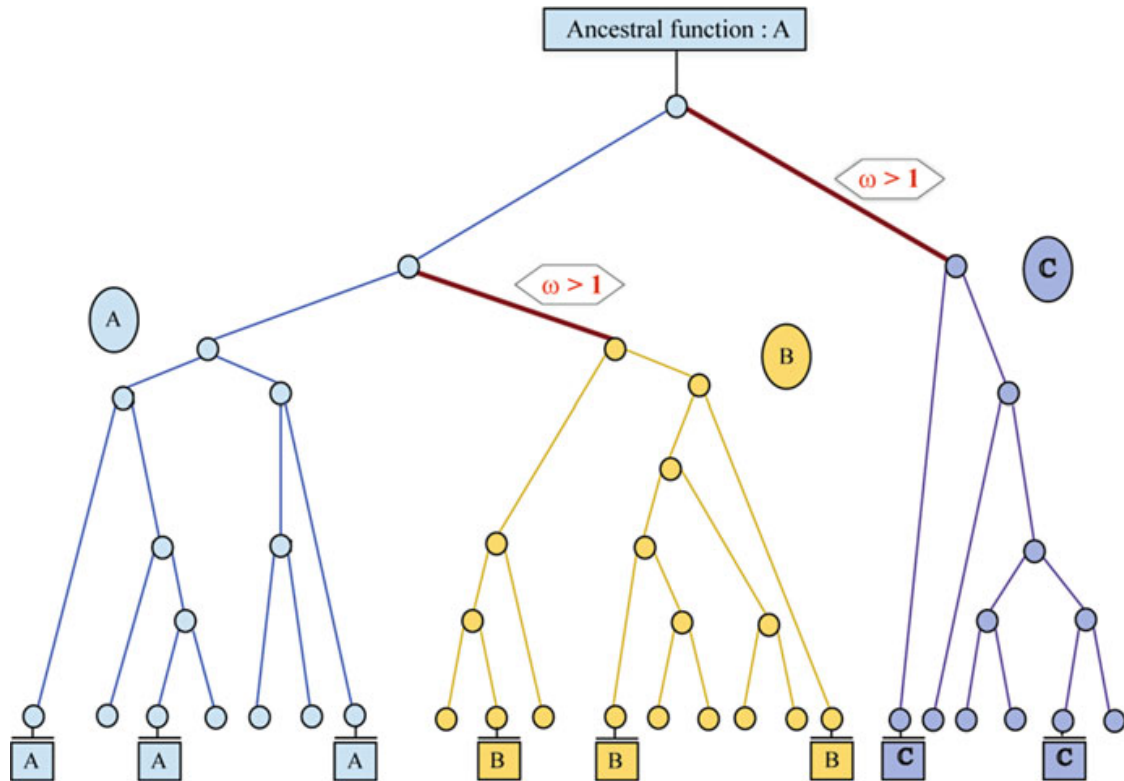


Fig. 5.1 Homology-based functional annotation. Functionally annotated leaves are labeled, respectively, as function A (blue), B (yellow), and C (dark blue). Putative function of non-annotated leaves is inferred after ancestral reconstruction based on propagation of functional information from leaves to the root of the phylogeny. Red branches: evolutionary and functional shift (using $\omega = dN/dS > 1$, i.e., Darwinian selection). (Adapted from Levasseur and Pontarotti 2008)

5.1.2 Strengthening Functional Annotation: Integration of Correlative Approaches

Functional prediction using “contextual information” is tricky because of (i) technical difficulty in detecting occurrence profiling and (ii) statistical methods required to correlate and infer function accurately. Co-occurrence and correlated gene profiles could result from phylogenetic inheritance among closely related species. Alternatively, co-occurrence could also result from individual adaptive functions, for instance when genes appear or are lost independently in several distinct lineages (Barker and Pagel 2005). Thus the probability of functional linkage between genes is proportional to the number of multiple independent phylogenetic events. A simplified example of co-occurrence and functional links is depicted in Fig. 5.2. Unlike the overall counting of presence or absence of genes, phylogenetic methods enable us to investigate ancestral states and decipher independent multiple evolutionary events.

Different methods for occurrence profiling have already been proposed, mainly on the basis of the parsimony principle and maximum likelihood (ML).

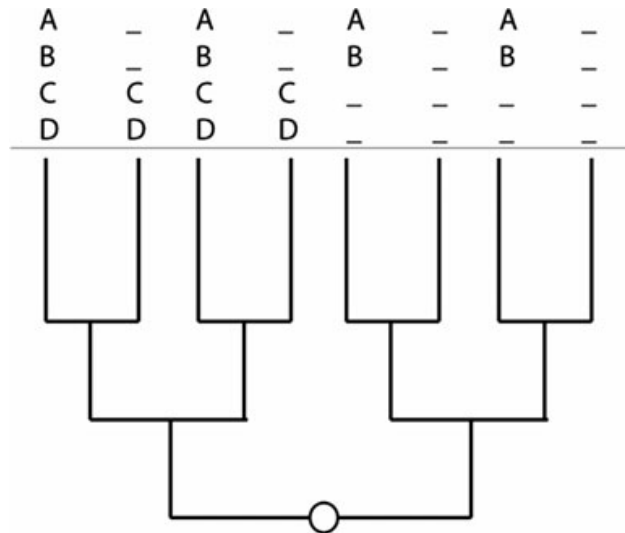


Fig. 5.2 Co-occurrence and functional link. Example of the need for comparative phylogenetic methods. Presence/absence of genes (A, B, C, D) is reported on the leaves of the phylogenetic tree. Here, multiple independent phylogenetic events of gain/loss of gene pairs (i.e., four independent events for genes A and B) are opposed to the apparent correlation arising from shared inheritance of gene pairs loss (resulting from one ancient event for genes C and D). The different steps can be summarized as follows: (i) detection of event: A is lost, (ii) convergence detection: A is lost several times, (iii) co-convergence detection: A and B are lost together several times. Subsequently, statistical tests are carried out. The function of non-annotated genes could be deduced from the correlated annotated genes

