

UNIVERSITE DE NANTES
FACULTE DE MEDECINE

Génomique intégrative du coactivateur transcriptionnel PRC

THESE DE DOCTORAT

Ecole doctorale: Biologie Santé Nantes Angers

Discipline: sciences de la vie et de la santé

Spécialité: recherche clinique, innovation technologique, santé publique

*présentée
et soutenue publiquement par*

CARAT Solenne

le 22 octobre 2010, devant le jury ci-dessous

Président: Dr REGNIER Mireille, LIX-Ecole polytechnique

Rapporteurs: Dr ROUGEMONT Jacques, Ecole Polytechnique
Fédérale de Lausanne
Pr THIEBLEMONT Catherine, Hôpital Saint-Louis

Examineurs: Dr POLLET Nicolas, Université d'Evry Val d'Essonne

**Directeur
de Thèse:** Dr HOULGATTE Rémi, Université de Nantes

Co-encadrant : Dr BOURDON Jérémie, Université de Nantes

Invitée : Dr SAVAGNER Frédérique, Université d'Angers

« *Ecoutez : de l'extérieur c'est déjà magnifique* »

Pierre, Le Père Noël est une ordure

Génomique Intégrative du coactivateur transcriptionnel PRC

L'avènement ces dernières années de technologies à haut débit comme les puces à ADN et plus récemment le « séquençage nouvelle génération » permet maintenant une analyse exhaustive du génome à différents niveaux de régulation parmi lesquels la mesure de l'expression de l'ensemble des gènes d'une cellule, les interactions ADN - facteurs de transcription, ou encore la méthylation de l'ADN. Il devient alors indispensable d'intégrer ces différentes données afin de pouvoir comprendre puis modéliser tous les mécanismes mis en jeu pour le fonctionnement global d'une cellule. Cette compréhension permettra alors de mieux appréhender les sources de dérégulation à l'origine de nombreuses pathologies.

L'objectif de mon travail de thèse est de comprendre l'impact du coactivateur transcriptionnel PRC (PGC-1 Related Coactivator) sur la coordination de la prolifération mitochondriale et cellulaire, à partir de cellules de la lignée tumorale humaine XTC.UC1, riche en mitochondries. Pour cela, PRC et les facteurs de transcription ERRA, NRF1, GABP, CREB et YY1 impliqués dans ces deux processus ont été étudiés par ChIP-chip. Les gènes positifs pour ces différents facteurs sont alors comparés au transcriptome des mêmes cellules sous l'effet d'un siRNA PRC.

L'intégration de ces différentes données de transcriptome et de ChIP-chip, couplées à des études de découverte de motifs, permettra ainsi de construire des réseaux de régulation transcriptionnelle nécessaires à la compréhension des voies d'action de PRC. Ces réseaux rendront possible la détection de cibles thérapeutiques permettant de corriger un état pathologique.

Mots clés : *génomique intégrative, ChIP-chip, transcriptome, découverte de motifs, PRC*

Integrative genomics of transcriptional coactivator PRC

Emergence over the last few years of high throughput technologies like DNA microarray and more recently next generation sequencing now allows an exhaustive analysis of genome with several regulation levels including expression measure of all cell genes, DNA – transcription factors interactions, and also DNA methylation. It thus becomes indispensable to integrate these different data to be able to understand and then to model all these mechanisms involved in global cell function. This understanding will then enable to apprehend the sources of deregulation involved in many pathologies.

The goal of my PhD is to understand the impact of PRC (PGC-1 Related Coactivator) transcriptional coactivator on the coordination of mitochondrial and cellular proliferation, from human tumor cell line XTC.UC1, rich in mitochondria. To do so, PRC and transcription factors ERRA, NRF1, GABP, CREB and YY1 involved in these two processes have been studied with ChIP-chip. Positive genes for these transcription factors are then compared with transcriptome of the same cells under siRNA PRC effects.

Integration of these different data of transcriptome and ChIP-chip, coupled with study of motifs discovery, will allow to construct the transcriptional regulatory networks necessary to the understanding of PRC pathways. These networks will make possible the detection of therapeutic targets allowing to correct pathological condition.

Keywords: *integrative genomics, ChIP-chip, transcriptome, motif discovery, PRC*

Solenne Carat

IRT-UN INSERM U915

8 quai moncoussu 44007 Nantes Cedex 1

Sommaire

Objectifs généraux de la thèse	11
Introduction.....	12
I- Du gène à la génomique fonctionnelle.....	12
1) La génomique	14
2) Les outils de la génomique fonctionnelle.....	15
a. Mesure de l'expression des gènes.....	17
b. Etude des facteurs de transcription	20
c. La découverte de motifs.....	25
d. Les outils choisis pour mon projet de thèse.....	26
II- La génomique intégrative	27
III- Etude des mécanismes de régulation transcriptionnelle de PRC	28
1) La mitochondrie	28
2) La biogenèse mitochondriale	29
a. Les facteurs de transcription NRF-1 et NRF-2	30
b. Le facteur $ERR\alpha$	31
c. Les facteurs de transcription ubiquitaire.....	31
d. Facteurs de transcription tissu spécifique	32
e. Les complexes de facteurs de transcription	32
3) Les co-activateurs nucléaires	33
a. PGC-1 α et PGC-1 β	33
b. PRC.....	34
Objectif du travail de thèse	35
Méthodes.....	36
I- Le transcriptome	36
1) Normalisation et clustering des données	37
2) Annotation fonctionnelle	38
II- Le ChIP-chip	39
1) Analyse des données de ChIP-chip	40
a. Normalisation des données	41
b. Sélection des sondes positives en ChIP	41
c. Traitement des réplicats et annotation fonctionnelle	42
2) Découverte de motifs	43

a. Les méthodes de découverte de motifs	44
b. Comparaison de motifs	45
c. Validation des motifs	48
III- Intégration de données hétérogènes	48
1) ChIP-chip + ChIP-chip.....	49
2) ChIP-chip + transcriptome	50
3) Construction de réseaux de régulation transcriptionnelle	50
Résultats.....	51
I- motifsComparator	51
II- Etude du coactivateur transcriptionnel PRC.....	73
1) Etude du transcriptome.....	73
2) Etude des ChIP-chip.....	76
a. Choix du type de puces	76
b. Sélection des sondes positives en ChIP	77
c. Reproductibilité des résultats.....	78
d. Sélection des gènes positifs en ChIP	81
e. Annotation fonctionnelle des gènes positifs en ChIP	82
f. Découverte de motifs dans les régions positives en ChIP	83
g. Validation des motifs.....	91
h. Intégration des résultats de ChIP des différents facteurs de transcription.....	93
i. Localisation des sites de fixation sur le génome	98
Discussion et Perspectives.....	100
Bibliographie	104
Annexes	108
I- Annotation fonctionnelle des gènes positifs en ChIP	108

Liste des tableaux

Tableau 1: Performance des 3 plateformes de NGS (adapté de [Mardis, 2008]).....	17
Tableau 2: codage des p-valeurs pour le clustering.....	49
Tableau 3: corrélation des réplicats des 6 facteurs de transcription	80
Tableau 4: effectif des gènes positifs pour les différents facteurs.....	81
Tableau 5: moyenne géométrique des p-valeurs des 6 facteurs de transcription	82
Tableau 6: annotation fonctionnelle des gènes positifs en ChIP	83

Liste des figures

Figure 1: l'évolution du concept du gène	13
Figure 2: Etapes pour l'hybridation d'une puce à ADN.....	18
Figure 3: Etapes du RNA-Seq	19
Figure 4: mécanismes de régulation de la transcription	20
Figure 5: étapes de ChIP (Chromatin ImmunoPrecipitation) et d'amplification génique.	22
Figure 6: principe du ChIP-Seq en fonction de la plateforme utilisée	23
Figure 7: comparaison de résultats de ChIP-chip et de ChIP-seq	25
Figure 8: plateforme de génomique intégrative.....	27
Figure 9: structure de la mitochondrie et chaîne respiratoire	29
Figure 10: fonctions mitochondriales régulées pas les facteurs NRF.....	30
Figure 11: fonctions mitochondriales régulées par les coactivateurs transcriptionnels PGC-1	33
Figure 12: plan expérimental	36
Figure 13: exemple de clustering hiérarchique.....	38
Figure 14: étapes du ChIP-chip	40
Figure 15: construction du profil moyen pour le Lowess.....	41
Figure 16: sélection des sondes positives en ChIP	42
Figure 17: préparation des séquences pour la découverte de motifs	44
Figure 18: principe de comparaison de 2 motifs	46
Figure 19: exemple de regroupement de motifs	47
Figure 20: Méthode de validation d'un motif.....	48
Figure 21: regroupement de motifs similaires avec motifsComparator	51
Figure 22: identification du motif ESR1	52
Figure 23: construction d'un motif consensus	53
Figure 24: clustering hiérarchique des gènes sensibles au siRNA PRC.....	74
Figure 25: motifs des 5 facteurs de transcription étudiés en ChIP	75
Figure 26: Localisation des sites de fixation de GABP	76
Figure 27: sélection des sondes positives en ChIP des 6 facteurs	78
Figure 28: hybridations multiples dues à la coupure aléatoire de l'ADN	79
Figure 29: corrélation des réplicats du facteur de transcription CREB	79
Figure 30: corrélation des réplicats des 6 facteurs étudiés en ChIP	80