As described in the work of Barker and Pagel (2005) and Barker et al. (2007), a common pattern of presence and absence across a range of distinct genomes could be integrated as a method for detecting functionally linked proteins. Thus correlated gains and losses of genes on a phylogenetic tree of species could improve the detection of functionally linked pairs of proteins, compared with the original across-species methods from Pellegrini et al. (1999). Several phylogenetic methods were compared in their work to evaluate the accuracy of their method. Methods were based on either Dollo parsimony (Farris 1977) or ML, including a general model, but also using a constrained model in which the rate of gain of genes is not estimated from the data, but set at a low value. The fixed value of the ML should model gene content evolution better, by preventing the modeling of multiple gains of the same gene in different parts of the phylogeny. In the parsimony case, the reconstructed ancestral states could be very uncertain and parsimony could be applied when rates of changes are rather low. Note that parsimony intervals are proposed to account for the uncertainty of the parsimony methods. For instance, Zhou et al. proposed a dynamic programming algorithm to calculate such parsimony intervals. The best 100 suboptimal ancestral states were determined, and the authors compared the number of correlated events, while allowing for the degree of suboptimality of the reconstructions (Zhou et al. 2006). By contrast, ML accounts for the branch length and uncertainty of topology in the tree, and the estimate of the likelihood values is an independent parameter (i.e., corresponding to all ancestral state possibilities). The authors conclude that all the phylogenetic methods except

unconstrained ML achieved higher specificity than the across-species approach (ML model being capable of greater accuracy and sensitivity than a Dollo parsimony-based approach) (Barker et al. 2007).

5.1.3 Toward Reliable Global Functional Annotation: The Need for Bioinformatics

Bioinformatics has unlocked vast amounts of genomic data and developed software applications based on increasingly powerful mathematical algorithms – which themselves produce large volumes of results –, but the amounts of data involved simply cannot be interpreted with any real depth using statistical correlations. We therefore need to develop smart software systems able to support researchers in their efforts, which means systems automatically handling the major routine component of their *in silico* research protocols, and helping analysts interpret the huge volumes of results generated. Such smart software systems could ease the most burdensome part of the workload, leaving researchers to channel their energy into the “sharp end” of their research.

In early 2002, evolutionary biologists were handling vast quantities of biological data made available through the Internet, and running an array of software tools based on probabilistic algorithms working on these data or on data derived from other mathematical tools. The models associated with these tools were all task-specific – sequence similarity, gene prediction, phylogenetic tree-building, and so on. However, they never integrated a large number of concepts employed in biological knowledge and reasoning into a single, integrative software solution. Hence individually, they were unable to answer complex questions posed by biologists or to verify their hypotheses. Consequently, we had to automatically chain mathematical computations through what bioinformaticians call pipelines.

According to the functional annotation strategies described above, homology and correlative approaches were integrated into specific bioinformatics platforms.

A bioinformatics strategy designed for homology-based functional annotation was first implemented by creating FIGENIX (Gouret et al. 2005). FIGENIX is a Java (java.sun.com) platform that automates simple pipeline schemes, such as basic phylogenetic tree-building from a protein sequence by (i) similarity searching against protein databases, (ii) simple filtering, (iii) alignment, and (iv) tree computation. Mathematical tool chaining, through this first version of FIGENIX or any of the pipeline systems available at the time, was unable to completely automate a process: this meant that biologists still had to intervene between computation phases to verify, correct, and synthesize data output from the mathematical tools and guide the workflow to the relevant part of the pipeline. The only way to resolve this automation issue was to introduce an expert system (with Prolog language; Warren et al. 1977) into FIGENIX to model a part of biologists’ knowledge and thus act as a human scientist as and when necessary. By introducing specific logical rules in the expert system, a pipeline was created and was dedicated to gene

predictions *via* an approach combining *ab initio* predictions and homology through a lab method. Tested against a known benchmark, the pipeline clearly proved successful. A complex phylogeny pipeline with 50 steps and a lot of expertise modeling was designed. The first version was stabilized in late 2003, and has since enabled the laboratory and its collaborators to produce thousands of phylogenetic trees from protein queries. These trees form the basis of our evolutionary research. This pipeline, along with others, was intensively used on laboratory projects, generating several published papers (Danchin et al. 2004, 2006, 2007; Paillisson et al. 2007; Levasseur et al. 2006, 2010). It continued to undergo improvements and enhancements, with upgrades including automatic detection of orthologs in the final process-synthesized tree by online recovery of functional data associated with these orthologs (GO (Ashburner et al. 2000), MGI (www.informatics.jax.org), NCBI (www.ncbi.nlm.nih.gov)), and EST integration (Balandraud et al. 2005). Part of the software developed, called PhyloPattern, emerged as a crucial independent component (Gouret et al. 2009). The aim of this tool was to reproduce human reading of phylogenetic trees, i.e., phylogenetic tree annotation and pattern recognition. Inside the phylogeny pipeline, this tool is used to detect incongruence or isolate specific subtrees, from which biases are then corrected. PhyloPattern now makes it possible to detect events in the history of species, genes, or any other characteristic (from domain to function and further), as well as highlighting artifacts in the phylogenetic trees. We are continuing to improve PhyloPattern as a free open-source JAVA/Prolog API.

5.2 From Pipelines to Multi-Agent Strategies

In 2005, it became clear that the “pipeline approach,” even with the controlling expertise introduced, remained limited to computation processes. In addition, functional annotation using the correlative approaches strategy required flexible and more sophisticated data processing architecture. Computation processes are essential, but are not really able to resolve complex tasks of interest to the laboratory, such as automatically highlighting genetic events in the human genome and detecting convergences and co-convergences among these events. Any solution to these issues needs to be driven by expertise through parallel and more “intelligent” processes than the rigid, deterministic pipelines. We also note that the “pipeline approach” does not extend to establishing an explicitly described semantic universe that would allow accurate meta descriptions of data. It thus remains impossible to raise the abstraction level of software tasks, and interfacing them with other software systems is not natural.

Integration of correlated gene profiles for functional annotation requires a three-step process: (i) specific detection of all evolutionary events, (ii) correlation using phylogenetic comparative methods leading to a compelling statistical results, and (iii) deducing the function of non-annotated genes from the correlated annotated genes.

5.3 Technical System Specifications

Accordingly, a new software system was conceived and is able to implement complete automation of actual full research via bottom-up (from biological data) strategies specified by the laboratory, rather than “just” complex computation workflows. We opted for the following research strategy: (i) working from known or computed features to find evidence for generating new hypotheses, (ii) attempting to verify hypotheses to transform them into features, (iii) correlating verified features to deduce new features, and so on. A set of characteristic specifications was drawn up:

- The treatments had to be flexible, modular, and parallelized.
- The strategies for identifying and verifying the facts had to be led by expertise.
- Communication with external software systems (online databases, web services) should systematically gather the relevant results produced by these platforms, such as Ensembl (Hubbard et al. 2009), NCBI, String (Szklarczyk et al. 2011), and ArrayExpress (Parkinson et al. 2011).
- The results had to be placed in an accurately described semantic universe that was not redundant but interfaced with data from external systems.
- Some modules had to work together and communicate directly, while others, such as modules for intelligent correlations of events, had to work in stand-alone mode directly on the mass of results produced by the full set of modules.
- The modules had also be able to work at different times.
- The system had to be resistant to failure; as such, very costly computational treatments should have to be run only once.

5.4 Technical State of the Art

The field of biology now has a number of software tools, approaches, standards, and publications that could be recycled for our needs. The type of system targeted here required establishing an integrated data model, placed between structured biological data (e.g., genomic databases) or unstructured data (publications) located inside or outside the laboratory, and the research strategies desired by laboratory researchers. Software systems clearly have to work with large-scale data banks, but what is most important now is to work with different kinds of data, many of which are not a direct representation of biological objects but are more abstract concepts.

We could therefore rule out relational database management systems, which are not powerful enough or flexible enough to describe semantics in biology. Some recently developed software tools such as the alignment expert system ALEXSYS (Aniba et al. 2009) are based on the UIMA framework (<http://sourceforge.net/projects/uima-framework/>), which offers a powerful architecture and is well-suited to the introduction of a virtual model on unstructured data, i.e., building meta-information from artifacts such as scientific publications (also see DiscoveryLink

(Hass et al. 2001) or BioMOBY (Wilkinson and Links 2002)). We are more focused on trying to directly model actual genomics or evolutionary concepts. Also, the UIMA approach is only “object-oriented,” and we believe that this kind of modeling architecture is not rich enough to integrate the complexity of biological paradigms, especially compared with approaches based on mathematical first-order logic ontology techniques such as Description Logic (<http://dl.kr.org/>). The W3C-standardized OWL language (<http://www.w3.org/TR/owl-ref/>) is an XML representation of DL. Initially applied to the semantic web, it is fast becoming a standard for ontology modeling. In DL, relations between classes are not limited to aggregation or inheritance links but can be formalized with logical formulae. However, we note that DL does not integrate concepts of inductive, temporal, or fuzzy logic, which in the long term could direct the natural extension of our systems.

Biology now has many ontologies (e.g., NCI Cancer Ontology: <http://www.mindswap.org/2003/CancerOntology/>). Some are defined in OWL but to our knowledge, none computationally exploit the descriptive capacity of description logic (DL). This situation is surely set to change. We note the existence of relational ontology (Smith et al. 2005), placed between “object” modeling and DL modeling, which attempts to standardize relations in biological ontologies. This point will be revisited below. There appears to be a continuing dichotomy between the activity of defining ontologies, considered as vocabularies by many biologists, and the establishment of DL-based software and databases within and between laboratories or institutes. We believe that this dichotomy is an error, as it has very adverse repercussions, such as poor software systems and bad interoperability.

As stated above, to fully automate *in silico* research strategies, the type of system we are targeting has to be less rigid and deterministic than pipelines. A natural candidate solution would be multi-agent systems. In bioinformatics, these systems are used essentially to model and simulate biological networks (reactive agents), although they are also used to parallelize mathematical computations through agents with very fine granularity. They are rarely employed for building integrative applications where “smart” agents work with biological information. Nevertheless, like the FIPA institute (<http://www.fipa.org/>), we are convinced that this kind of architecture built from cognitive agents (with large granularity) communicating inside an ontological semantic universe can be applied to bioinformatics automation. The JADE software framework (<http://jade.tilab.com/>) is a Java implementation of FIPA specifications. At our lab, we used JADE to develop a first prototype multi-agent system named CASSIOPE (Rascol et al. 2009), dedicated to highlighting conserved synteny.