Figure 31: réseaux de régulation des 6 facteurs de transcription.....	82
Figure 32: dendrogramme des motifs découverts pour les gènes positifs en CREB.....	85
Figure 33: dendrogramme des motifs découverts pour les gènes positifs en $ERR\alpha$	86
Figure 34: dendrogramme des motifs découverts pour les gènes positifs en GABP.....	87
Figure 35: dendrogramme des motifs découverts pour les gènes positifs en NRF1	88
Figure 36: dendrogramme des motifs découverts pour les gènes positifs en PRC.....	89
Figure 37: dendrogramme des motifs découverts pour les gènes positifs en YY1	90
Figure 38: identification des motifs dans les jeux de séquences positives en ChIP	91
Figure 39: validation du motif correspondant à YY1	92
Figure 40: invalidation d'un motif répété	93
Figure 41: corrélation des meilleures sondes par gènes entre facteurs.....	94
Figure 42: corrélation des gènes positifs en PRC et $ERR\alpha$	95
Figure 43: clustering des gènes positifs en ChIP	96
Figure 44: réseau d'intersection de listes de gènes positifs en CHIP.....	97
Figure 45: distribution des distances des sondes positives en ChIP aux départs de transcription	98

Liste des abréviations

ADN: Acide **D**esoxyribo**N**ucléique
ARN: Acide **R**ibo**N**ucléique
ATP: Adénosine **T**ri**P**hosphate
BAC: **B**acterial **A**rtificial **C**hromosome
CGH: **C**omparative **G**enomic **H**ybridization
ChIP-chip: **C**hromatin **I**mmuno**P**recipitation on **chip**
ChIP-seq: **C**hromatin **I**mmuno**P**recipitation followed by **seq**uencing
CREB: **c**AMP **R**esponse **E**lement **B**inding
CRM: **C**is **R**egulatory **M**odule
ENCODE: **E**NCyclopedia **O**f **D**N**A** **E**lements
ERR α : **E**strogen **R**elated **R**eceptor **A**lpha
GO: **G**ene **O**ntology
GWAS: **G**enome-**W**ide **A**ssociation **S**tudy
HMP: **H**uman **M**icrobiome **P**roject
HSF : **H**eat **S**hock **F**actor
LINE: **L**ong **I**nterspersed **N**uclear **E**lement
lowess: **l**ocally **w**eighted scatterplot smoothing
miRNA: **m**icro **R**ibo**N**ucleotide **A**cid
mTERF: **m**itochondrial transcription **T**ermination **F**actor
NGS: **N**ext **G**eneration **S**equencing
NHR: **N**uclear **H**ormone **R**eceptor
NRF1/NRF2: **N**uclear **R**espiratory **F**actor **1/2**
ORF: **O**pen **R**eadin**G** **F**rame
PGC-1: **P**eroxisome proliferator-activated receptor **G**amma **C**oactivator **1**
PPAR: **P**eroxisome **P**roliferator-**A**ctivated **R**eceptor
PRC: **P**GC-1 **R**elated **C**oactivator
PWM: **P**osition **W**eight **M**atrice
RNA-seq: **R**ibo**N**ucleotide **A**cid sequencing
RORA: **R**etinoid related **O**rphan **R**eceptor **A**lpha
SINE: **S**hort **I**nterspersed **N**uclear **E**lement
siRNA: **s**ilencing **R**ibo**N**ucleotide **A**cid
SNP: **S**ingle **N**ucleotide **P**olymorphism

Sp1: Specificity **p**rotein **1**

TFAM: Transcription **F**actor **A**, **M**itochondrial

TFB1M: Transcription **F**actor **B1**, **M**itochondrial

TFB2M: Transcription **F**actor **B2**, **M**itochondrial

TFBS: Transcription **F**actor **B**inding **S**ite

TIM: Translocase **I**nnner **M**embrane

TOM: Translocase **O**uter **M**embrane

TSS: Transcription **S**tart **S**ite

YY1: Ying **Y**ang **1**

Objectifs généraux de la thèse

L'objectif de mon travail de thèse a été de mettre en place des stratégies de génomique intégrative pour étudier de manière globale les mécanismes de régulation transcriptionnelle d'une fonction donnée.

J'ai pour cela utilisé des outils de la génomique fonctionnelle pour étudier le niveau d'expression de tous les gènes d'une cellule (transcriptome) ainsi que les facteurs de transcription intervenant dans la régulation de ces gènes (ChIP-chip). Le but ultime de mon travail était de préciser les méthodes et les outils nécessaires à l'intégration de données hétérogènes afin de pouvoir par la suite construire des réseaux de régulation transcriptionnelle permettant de comprendre et de modéliser les mécanismes sous-jacents aux pathologies étudiées. Cette compréhension est une étape importante pour identifier puis tester *in silico* des cibles thérapeutiques potentielles qui permettent de rétablir le niveau d'expression physiologique des gènes dérégulés.

Le modèle choisi pour éprouver ces stratégies est la lignée cellulaire XTC.UC1, modèle humain de tumeurs riches en mitochondries. La question principale que j'ai abordée est de comprendre le rôle du coactivateur transcriptionnel PRC dans la coordination de l'expression des gènes mitochondriaux et nucléaires, ce qui est permis par ce modèle.

Introduction

I- Du gène à la génomique fonctionnelle

Le gène est le support de l'information héritable. Cette notion de transmission des caractères à la descendance, proposée par Gregor Mendel dès 1866, a beaucoup évolué depuis. Le terme « gène » a été pour la première fois proposé par Wilhelm Johannsen en 1909, et Thomas Hunt Morgan a démontré l'année suivante qu'ils étaient structurés en chromosomes. Ces nouvelles connaissances ont donné lieu à la construction de la première carte physique des gènes. En 1941, George Wells Beadle et Edward Lawrie Tatum sont les premiers à montrer la relation entre mutations et modification du métabolisme. C'est de ces travaux que la notion « un gène donne une protéine » est née. La découverte de la structure de l'ADN par James Dewey Watson et Francis Crick en 1953 [Watson et Crick, 1953] permet alors de comprendre comment l'information génétique est transmise de génération en génération. Cette découverte de la double hélice d'ADN, composées des 4 bases azotées complémentaires Adénine - Thymine, et Guanine - Cytosine permet de comprendre les mécanismes de copie d'un brin d'ADN, et donc les mécanismes d'apparition des mutations. En 1965, le code génétique proposé par Nirenberg [Nirenberg et al., 1965], à partir de la découverte de la structuration en codons de l'ADN par Gamow, permet de comprendre les mécanismes de transcription et de traduction des gènes en protéines. Au cours de la décennie suivante, Walter Fiers séquence pour la première fois un gène, celui du bactériophage MS2 codant pour la protéine de capsid [Fiers et al., 1971], puis le génome complet de ce même organisme.

Les nouvelles connaissances sur les séquences des gènes permettent à Phillip Sharp et Richard J. Roberts d'identifier des gènes capables de produire plusieurs protéines, c'est la découverte de l'épissage alternatif [Gerstein et al., 2007]. Dès lors, le concept « un gène, une protéine » est aboli. Les progrès considérables de la biologie moléculaire, et le début de la détection *in silico* de gènes à partir des ORF (Open Reading Frame) rendent alors possibles des études plus complètes du génome, c'est le début de la génomique. Toute cette évolution est retracée dans la Figure 1.

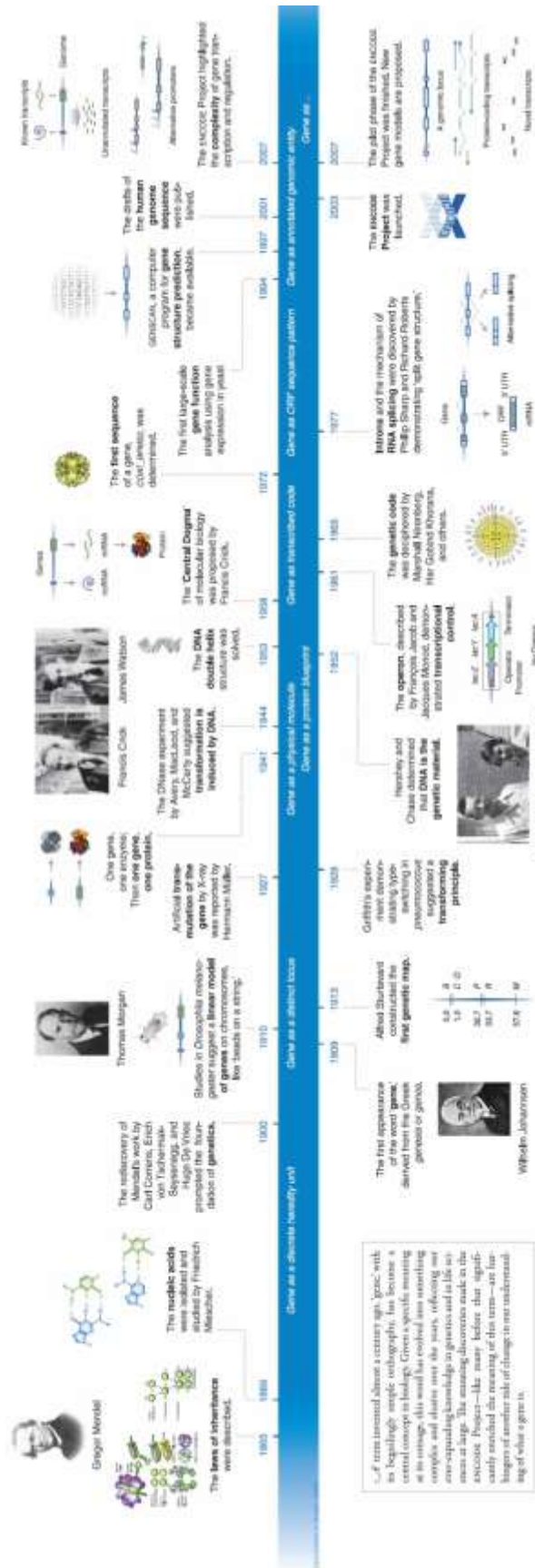


Figure 1: l'évolution du concept du gène

Historique des découvertes ayant fait évoluer les concepts du gène, depuis 1865. D'après **[Gerstein et al., 2007]**

1) La génomique

La génomique est l'étude exhaustive des génomes et en particulier de l'ensemble des gènes, de leurs dispositions sur les chromosomes, de leurs séquences, de leurs fonctions et de leurs rôles. Elle se décompose en deux orientations complémentaires: la génomique structurale et la génomique fonctionnelle. Les débuts de la génomique remontent au séquençage complet du génome du Phage Φ -X174 par Frederick Sanger en 1977 [Sanger *et al.*, 1977]. De nombreux programmes internationaux d'étude systématique des génomes sont alors lancés dans les années 80, et les premières cartes génétiques, géniques et physiques des génomes sont construites.

Le premier axe de la génomique qui a été développé est celui de la génomique structurale. Elle comprend principalement le séquençage des génomes, la détermination de polymorphismes et la détection d'altérations géniques. Cette branche de la génomique permet, entre autres, de reclasser l'ensemble des espèces en fonction de l'évolution de leurs génomes, d'identifier précisément la position des gènes sur les génomes, et d'étudier les polymorphismes liés aux pathologies humaines. Plusieurs projets internationaux d'envergure ont débuté récemment dans ce domaine. On peut tout d'abord parler du projet génome humain, lancé en 1990, visant au séquençage complet de ce génome, et à la caractérisation de l'ensemble de ses gènes. Ce projet s'est achevé en 2003, avec plus de 2 ans d'avance sur les premières prévisions grâce aux avancées technologiques majeures du séquençage. Les technologies de séquençage de « nouvelle génération » (NGS) ont notamment permis à un nouveau grand projet de voir le jour : le projet 1000 génomes. Ce projet ambitieux, dont les études pilotes ont débuté en 1998, a pour objectif de séquencer le génome complet de plus de 1000 individus, afin d'étudier les polymorphismes de la population humaine. Le projet consiste à séquencer les génomes d'hommes et de femmes issus de 20 populations différentes à travers le monde, afin de déceler des polymorphismes ayant des fréquences d'apparition d'au moins 1% dans la population générale. Les données issues de ce projet seront accessibles gratuitement pour l'ensemble de la communauté scientifique afin de servir notamment pour les analyses d'association génome entier (GWAS). Les données issues du séquençage, qui s'accroissent considérablement de jour en jour, ont notamment permis de mettre en évidence des gènes impliqués dans certaines pathologies, mais une analyse fonctionnelle de ces gènes cibles est maintenant indispensable pour comprendre les mécanismes mis en jeu.

La génomique fonctionnelle est arrivée dans un second temps, pour identifier la fonction des gènes, ainsi que pour étudier la variation de leur expression. La génomique fonctionnelle a pour ambition de comprendre tous les mécanismes dynamiques de régulation transcriptionnelle à l'échelle d'un tissu ou d'un organe pour expliquer le phénotype observé. Au-delà de l'aspect fondamental de ces études, elle permettra également de comprendre les dérégulations sous-jacentes aux pathologies étudiées, et d'identifier de nouvelles cibles thérapeutiques. Plusieurs projets internationaux ont également récemment vu le jour dans ce domaine. Le projet ENCODE (ENCyclopedia Of DNA Elements [ENCODE Project Consortium, 2004]), débuté en 2003, s'est fixé pour objectif d'identifier, de classer et de caractériser tous les éléments fonctionnels du génome humain. Le projet pilote, concernant l'analyse de seulement 1% du génome, s'est achevé en 2007 [Birney *et al.*, 2007; Weinstock, 2007]. Une des principales découvertes de ce projet pilote concerne la proportion du génome transcrit en ARN primaire (93%), alors qu'un nombre infiniment plus petit est ensuite traduit en protéines. Cette découverte montre l'étendue du rôle des ARNs et l'intérêt de les étudier en détail. Cette étude pilote a permis de tester et de développer les méthodes à haut débit nécessaires à ces analyses fonctionnelles globales et indispensables à l'étude des 99% restant du génome. Un autre projet international, le Projet Microbiome Humain, a pour objectif d'étudier les échanges inter-espèces microbiennes ainsi que les interactions microbes-homme, dans les états physiologiques. Le projet HMP (Human Microbiome Project) cherche à étudier l'ensemble des très nombreuses bactéries dont l'homme est l'hôte, et d'essayer de comprendre leur impact sur le développement, la physiologie, le système immunitaire et la nutrition humaine. Cette étude de métagénomique est rendue possible grâce au séquençage de plus de 600 génomes bactériens.

Toutes ces études de génomique, qu'elles soient structurales ou fonctionnelles, ont pour point commun de s'intéresser à l'intégralité du génome, et nécessitent donc la mise au point de technologies permettant d'analyser, en une seule fois et si possible dans les conditions physiologiques, le génome global.

2) Les outils de la génomique fonctionnelle

Le génome humain est composé de plus de 3 milliards de paires de bases, et d'environ 30 000 gènes, dont plus de la moitié sont de fonction inconnue. En faire une étude complète, que ce soit en terme de séquences ou de gènes, nécessite donc des

méthodes à haut débit, réalisables rapidement et à moindre coût. Les deux principales méthodes hauts débits utilisables en génomique sont les puces à ADN et le séquençage.

Les puces à ADN ont été utilisées pour la première fois en 1995, notamment par l'équipe de Patrick O'. Brown [Schena *et al.*, 1995] pour l'analyse de l'expression des gènes d'*Arabidopsis Thaliana*. Les puces à ADN se composent d'un support solide, généralement en verre ou en silicium, sur lequel est positionné un réseau de sondes ordonnées. Les 2 principales techniques de fixation des sondes sur le support sont le dépôt et la synthèse *in situ*. La technique de synthèse *in situ* présente l'avantage d'une reproductibilité parfaite lors de la production de plusieurs lames, ce qui n'était pas le cas avec la méthode par dépôt. Elle permet également d'augmenter davantage la résolution des puces à ADN. Le nombre de sondes, variable en fonction du fournisseur, peut atteindre plusieurs millions sur une seule lame. L'ARN ou l'ADN de l'échantillon que l'on veut étudier va être amplifié et marqué, généralement par un fluorochrome, puis hybridé sur la puce à ADN. Le principe de fonctionnement des puces à ADN repose sur la découverte de complémentarité des bases par Watson et Crick [Watson *et Crick*, 1953]. En effet, chaque ADNc présent dans l'échantillon étudié va s'hybrider sur la sonde complémentaire à sa séquence sur la puce à ADN. Suite à l'hybridation et au lavage de la lame, une lecture laser permet de quantifier l'émission de fluorescence des sondes. Les puces à ADN, initialement utilisées principalement pour mesurer l'expression des gènes (transcriptome), ont maintenant de nombreuses autres applications. On les utilise notamment pour des études de CGH, de SNP, de miRNA, de ChIP-chip ou encore de méthylation de l'ADN.

Le séquençage est une technique plus ancienne, mais dont les progrès récents ont rendu ses champs d'applications beaucoup plus importants. Les premières techniques de séquençage du génome humain nécessitaient l'utilisation de milliers de BAC (Bacterial Artificial Chromosome), chacun contenant un fragment d'ADN d'environ 100 kb. Ces BAC devaient ensuite être amplifiés par culture bactérienne, isolés puis coupés en plusieurs petits fragments de 2 à 3 kb qui sont alors clonés dans des plasmides. La séquence de chaque BAC est couverte en moyenne 8 fois par les plasmides afin de pouvoir par la suite reconstruire la séquence complète du génome. Depuis le projet génome humain, il est possible de fragmenter directement l'ADN génomique et de le cloner dans des plasmides, sans passer par l'étape des BAC. La technique de séquençage utilisée dans ces 2 cas de figure reste celle de Sanger [Sanger *et al.*, 1977], avec un système de capillarité. Les techniques de séquençage de nouvelles générations (NGS),

disponibles depuis 2004, révolutionnent actuellement le séquençage, non seulement en terme de temps, de fiabilité et de coûts, mais aussi en terme d'applications.

Les trois plateformes basées sur cette technologie sont le 454 de Roche [M. Margulies *et al.*, 2005] utilisant la technologie du pyroséquençage, le Solexa d'Illumina avec le séquençage par synthèse et le SOLiD d'Applied Biosystem utilisant une technique de séquençage par ligature. Chacune de ces 3 plateformes a quelques autres spécificités propres décrites dans le Tableau 1 ci-dessous.