Recently, eHive emerged from EBI as a new workflow system (Severin et al. 2010). It is built as a multi-agent “blackboard” architecture. Here, the blackboard, i.e., the communication area between agents, is reduced to chaining rules between agents. Thus the tasks produced by the system are driven by predefined functional relations between agents and not by the autonomous interpretation, by agents, of the data resulting from other agents’ work. The Ehive blackboard database has a rigid structure with no data modeling. Also, agents’ source code is written with the Perl

language, which albeit very widely used in bioinformatics remains very poor in expertise and knowledge modeling.

As stated earlier, we are seeking to deploy expertise-driven research strategies, which means that all agents need to be built with expert-system architectures. Rule engines do exist – one example is Jess (www.jessrules.com) – but it would be preferable to write our own engine in Prolog language to reap the benefit of tools we developed previously, especially PhyloPattern. After years of hands-on experience, we can confirm that the Prolog language is very well-suited to bioinformatics. Its benefits for the target system include: (i) a natural capacity to generate all the solutions for a question, (ii) easy and native manipulations of lists and tree structures, which are intensively used in bioinformatics data, (iii) development of expert systems in backward- and/or forward-chaining mode (verification and/or production of facts), (iv) formalisms (e.g., ontological relations) representable directly in the language’s syntax, (v) brevity and simplicity of knowledge descriptions, and (vi) interpreted language that strengthens the experimental aspect of certain developments.

5.5 System Architecture

Our system was called DAGOBAB. It is shaped as a multi-agent software (see Fig. 5.3), with a voluntarily hybrid model summing of a model called “Belief Desire Intention” with a model called “Blackboard” (Ferber 1995). The BDI model is suitable for cognitive agents with high granularity and therefore high “intelligence.” In the BDI model, agents have a plan formed for our purposes by logical rules. This highly flexible rule system is used by each agent to implement a specific strategy, but can also be used as a traditional expert system to produce high-level facts deduced from simpler facts. For example, an agent capable of sifting through actions to detect several equally probable genetic events from a phylogenetic tree will be able to retain only one event, through a set of logical rules associated with a set of criteria.

The semantics for one rule is defined as follows:

$$\begin{aligned} & \cdot \text{Action}_1 \dots \text{Action}_k \\ & \text{ConditionFact}_1 \dots \text{ConditionFact}_n \rightarrow \text{ConclusionFact}_1 \dots \text{ConclusionFact}_m \\ & \text{ToBeRemovedFact}_1 \dots \text{ToBeRemovedFact}_z \end{aligned}$$

The meaning is “if all condition facts (n) are known by the agent (\subset Belief) and if at least one of the conclusion facts (m) is not present and if the agent is capable of achieving all actions (k) (\subset Intention) successfully, then all conclusions (m) (\subset Desire) are considered truthful, and all indicated facts (z) are removed from the agent’s knowledge.”

Here is an example rule, used in the DAGOBAB agent dedicated to searching for domain architecture events. We suppose that for a specific protein with the domain architecture A-B-C, DAGOBAB detects an event that produced the B-C part of the architecture by analyzing the phylogenetic tree of domain B, and we suppose