Tableau 1: Performance des 3 plateformes de NGS (adapté de [Mardis, 2008])

	454	Solexa	SOLiD
technique de séquençage	pyroséquençage	séquençage par synthèse	séquençage par ligation
méthode d'amplification	PCR en emulsion	amplification par pont	PCR en emulsion
paired ends / séparation	oui / 3 kb	oui / 200 pb	oui / 3 kb
MB / run	100 Mb	1300 Mb	3000 Mb
temps / run	7h	4 jours	5 jours
longueur des lectures	250 pb	32-40 pb	35 pb
coût par run	6423 €	6812 €	13280 €
coût par Mb	64 €	4,5 €	4,4 €

Deux autres plateformes sont en cours de développement, le SMRT de Pacific Biosciences et l'Heliscope d'Helicos Bioscience. Toutes ces techniques ont la particularité de produire des fragments séquencés assez courts par rapport à l'utilisation de la technique classique de Sanger. Ceci a donc nécessité le développement d'outils de bioinformatique pour l'assemblage des séquences et la détection d'erreurs [J. R. Miller *et al.*, 2010]. De part sa rapidité, et son coût de plus en plus faible, le séquençage « nouvelle génération » est maintenant couramment utilisé pour le séquençage et le reséquençage des génomes, mais également pour étudier les interactions ADN-protéine avec la technologie du ChIP-seq, l'expression des gènes avec le RNA-Seq, ou encore la découverte de petits ARNs.

a. Mesure de l'expression des gènes

La mesure de l'expression des gènes se fait par la mesure du nombre de copies d'ARNm transcrit pour chaque gène, appelé transcriptome. Dans une situation donnée, plus le gène est transcrit, plus la quantité d'ARNm sera importante. La mesure

d'expression se fait généralement en comparant 2 situations particulières. Il est possible par exemple de comparer des individus sains et des individus atteints d'une pathologie particulière, des souris KO pour un gène donné contre des souris sauvages, ou encore des cellules d'un tissu contre des cellules d'un autre tissu.

Les puces à ADN sont historiquement l'outil de référence pour l'étude du transcriptome. Une fois les ARNm extraits des cellules, ils sont amplifiés et marqués par des fluorochromes. Les échantillons sont alors hybridés sur une puce à ADN. Ces étapes sont reprises dans la Figure 2.

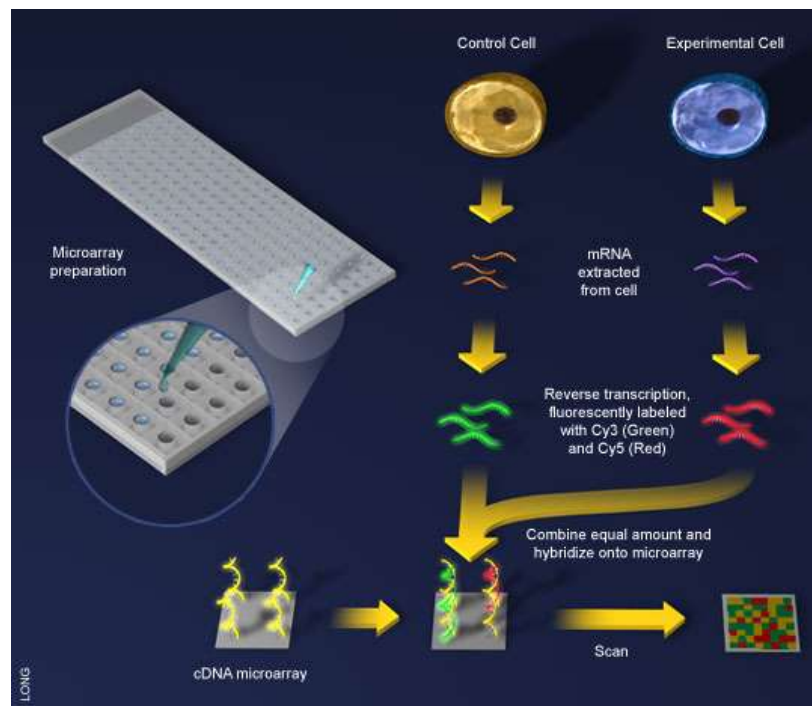


Figure 2: Etapes pour l'hybridation d'une puce à ADN

Les ARN messagers des deux échantillons sont extraits, marqués par reverse transcription (ADNc) par des fluorochromes (cyanines) puis hybridés sur puces à ADN.

La lame est alors scannée et les données d'expressions de chacune des sondes sont enregistrées. Des méthodes de normalisation et de traitement des données permettent alors d'identifier les gènes dont l'expression est différentielle entre les 2 situations étudiées.

Le RNA-Seq, une autre technique plus récente, est une des applications issue des séquenceurs nouvelle génération. Les principales étapes de cette technique sont présentées dans la Figure 3. Les ARN totaux, ou des ARN fractionnés, sont convertis en une librairie d'ADNc avec des adaptateurs aux extrémités. Chaque fragment est alors séquencé, après une éventuelle étape d'amplification, à partir d'une ou des deux extrémités (paired ends) des fragments. Les lectures, de 30 à 300pb en fonction de la

plateforme utilisée, sont alors alignées sur le génome de référence et divisées en 3 classes : celles correspondant à un exon, celles correspondant à une jonction exon-exon, et celles correspondant à une queue polyA. Ces 3 types de lectures servent à produire le profil d'expression du gène.

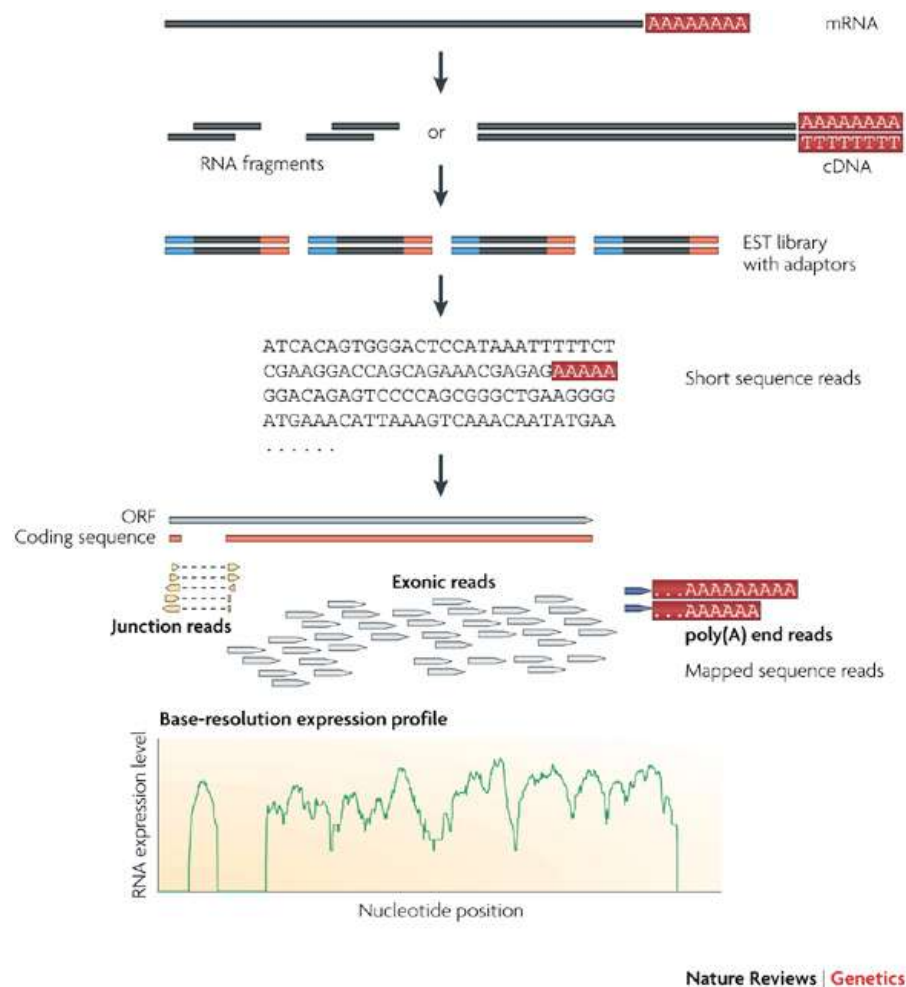


Figure 3: Etapes du RNA-Seq

Les ARN totaux sont convertis en librairie d'ADN complémentaire avec des adaptateurs qui sont ensuite séquencés puis alignés sur le génome de référence. Cet alignement permet de construire le profil d'expression ayant une résolution d'une base [Z. Wang et al., 2009].

Les puces à ADN présentent en premier lieu l'avantage d'être une technique relativement ancienne (15 ans) et donc bien maîtrisée autant sur le plan expérimental que bioinformatique, avec des méthodes d'analyses maintenant standardisées. Elles sont également 10 à 100 fois moins chères, à résolution comparable, qu'une même analyse réalisée par RNA-Seq, notamment à cause des transcrits très abondants (5% des gènes représentent 75% des lectures), problème que l'on ne rencontre pas avec les puces à ADN. Cependant, même si les puces à ADN permettent une mesure quantitative du taux

d'expression de chaque gène, la technique du RNA-Seq permet une mesure plus précise du nombre de copies pour les gènes d'expression intermédiaire. Enfin, le RNA-Seq présente l'avantage de pouvoir découvrir de nouveaux ARNm, ce qui n'est pas le cas des puces à ADN puisqu'elles ne contiennent que les sondes correspondant aux transcrits connus et séquencés.

b. Etude des facteurs de transcription

Les facteurs de transcription sont des protéines qui agissent sur le niveau d'expression des gènes, en régulant la transcription du gène en ARNm. Ils se lient à l'ADN grâce à leur domaine de liaison à l'ADN, ou en complexe avec d'autres facteurs de transcription (co-facteur), et peuvent activer ou inhiber l'expression des gènes. La Figure 4 présente les différents mécanismes de régulation de la transcription par les facteurs de transcription.