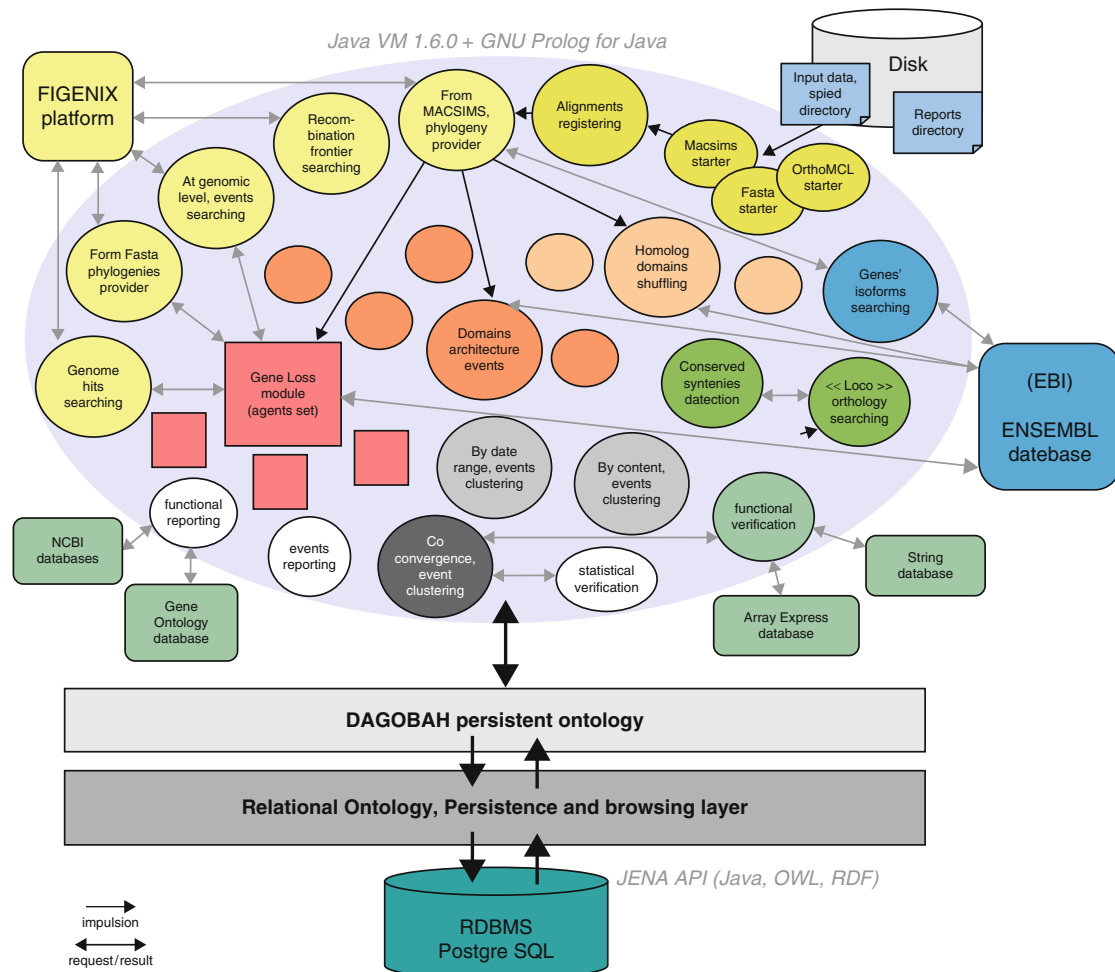


Fig. 5.3 DAGOBAH multi-agents system architecture. All agents (*disks*) or modules (*squares*) that compose DAGOBAH are contained in the large blue ovoid. Around it are displayed the external software systems interacting with the agents by the network. At the bottom of the scheme is shown the ontological database, containing the biological results produced and shared by the agents

that DAGOBAH hesitates between identifying the event as a shuffling or a gain. A simple rule, if it is applicable, allows DAGOBAH to definitely assert there is a gain (see Fig. 5.4):

- $verify_similarity_of_signal_between(P1, P2, [B, C])$
- $event_found_under_ancestral_node(N),$
- $apomorphic_chosen_protein(P1, [A, B, C]), \rightarrow gain_event_found(N, P1, [C])$
- $plesiomorphic_chosen_protein(P2, [A, B])$
- $event_found_under_ancestral_node(N)$

The “Blackboard” model introduces an area of information shared by agents, i.e., any important result produced by an agent is placed on the blackboard. The blackboard architectural model chosen in DAGOBAH is defined as a persistent ontology (an ontological database) representing the semantic universe in which the agents work. These results are used by other agents, unless they are forced to

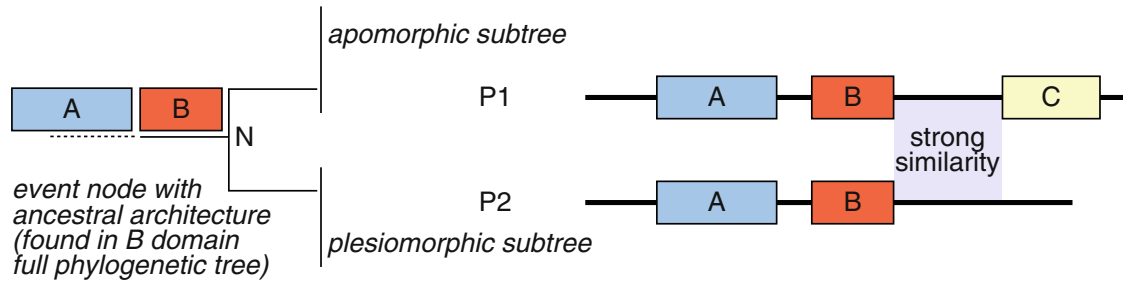


Fig. 5.4 A virtual example for a domains architecture event. Here again event is confirmed because the genomic signal between domains B and C on the apomorphic sequence is strongly conserved after domain B on the plesiomorphic sequence