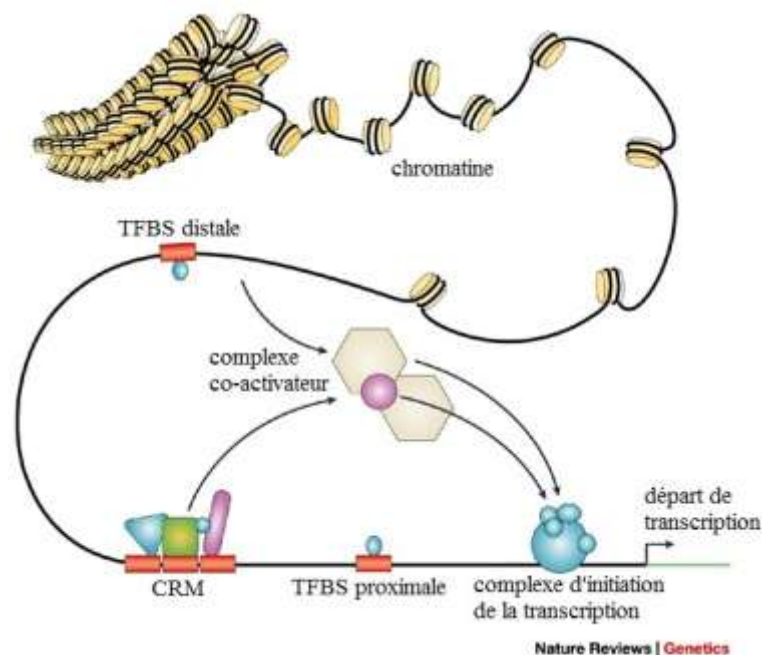


Figure 4: mécanismes de régulation de la transcription

Représentation des principales zones de régulation de la transcription : le complexe d'initiation de la transcription, le TFBS (Transcription Factor Binding Site) proximal, les CRM (Cis Regulatory Module), et le TFBS distal. Image adaptée de [Wasserman et Sandelin, 2004]

Certains facteurs de transcription, appelés facteurs de transcription généraux, sont indispensables pour l'initiation de la transcription. Il s'agit notamment du facteur RNAPol II, et des facteurs de la famille TFII qui forment le complexe d'initiation de la transcription, localisé près du départ de transcription [M. C. Thomas et Chiang, 2006].

D'autres facteurs de transcription, plus spécifiques, permettent de réguler finement la transcription des gènes, afin d'exprimer les bons gènes au bon moment avec le niveau nécessaire à chaque cellule, en fonction du tissu et de l'état physiologique. Ils peuvent se fixer sur le promoteur proximal, ou sur des régions plus distales du départ de transcription, sur leur site de fixation spécifique (TFBS). Des modules cis-régulateurs (CRM) de facteurs de transcription, des co-facteurs, et des facteurs de transcription distaux forment alors un complexe capable de stabiliser la machinerie d'initiation de la transcription afin de permettre la transcription proprement dite. La structure tridimensionnelle de la chromatine joue un rôle important pour favoriser ou non la formation des complexes.

Un certain nombre de sites de fixation de facteurs de transcription sont répertoriés dans les bases de données, telles que JASPAR [Sandelin *et al.*, 2004] et Transfac [Wingender *et al.*, 1996]. Chaque site de fixation y est représenté par une matrice poids position (PWM), montrant l'importance relative de chaque base à chaque position pour la fixation du facteur de transcription. Ces sites sont représentés par des logos [Schneider *et Stephens*, 1990] caractérisant les motifs d'ADN reconnus par les facteurs de transcription. Il existe quatre familles distinctes de facteurs de transcription, qui reconnaissent des séquences spécifiques. Ces quatre familles se distinguent par la conformation tridimensionnelle de leurs domaines de liaison à l'ADN : Hélice – Tour – Hélice, protéines à doigts de zinc, facteurs à glissière de leucine et Hélice – Boucle – Hélice. De plus, certains sites de fixation spécifiques peuvent être reconnus, comme les motifs palindromiques ou encore les motifs en dimères. En fonction de l'orientation et de l'ordre des dimères de facteurs de transcription, le niveau de régulation de la transcription peut varier.

L'étude des facteurs de transcription va permettre d'identifier les gènes cibles d'un facteur de transcription donné, et de localiser précisément son site de fixation. Le pré-requis indispensable pour pouvoir effectuer cette étude est l'existence d'un anticorps spécifique du facteur de transcription que l'on souhaite étudier. Il est alors possible d'identifier, dans les conditions physiologiques, tous les gènes cibles du facteur de transcription choisi. Deux techniques principales permettent d'étudier les facteurs de transcription à grande échelle, le ChIP-chip et le ChIP-seq. La première étape, commune aux deux techniques, consiste en une immunoprécipitation de la chromatine, avec un anticorps spécifique du facteur de transcription d'intérêt. La deuxième étape consiste en l'hybridation de toutes les séquences d'ADN immunoprécipitées sur une puce à ADN

pour le ChIP-chip, ou au séquençage de toutes ces séquences pour le ChIP-seq. L'étape d'immunoprécipitation de la chromatine (ChIP) est décrite dans la Figure 5.

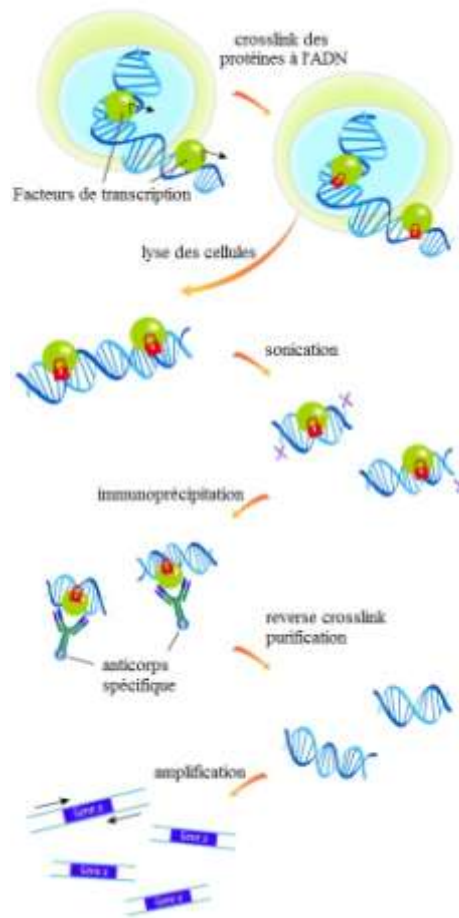


Figure 5: étapes de ChIP (Chromatin ImmunoPrecipitation) et d'amplification génique

Suite à la liaison covalente des protéines à l'ADN, les cellules sont lysées, l'ADN est fragmenté par sonication puis une immunoprécipitation spécifique permet de sélectionner spécifiquement les fragments d'ADN contenant les sites de fixation des facteurs de transcription. Le reverse crosslink et la purification permettent d'obtenir des fragments d'ADN prêts pour l'amplification. Image adaptée de : www.RnDSYSTEMS.com

Les facteurs de transcription sont fixés de manière covalente à l'ADN dans la cellule avec du formaldéhyde. Les cellules sont alors lysées et l'ADN est extrait. L'ADN est alors fragmenté par sonication ou par digestion enzymatique afin d'obtenir des petits segments d'ADN, d'environ 500 pb, qui vont ensuite être immunoprécipités avec l'anticorps spécifique. Les fragments d'ADN ainsi sélectionnés seront appelés IP. Une étape de reverse crosslink et de purification permet de dissocier les complexes ADN - facteur de transcription - anticorps, pour ne garder que l'ADN nu. Ces fragments d'ADN seront finalement amplifiés.

Pour la technique du ChIP-chip, l'échantillon contrôle pour l'hybridation subit toutes les étapes décrites précédemment, hormis l'immunoprécipitation. Il correspond

donc à des fragments de la totalité de l'ADN génomique. Ces 2 échantillons sont amplifiés, marqués chacun par un fluorochrome différent, puis hybridés sur une puce à ADN. Différents types de puces sont disponibles, avec des sondes conçues soit dans les îlots CpG, soit dans les promoteurs proximaux des gènes, soit dans le génome complet. Ils permettent une étude plus ou moins complète du génome. Les puces à îlots CpG, tout comme les puces à promoteurs, permettent d'étudier les facteurs de transcription qui se fixent dans les régions majeures de la régulation de l'expression de gènes. Cependant, certains facteurs de transcription peuvent se fixer très en amont du départ de transcription, jusqu'à 50kb (en moyenne) pour le facteur de transcription ESR1 par exemple [Carroll *et al.*, 2006]. En fonction du facteur de transcription que l'on étudie, il peut donc être intéressant d'utiliser des puces à ADN couvrant la totalité du génome, afin de ne pas omettre de sites de fixation. Cependant, ces puces ont un coût beaucoup plus important que les puces à promoteur.

Le ChIP-Seq, dont les premiers résultats ont été publiés en 2007 [Robertson *et al.*, 2007; D. S. Johnson *et al.*, 2007], nécessite, suite à l'immunoprécipitation de la chromatine, une étape de séquençage massif. La technique de séquençage massif diffère en fonction de la plateforme utilisée, comme illustré dans la Figure 6 [P. J. Park, 2009], même si le concept reste identique.

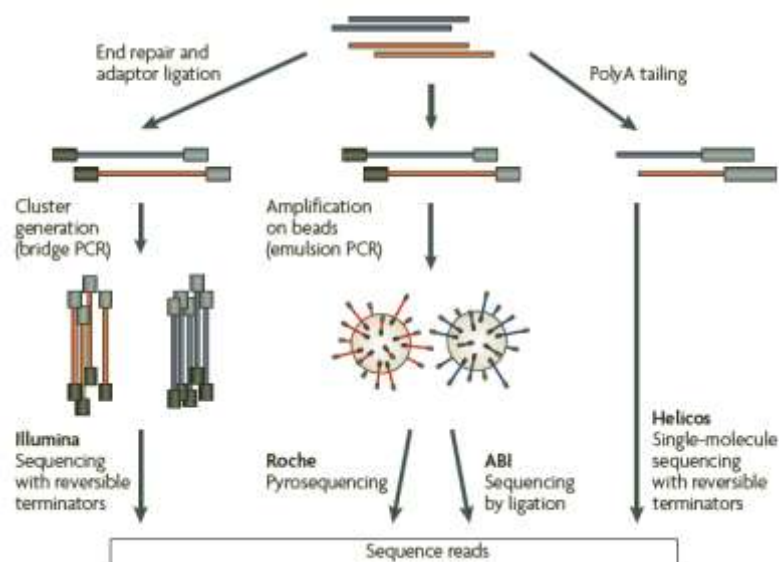


Figure 6: principe du ChIP-Seq en fonction de la plateforme utilisée

L'ADN purifié peut être séquençé par les différentes plateformes de NGS (Next Generation Sequencing). Le concept reste cependant le même pour les différentes plateformes : un adaptateur est fixé aux extrémités des fragments d'ADN qui sont ensuite clonés. L'étape de séquençage s'effectue sur tous les échantillons en parallèle. Image adaptée de [P. J. Park, 2009].

Ces 2 techniques d'analyse des sites de fixation des facteurs de transcription à grande échelle présentent chacune des avantages et des inconvénients. Le ChIP-chip présente dans un premier temps l'intérêt d'utiliser une technique éprouvée depuis de nombreuses années, la technique des puces à ADN, dont les biais sont maintenant connus, et pour la plupart maîtrisés, grâce notamment à des méthodes de normalisation. Les principaux biais étant des problèmes d'hybridation non spécifique, et une gamme dynamique non linéaire. Un autre problème issu de l'utilisation des puces à ADN est l'impossibilité d'étudier une région chromosomique où aucune sonde n'a été conçue, contrairement au séquençage massif. Cependant, la technique des puces à ADN est maintenant bien répandue, le matériel nécessaire à l'hybridation des puces présent dans de nombreux laboratoires, et le coût est globalement moins élevé que pour réaliser le ChIP-Seq. De plus, les dernières méthodes de synthèse des sondes sur les puces à ADN permettent d'augmenter considérablement la résolution de celles-ci, et de pouvoir ainsi couvrir l'ensemble du génome sur un nombre de lames très faible. Le ChIP-Seq, qui va continuer à progresser avec l'essor du séquençage massif, reste pour le moment une technique coûteuse, et plus difficilement accessible. De plus, les méthodes d'analyse restent encore à améliorer, notamment pour l'alignement des séquences localisées dans les régions riches en GC. De plus, la réalisation d'un contrôle peut être extrêmement coûteuse en ChIP-seq, à cause du nombre de séquences nécessaires du fait du nombre important des séquences répétées du génome ; alors que la technologie bi-couleur du ChIP-chip permet d'hybrider l'échantillon contrôle sans surcoût.

Cependant, plusieurs études actuelles [Robertson *et al.*, 2007; Gilchrist *et al.*, 2009] s'accordent pour dire que les résultats les plus significatifs des 2 techniques concordent, comme illustré dans la Figure 7. Le ChIP-chip et le ChIP-seq de quatre facteurs de transcription de la drosophile permettent d'identifier les mêmes régions de régulation.

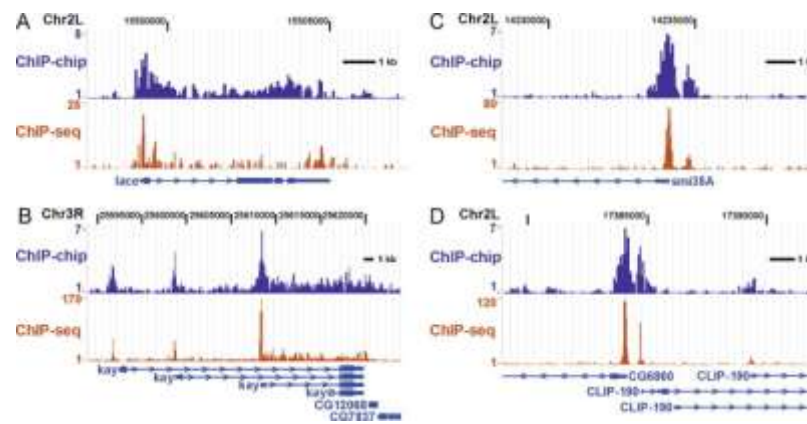


Figure 7: comparaison de résultats de ChIP-chip et de ChIP-seq

Les résultats de ChIP-chip (en bleu) et de ChIP-seq (en rouge) correspondant à quatre facteurs de transcription distincts de la drosophile ne montrent pas de différence significative pour la localisation des sites de fixation sur l'ADN [Gilchrist *et al.*, 2009].

c. La découverte de motifs

Les méthodes de découverte de motifs sont utilisées pour identifier les sites de fixation des facteurs de transcription sur l'ADN. Ces méthodes sont utilisées aussi bien sur les résultats de ChIP-chip que de ChIP-seq, afin de localiser précisément le site de fixation. L'objectif des méthodes de découverte de motifs est d'identifier un motif, une courte séquence d'ADN, statistiquement surreprésentée dans un jeu de séquences par rapport à ce que l'on attendrait par hasard. Il sera alors légitime de penser que ce motif est un bon candidat pour représenter un site de fixation de facteur de transcription. Cette recherche est effectuée, dans le cas d'une étude de ChIP, sur les séquences positives. De nombreux outils basés sur plusieurs méthodes, tels que les algorithmes EM (Expectation Maximisation), échantillonnage de Gibbs ou le comptage de mots, ont été développés ces dernières années pour identifier ces sites de fixation. L'algorithme EM est basé sur une procédure d'optimisation locale pour maximiser une fonction de voisinage avec des variables cachées. Comme toutes les méthodes d'optimisation locale, cet algorithme est fortement dépendant du point d'initialisation, et les résultats sont donc différents à chaque nouvelle itération. Le principal outil de découverte de motif utilisant cet algorithme EM est le logiciel MEME [Bailey *et Elkan*, 1994]. La méthode d'échantillonnage de Gibbs est une technique classique pour réaliser de l'inférence de probabilité. Contrairement aux méthodes EM, ces méthodes sont globales et relatives à une distribution de paramètres. Cependant, la réévaluation de ces paramètres sur des jeux aléatoires est très gourmande en calculs en raison du nombre important d'itérations nécessaires. L'outil AlignACE [Hughes *et al.*, 2000] est un exemple de logiciel basé sur cette méthode pour la découverte de motifs. Enfin, plusieurs méthodes de comptage de mots intégrant des

données complémentaires nécessaires à la détection de motifs pertinents, tels que la conservation phylogénique ou encore la validation par ChIP-chip, sont disponibles. On peut notamment citer MDscan [X. S. Liu *et al.*, 2002] et MDmodule, issus de la suite MotifRegressor [Conlon *et al.*, 2003], qui utilisent des résultats de ChIP-chip pour optimiser et raffiner la découverte de motifs.

Parallèlement à ces outils, 2 bases de données principales de motifs ont été développées afin de recenser et d'identifier les sites de fixation des facteurs de transcription connus. Ces bases de données, JASPAR [Sandelin *et al.*, 2004] et Transfac [Wingender *et al.*, 1996], contiennent plusieurs centaines de sites de fixation, validées de manière expérimentale ou bioinformatique sur un jeu de séquences plus ou moins important. La base de données JASPAR, dont la principale particularité est d'être entièrement publique, a récemment été filtrée et validée afin d'éliminer les nombreuses redondances qu'elle contenait dans les anciennes versions. La base de données Transfac possède quant à elle une seule partie accessible gratuitement, mais basée uniquement sur les résultats datant d'avant 2005. Cependant, la suite BIOBASE, de laquelle est issue Transfac, propose de nombreux autres outils dévolus à l'analyse des facteurs de transcription et de la régulation de l'expression des gènes.

d. Les outils choisis pour mon projet de thèse

Pour mon projet de thèse, j'ai choisi d'utiliser les puces à ADN et le ChIP-chip pour effectuer une étude de génomique fonctionnelle.