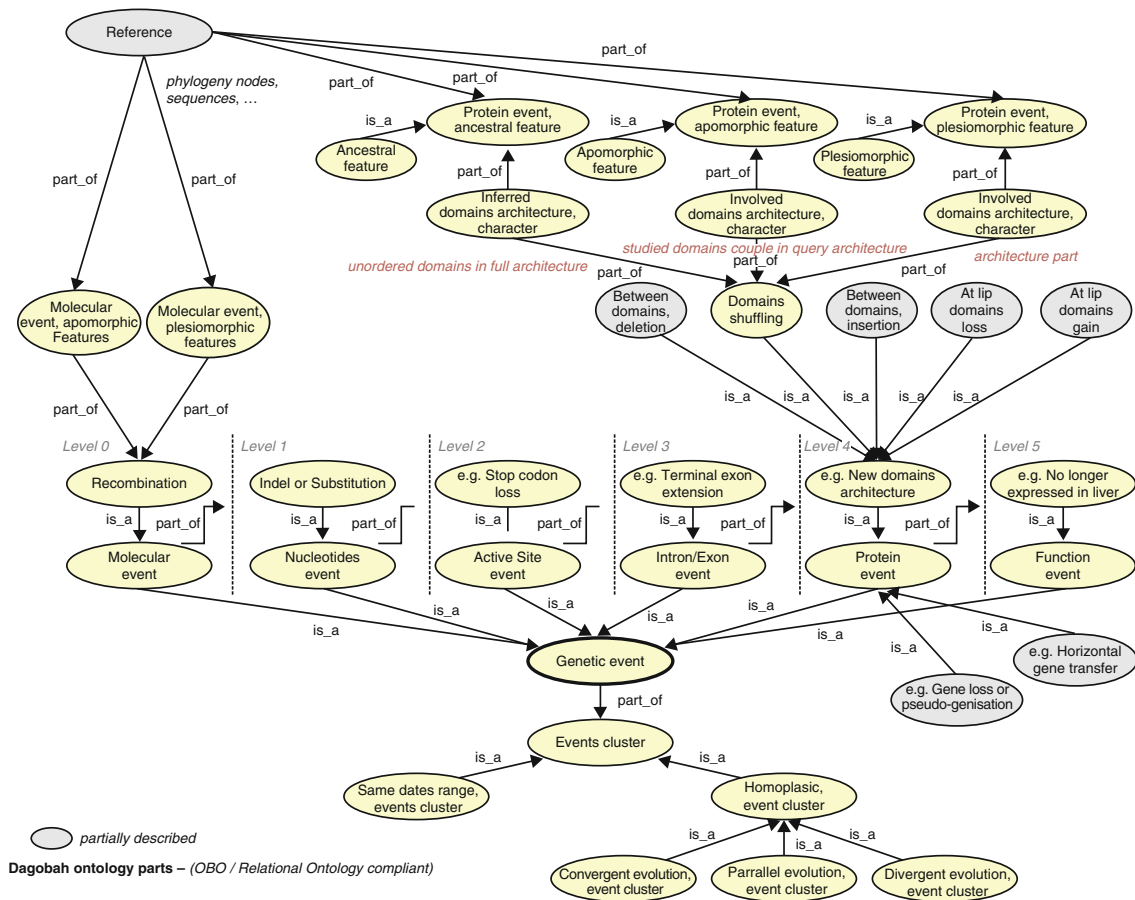


Fig. 5.5 The core of DAGOBAH ontology. Some genetic event classes laid out by their reading level are presented. As an example, we give all the classes participating in a nonhomologous domain shuffling event, induced by a recombination event. Clustering classes are also displayed with their inheritance relationships

explicitly and systematically exchange them. Figure 5.5 illustrates the main parts of the DAGOBAH ontology. Genetic event classes are grouped by reading level. For example, a recombination event can be described at a “protein” level if we are talking about domains involved in recombination, but also at a “molecular” level if we are talking about the position of the recombination on a chromosomal region. Ancestral, apomorphic, and plesiomorphic features associated with an event are

always explicitly expressed. This model is also particularly well-suited to studying automatic correlations of genetic events, and is able to correlate several events detected by DAGOBDAH and temporally localized between speciation event pairs. For example, DAGOBDAH may find that two genes A and B are lost twice “together” for two different lineages, which could prove very interesting in a functional perspective. In this case, if the “function” of gene A is known and the “function” of B is not, we can assume that the B gene may be involved in the “same” function as A. “By Dates” event clusters and homoplastic event clusters are the sources of a co-convergent event clustering process in DAGOBDAH. For example, a “convergent evolution event cluster” is produced for events that have the same apomorphic feature objects.

The DAGOBDAH ontological database must not have redundancy vs. external databases (like Ensembl; Hubbard et al. 2009). Consequently, we only model, by classes and relations, those concepts associated with specific laboratory research themes, and references were kept only to biological data or results held in external databases. The current DAGOBDAH ontology adopts the Relational Ontology standard, although in the future we will probably abandon this standard so as to fully exploit the capabilities of Description Logic.

5.6 DAGOBDAH Functionalities and Summarized Strategies

As described in Fig. 5.2, the strategies used in DAGOBDAH can be conceptually subdivided into these different steps: (i) detection of evolutionary events, i.e., gain or loss of genes, shuffling, etc. (ii) detection of convergence between one or more gene pairs, (iii) detection of co-convergence between linked genes, (iv) search for functionally annotated gene and infer the function of correlated non-annotated gene. These four steps can be considered as forming the core of the phylogenetic comparative methods.

5.7 Detection of Events (New Architecture Appearance)

The current DAGOBDAH version offers a broad panel of functions, ranging from automatic detection of genetic events to homologous domain shuffling, nonhomologous domain shuffling, insertion, deletion, gain and loss, plus gene losses and pseudogenization, and on to horizontal gene transfer and duplications (compilation on gene and species trees). A simplified summary of DAGOBDAH’s general strategy for event detection is:

1. Use “domain-annotated” protein alignments built from a query protein to outsource phylogeny trees building (domain trees and protein trees) to the FIGENIX platform.
2. Automatically read these trees with PhyloPattern to highlight possible events.

3. Seek to verify and clarify the putative events at a genomic level.

For new protein domain architecture events, actual examples of putative events in trees are given in the PhyloPattern publication. For this kind of event, a dedicated DAGOBAB agent studies each consecutive domain pair in the query protein architecture to investigate whether the association is the result of an event. Ideally, it finds an event's phylogenetic pattern (see Fig. 5.4) on each domain phylogenetic tree, which strengthens the event hypothesis.

The full confirmation of the event is achieved at genomic level by searching for an alignment break position between two DNA segments – one associated with the most representative apomorphic sequence and the other associated with the most representative plesiomorphic sequence. DNA segments are extracted between the domains involved (see Fig. 5.6). The most representative apomorphic sequence is chosen as the one nearest the parent node (the agent uses neighbor joining for branch lengths), while the most representative plesiomorphic sequence is chosen as the one whose domain architecture is closest to the ancestral node architecture (Dollo, Sankoff, and Mirkin parsimony algorithms (Sankoff 1975; Farris 1977; Mirkin et al. 2003) are integrated into PhyloPattern and used by the agent to infer ancestral domain architectures). If several plesiomorphic sequences share the same architecture comparison “score,” the agent chooses a sequence from the nearest species in the species tree.

Gene losses and pseudogenization are studied by a set of agents in DAGOBAB, which form a module named GeneLoss. It starts the study by searching for missing species in the biggest ortholog group of the query protein tree. Each species is then studied by independent agents.

Describing the strategy in schematic terms, agents set out to determine whether the species is really missing, whether a new gene should be annotated, or whether there are some mutations or indels that can explain a pseudogenization process.