Ce choix s'explique tout d'abord par la présence, au sein de mon équipe, d'une plateforme de puces à ADN. Tout le matériel nécessaire à l'hybridation des puces à ADN était donc déjà disponible sur place, et l'équipe possède une grande expérience dans le domaine de la transcriptomique notamment, aussi bien pour la partie expérimentale que pour la partie bioinformatique. Le ChIP-chip, qui était en cours de mise au point sur la plateforme au début de ma thèse, est maintenant une technique bien maîtrisée sur le plan expérimental et bioinformatique.

De plus, la technique des puces à ADN étant plus ancienne que le séquençage nouvelle génération, nous avons pu bénéficier d'une grande expérience dans le traitement du signal des puces à ADN.

Enfin, comme plusieurs études ont pu le montrer, les résultats les plus probants en terme de fixation des facteurs de transcription sont comparables entre les 2 technologies, il ne nous semblait donc pas indispensable d'introduire une nouvelle technologie sur la

plateforme, qui aurait engendré des coûts très importants et de nombreuses mises au point avant d'être opérationnelle.

II- La génomique intégrative

La génomique fonctionnelle a pour objectif d'étudier le fonctionnement global des gènes à l'échelle de la cellule, ou du tissu. Les méthodes couramment utilisées permettent d'étudier différents éléments intervenant dans la régulation transcriptionnelle. Il est par exemple possible d'étudier le niveau d'expression de tous les gènes d'un tissu, ou d'identifier l'ensemble des cibles d'un facteur de transcription donné, ou encore d'étudier le niveau de méthylation de l'ADN. Pour comprendre l'ensemble des mécanismes régissant la régulation de l'expression des gènes d'un tissu, il est indispensable d'étudier simultanément toutes ces informations afin de pouvoir construire des réseaux de régulation transcriptionnelle. C'est un des objectifs de la génomique intégrative. Cette approche passe par l'analyse, à différents niveaux, d'un même modèle d'étude afin d'appréhender l'ensemble des mécanismes sous-jacents. Une plateforme de génomique intégrative a d'ailleurs été mise en place [Shinde et al., 2010] afin de faciliter l'accès à toutes les données génomiques publiques et de pouvoir les analyser afin de produire des connaissances nouvelles (Figure 8).

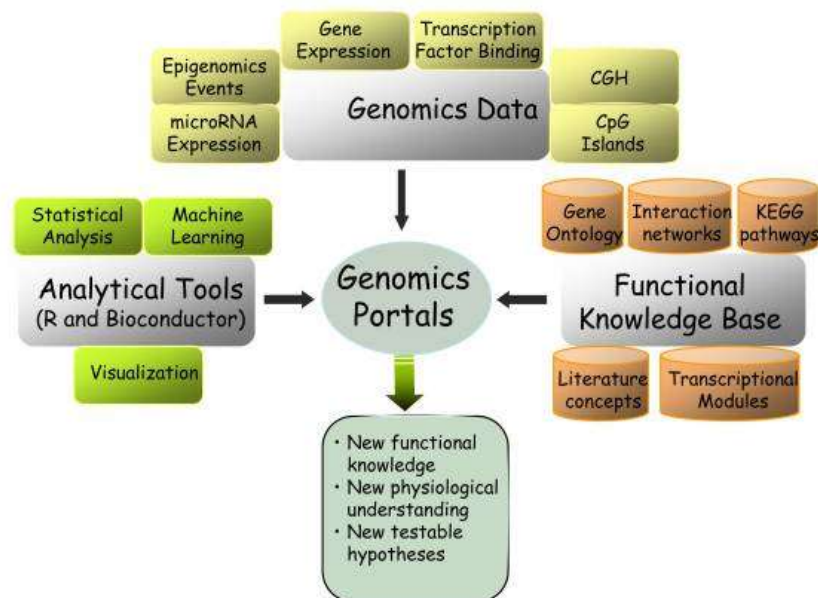


Figure 8: plateforme de génomique intégrative

Structure conceptuelle d'un portail de génomique intégrative regroupant des données de génomique, des connaissances fonctionnelles et des outils d'analyse [Shinde et al., 2010].

Pouvoir modéliser l'ensemble des mécanismes de régulation transcriptionnelle d'un tissu sain permettrait de pouvoir identifier plus facilement des dysfonctionnements dans un tissu malade, et de prévoir l'impact d'une modification sur l'ensemble du fonctionnement de la cellule.

La génomique intégrative est également un préalable indispensable à toute étude de biologie des systèmes. En effet, la biologie des systèmes a pour objectif d'étudier l'organisme vivant dans son ensemble, ainsi que ses interactions avec l'environnement. Elle aura donc besoin des données issues de la génomique, mais aussi de la protéomique et de la métabolomique, afin de comprendre les interactions entre les cellules, les organes, et l'environnement. L'étude du système biologique passera alors nécessairement par une modélisation mathématique du vivant, afin de pouvoir évaluer l'impact de différents évènements sur l'organisme complet étudié.

III- Etude des mécanismes de régulation transcriptionnelle de PRC

1) La mitochondrie

La mitochondrie est un organite localisé dans le cytoplasme des cellules eucaryotes, et dont la taille est de l'ordre du micromètre. Son rôle physiologique est primordial, puisque ce sont les mitochondries qui, par le processus de phosphorylation oxydative, produisent la principale source d'énergie de la cellule. La mitochondrie (Figure 9) est composée de 2 compartiments, la matrice et l'espace intermembranaire, séparés par une membrane interne et une membrane externe. La chaîne respiratoire, composée des complexes enzymatiques I à IV, et de l'ATP synthase constitue le système de phosphorylation oxydative. Ces 5 complexes sont enchâssés dans la membrane interne mitochondriale.

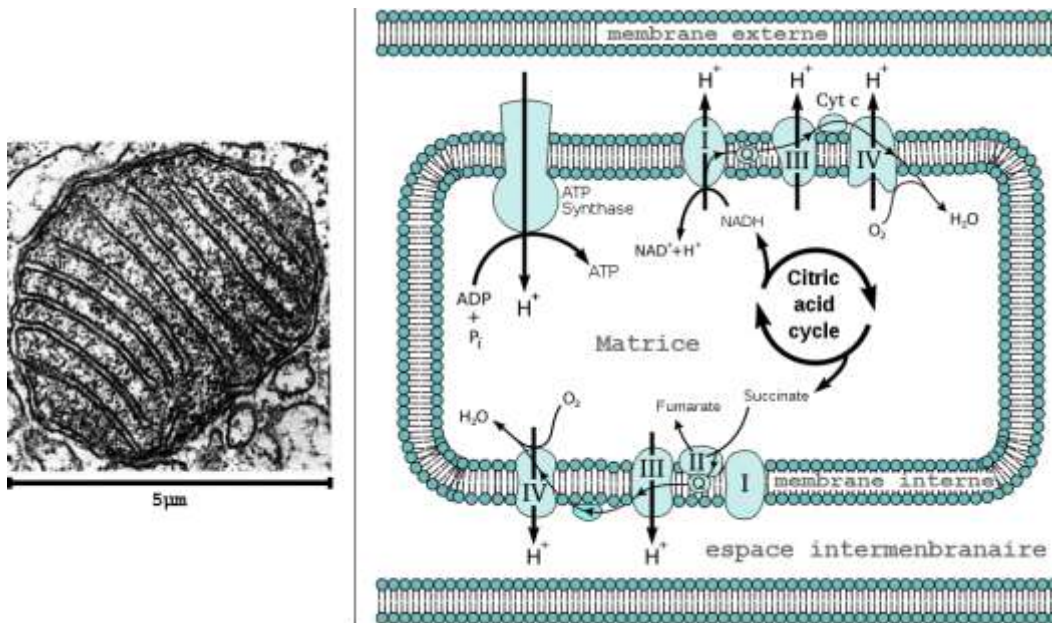


Figure 9: structure de la mitochondrie et chaîne respiratoire

Image de microscopie électronique d'une mitochondrie, et schéma représentant la structure interne de la mitochondrie avec les mécanismes de la phosphorylation oxydative.

2) La biogenèse mitochondriale

Le protéome mitochondrial comprend de 1100 à 1500 protéines. Le génome mitochondrial ne code que 13 de ces protéines correspondant à des sous-unités de certaines enzymes de la chaîne respiratoire. Toutes les autres protéines sont codées par le génome nucléaire. La biogenèse mitochondriale nécessite donc une très bonne coordination de la transcription des gènes nucléaires et mitochondriaux. La transcription de ces gènes localisés dans deux compartiments cellulaires différents dépend de facteurs de transcription nucléaire. Certains de ces facteurs de transcription sont transloqués dans la mitochondrie, permettant ainsi la transcription bidirectionnelle du génome circulaire mitochondrial. D'autres facteurs agissent au niveau du noyau sur les gènes nucléaires codant pour la majorité des sous-unités de la chaîne respiratoire, ainsi que sur les gènes impliqués dans son assemblage et son fonctionnement. La co-expression de ces facteurs est contrôlée, directement ou non, par les co-activateurs transcriptionnels de la famille PGC-1 (**P**eroxisome **p**roliferator-activated receptor **G**amma **C**oactivator-**1**) dont l'expression dépend de nombreux facteurs environnementaux.