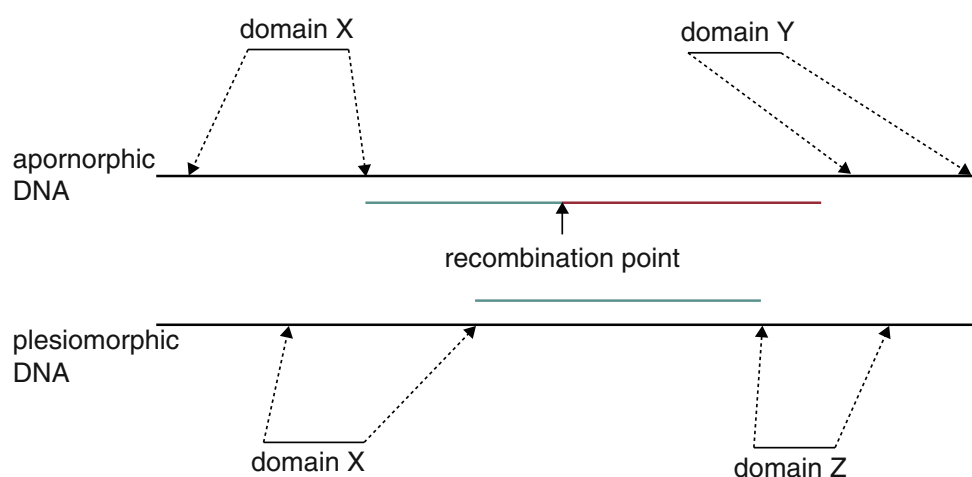
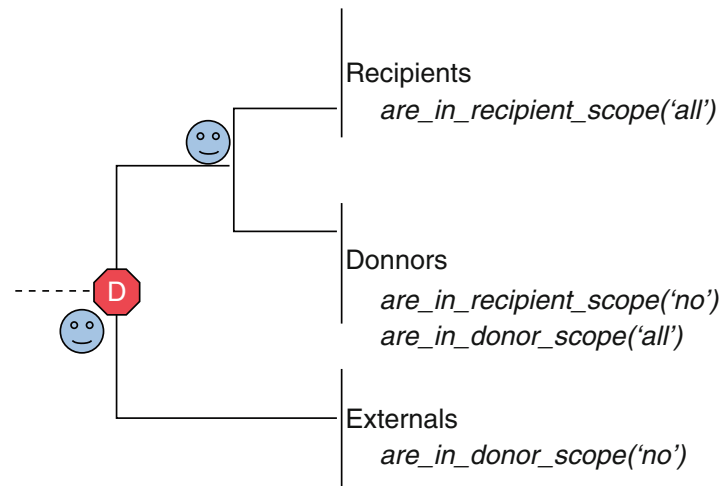


Fig. 5.6 Summary of the verification of a domain new architecture event at a genomic level. The DNA segments between domains on the apomorphic and the plesiomorphic sequences are intelligently extracted from chromosomes or scaffolds; they are then aligned and the recombination point is searched for as an alignment break

Fig. 5.7 A pattern to detect horizontal gene transfers from a phylogenetic gene tree. This means a duplication node, because the subtree does not have to match the species tree. The “donor” subtree must contain only species of a specific scope, and not from the “recipient” scope and *vice versa*



Full complex GeneLoss module strategy and results will be published separately at a later date.

Horizontal gene transfer events are detected from the query protein tree. A recipient species scope and a donor species scope are defined so as to orient the search. The dedicated agent uses PhyloPattern to annotate each internal node of the tree with two tags: `are_in_recipient_scope_species` and `are_in_donor_scope_species`, which can take three values: “no” if no species of a subtree falls in a scope, “some” if some species of a subtree fall in a scope, or “all” if all the species of a subtree fall in a scope. Then, *via* PhyloPattern, the agent applies a specific phylogenetic pattern (see Fig. 5.7) that directly gives the branch with potential HGT events.

The expert idea behind this pattern is to search the gene tree to find recipient species closer to donor species than other species that are normally placed between the recipient and donor species in the species tree.

5.8 Convergence and Co-Convergence Detection

Another important function in DAGOBAH is event convergence and co-convergence detection as conceptually described in the correlative approaches described above. Convergence identification is easy to obtain from the DAGOBAH ontological database, as a dedicated agent groups events into homoplasic convergent clusters. For example, two events are in the same convergent cluster if they have the same apomorphic character. The definition of an apomorphic character can easily be user-defined as a Prolog “ontological” pattern. The clustering mechanism is independent of the pattern definition. Co-convergence detection is a more complex task. It starts by homoplasic clustering, after which an agent produces date range clustering. Inside DAGOBAH, events are dated with tuples:

[TaxidSpeciationBefore, NumberOfDuplicationsBefore, NumberOfDuplicationsAfter, TaxidSpeciationAfter]

This tuple is determined by taking the nearest speciation event (SBE) before the event (E) on its parent branch. `NumberOfDuplicationsBefore` equals the number of duplication events on the branch between SBE and E. `TaxidSpeciationBefore` is the common parent taxid of all species in the SBE subtree. The same approach is then reapplied for the next speciation event. Date range clustering is also “user-defined” through date range patterns. Two events whose dates fit the same date pattern are pooled in the same date range cluster.

Co-convergence clusters are built with a hierarchical clustering method. A minimum co-convergent cluster is formed by four events: Eh1, Eh2, Eh1', Eh2'. Eh1 and Eh1' have to be in the same homoplastic cluster, while Eh2 and Eh2' have to be in another homoplastic cluster. Eh1 and Eh2 have to be in the same date range cluster, while Eh1' and Eh2' have to be in another date range cluster.

We can model this basic cluster as a square:

```
--- Eh1, Eh2,  
--- Eh1', Eh2'
```

The clusters can be rectangular, if they come from more date clusters than homoplastic clusters (shape 1) or the opposite (shape 2). The hierarchical clustering method enables us to build the biggest possible clusters, and implies the definition of a distance method between two clusters. Our distance method favors clusters with shape 1 rather than shape 2.

Once the biggest clusters are determined, the agents seek to verify them, both statistically, *via* the Pagel method (Pagel 1994), and functionally, using the String database (Szklarczyk et al. 2011) to see whether proteins associated with events in the same homoplastic cluster belong to the same protein interactions network, and using the ArrayExpress database (Parkinson et al. 2011) to see whether proteins associated with events in the same homoplastic cluster concern the same expression experiments.

In conclusion, DAGOBAH is designed to exploit the modern functional annotation strategies and specially the evolutionary-based biology concepts. In addition, it could be addressed to various general biological questions such as searches of conserved synteny regions from a given region associated to a species to another target species.

All public results produced by DAGOBAH are openly available on the IODA Web site (<http://ioda.univ-provence.fr/>).

Acknowledgments This research was supported by the contract MIE (Maladies Infectieuses Emergentes-Programme Interdisciplinaire, CNRS) and ANR EvolHHuPro (ANR-07-BLAN-0054-01).

References

Aniba MR, Siguenza S, Friedrich A, Plewniak F, Poch O, Marchler-Bauer A, Thompson JD (2009) Knowledge-based expert systems and a proof-of-concept case study for multiple sequence alignment construction and analysis. *Brief Bioinform* 10:11–23

- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 25:25–29
- Balandraud N, Gouret P, Danchin EG, Blanc M, Zinn D, Roudier J, Pontarotti P (2005) A rigorous method for multigenic families' functional annotation: the peptidyl arginine deiminase (PADs) proteins family example. *BMC Genomics* 6:153
- Barker D, Pagel M (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol* 1:e3
- Barker D, Meade A, Pagel M (2007) Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* 23:14–20
- Collette Y, Gilles A, Pontarotti P, Olive D (2003) A co-evolution perspective of the TNFSF and TNFRSF families in the immune system. *Trends Immunol* 24:387–394
- Danchin E, Vitiello V, Vienne A, Richard O, Gouret P, McDermott MF, Pontarotti P (2004) The major histocompatibility complex origin. *Immunol Rev* 198:216–232
- Danchin EG, Gouret P, Pontarotti P (2006) Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. *BMC Evol Biol* 6:5
- Danchin EG, Levasseur A, Rascol VL, Gouret P, Pontarotti P (2007) The use of evolutionary biology concepts for genome annotation. *J Exp Zool B Mol Dev Evol* 308:26–36
- Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* 8:163–167
- Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol* 1:e45
- Farris JS (1977) Phylogenetic analysis under Dollo's law. *Syst Zool* 26:77–88
- Ferber J (1995) Les systèmes multi-agents. InterEdition, Paris
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545
- Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EG (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinform* 6:198
- Gouret P, Thompson JD, Pontarotti P (2009) PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. *BMC Bioinform* 19 10:298
- Haas LM, Schwarz, Kodali P, Kotlar E, Rice JE, Swope WC (2001) DiscoveryLink: A system for integrated access to life sciences data sources. *IBMSJ* 40:489–511.
- Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009) Ensembl. *Nucleic Acids Res* 37:D690–D697
- Levasseur A, Pontarotti P (2008) An overview of evolutionary biology concepts for functional annotation: advances and challenges. In: Pontarotti P (ed) *Evolutionary biology from concept to application*. Springer, Berlin, pp 209–215
- Levasseur A, Pontarotti P (2011) The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biol Direct* 6:11
- Levasseur A, Gouret P, Lesage-Meessen L, Asther M, Asther M, Record E, Pontarotti P (2006) Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase a family. *BMC Evol Biol* 6:92
- Levasseur A, Saloheimo M, Navarro D, Andberg M, Pontarotti P, Kruus K, Record E (2010) Exploring laccase-like multicopper oxidase genes from the ascomycete trichoderma reesei: a functional, phylogenetic and evolutionary study. *BMC Biochem* 11:32

- Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3:2
- Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc R Soc Lond B* 255:37–45
- Paillisson A, Levasseur A, Gouret P, Callebaut I, Bontoux M, Pontarotti P, Monget P (2007) Bromodomain testis-specific protein is expressed in mouse oocyte and evolves faster than its ubiquitously expressed paralogs BRD2, -3, and -4. *Genomics* 89:215–223
- Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res* 39:D1002–D1004
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96:4285–4288
- Rascol VL, Levasseur A, Chabrol O, Grusea S, Gouret P, Danchin EG, Pontarotti P (2009) CASSIOPE: an expert system for conserved regions searches. *BMC Bioinform* 10:284
- Ronquist F (2004) Bayesian inference of character evolution. *Trends Ecol Evol* 19:475–481
- Sankoff D (1975) Minimal mutation trees of sequences. *SIAM J Appl Math* 28:35–42
- Severin J, Beal K, Vilella AJ, Fitzgerald S, Schuster M, Gordon L, Ureta-Vidal A, Flicek P, Herrero J (2010) eHive: an artificial intelligence workflow system for genomic analysis. *BMC Bioinform* 11:240
- Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C (2005) Relations in biomedical ontologies. *Genome Biol* 6:R46
- Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* 25:210–216
- Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguéz P, Doerks T, Stark M, Müller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 39: D561–D568
- Warren DH, Pereira LM, Pereira F (1977) Prolog - the language and its implementation compared with Lisp. *Proceedings of the 1977 symposium on artificial intelligence and programming languages*
- Wilkinson MD, Links M (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform* 3:331–341
- Zhou Y, Wang R, Li L, Xia XF, Sun Z (2006) Inferring functional linkages between proteins from evolutionary scenarios. *J Mol Biol* 359:1150–1159