On distingue 2 types de facteurs de transcription intervenant dans la biogenèse mitochondriale. La première catégorie concerne les facteurs de transcription transférés dans la mitochondrie pour effectuer la transcription de l'ADN mitochondrial. Il s'agit des facteurs TFAM, TFB1M, TFB2M et du facteur mTERF pour la terminaison de la

transcription. Ces protéines sont transloquées du noyau vers la mitochondrie par l'intermédiaire des complexes TOM (Translocase of the **O**uter **M**embrane) et TIM (Translocase of the **I**nnner **M**embrane).

L'autre catégorie de facteurs de transcription concerne des protéines qui gouvernent l'expression coordonnée des gènes impliqués dans la fonction de phosphorylation oxydative. Il s'agit notamment des facteurs NRF-1, GABP (aussi appelé NRF-2), $ERR\alpha$, CREB, YY1, Sp1 et PPARs.

a. Les facteurs de transcription NRF-1 et NRF-2

Les facteurs de transcription NRF-1 et NRF-2 (**N**uclear **R**espiratory **F**actors 1 et 2) agissent sur la majorité des gènes codant pour des sous-unités des complexes de la chaîne respiratoire [Gleyzer *et al.*, 2005]. De plus, ils régulent l'expression des facteurs de transcription TFAM, TFB1M et TFB2M responsables de l'augmentation de la transcription de l'ADN mitochondrial. Les différentes fonctions mitochondriales régulées par les facteurs NRF-1 et NRF-2 sont présentées dans la Figure 10.

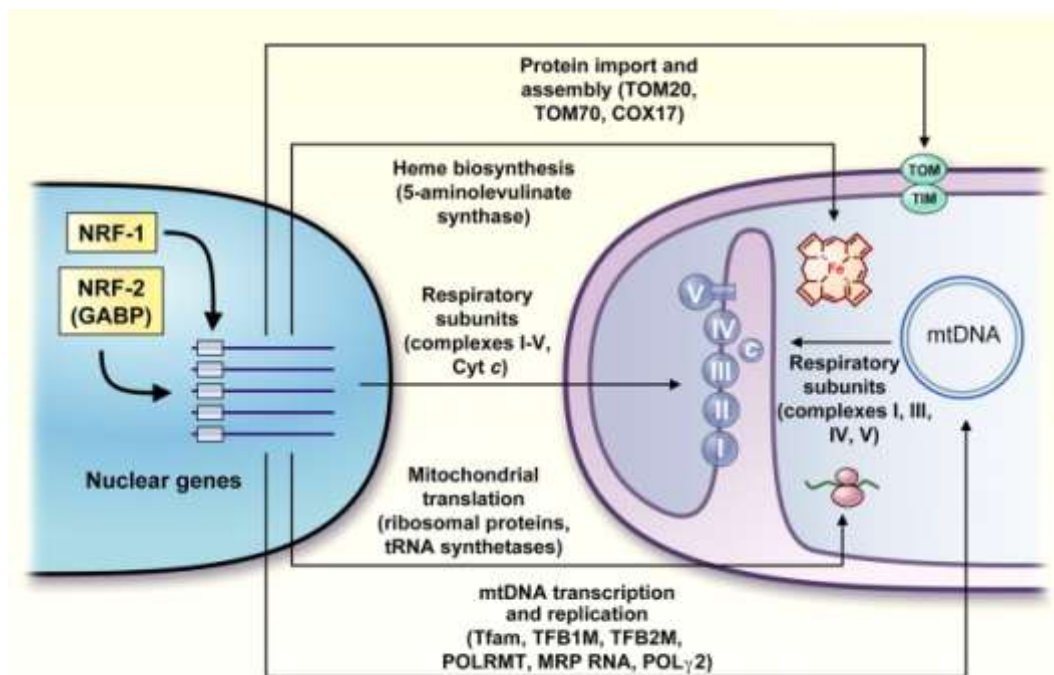


Figure 10: fonctions mitochondriales régulées pas les facteurs NRF

Les facteurs NRF (Nuclear Respiratory Factor) contribuent directement ou indirectement à l'expression de nombreux gènes nécessaires à la respiration mitochondriale. De plus, ils induisent l'expression de facteurs clés tels que TFAM, TFB1M, TFB2M, nécessaires à la transcription du génome mitochondrial. Ils jouent enfin un rôle important pour l'import des protéines nucléaires dans la mitochondrie par l'intermédiaire des complexes TOM (Translocase of the Outer Membrane) /TIM (Translocase of the Inner Membrane) [Richard C. Scarpulla, 2008].

NRF-1 a été identifié comme impliqué dans la régulation positive de nombreux gènes régissant le fonctionnement de la chaîne respiratoire [Evans et R C Scarpulla, 1990]. De plus, la phosphorylation de la partie N-terminale de NRF-1 dans les cellules en prolifération induit une augmentation de ses activités de liaison à l'ADN et de trans-activation [Gugneja et R C Scarpulla, 1997]. NRF-1 joue également un rôle important pour la coordination de l'expression des sous-unités codées par les 2 génomes en régulant l'expression des facteurs de transcription mitochondriaux (TFAM, TFB1M et TFB2M). Il est donc un facteur majeur dans l'orchestration de la biogenèse mitochondriale. Enfin, NRF-1 est également impliqué dans l'expression de nombreux gènes non impliqués dans la fonction mitochondriale, contrôlant le cycle cellulaire et la prolifération.

Des sites de liaison à l'ADN du facteur de transcription NRF-2 ont été découverts dans les promoteurs de nombreux gènes codant pour des composants de la chaîne respiratoire, pour l'import des protéines nucléaires dans la mitochondrie via les complexes TOM/TIM ainsi que dans les promoteurs des facteurs de transcription mitochondriaux [Ongwijitwat et al., 2006].

b. Le facteur ERR α

ERR α (Estrogen Related Receptor α) est un facteur de transcription de la famille des récepteurs nucléaires. C'est un des rares facteurs de cette famille qui n'a pas de ligand spécifique pour son activation. De ce fait, l'activité transcriptionnelle de ce facteur est régulée par des interactions physiques qu'il a avec des co-facteurs transcriptionnels [Villena et Kralli, 2008]. L'activité d'ERR α est donc soumise au niveau d'expression de ces co-facteurs induits soit par l'état physiologique, soit par la spécificité du tissu étudié. ERR α régule l'expression de nombreux gènes impliqués dans le mécanisme de phosphorylation oxydative, et est souvent associé aux facteurs NRF-1 et 2 [van Waveren et Moraes, 2008]. Ce facteur régule également de nombreuses fonctions mitochondriales telles que l'oxydation des acides gras, le cycle de Krebs, la production et l'export d'ATP, ainsi que la transcription et la traduction de l'ADN mitochondrial [Giguère, 2008].

c. Les facteurs de transcription ubiquitaire

Les facteurs de transcription CREB, YY1 et Sp1 sont des facteurs de transcription ubiquitaires, qui jouent notamment un rôle dans la régulation de la biogenèse mitochondriale. L'augmentation de la phosphorylation de Sp1, un facteur de transcription

surtout impliqué dans la régulation des gènes de ménage, influence la régulation de nombreux gènes mitochondriaux en cas de modification des conditions physiologiques [Zaid *et al.*, 1999]. Le rôle de Sp1 dans cette adaptation reste cependant encore à élucider. Le facteur CREB, impliqué dans la régulation de nombreuses fonctions biologiques, interagit notamment avec NRF-1 pour induire l'expression du cytochrome c. Enfin, YY1 est impliqué dans la régulation positive et négative de gènes codant des sous-unités de COX. Le site de fixation de ce facteur a également été retrouvé à proximité des facteurs de transcription NRF-1 et 2 et ERR α dans les promoteurs de gènes mitochondriaux [van Waveren *et Moraes*, 2008].

d. Facteurs de transcription tissu spécifique

Certains facteurs de transcription impliqués dans la biogenèse mitochondriale ont une activité spécifique du tissu dans lequel ils sont exprimés. On peut notamment parler des facteurs PPARs. Les récepteurs nucléaires PPARs induisent spécifiquement la biogenèse mitochondriale au niveau de la graisse blanche pour PPAR γ et du muscle squelettique pour PPAR δ . Ils sont principalement impliqués dans le métabolisme des lipides.

e. Les complexes de facteurs de transcription

Tous ces facteurs de transcription peuvent interagir en complexes, ce qui leur confère une spécificité plus importante pour la régulation de fonctions particulières (Figure 11). Par exemple, ERR α en interaction avec PPAR α , régule spécifiquement des gènes mitochondriaux impliqués dans l'oxydation des acides gras. En revanche, en complexe avec les facteurs NRFs, il intervient dans l'expression des gènes de la phosphorylation oxydative. L'expression de ces facteurs est également tissu-spécifique et dépendante de divers signaux extérieurs. De plus, il y a de nombreuses boucles de régulation et d'autorégulation entre ces différents facteurs. En effet, l'augmentation de l'expression de l'un des facteurs engendre généralement une modification de l'expression des gènes codant pour les autres facteurs [Hock *et Kralli*, 2009]. C'est notamment le cas lors d'une surexpression de ERR α , qui induit alors l'expression de PPAR α et GABP, ainsi que sa propre expression. Tous ces phénomènes complexes de régulation sont gouvernés par les co-activateurs nucléaires qui permettent une bonne adaptation aux conditions extérieures.

3) Les co-activateurs nucléaires

Si le rôle des facteurs de transcription gouvernant les interactions entre le noyau et les mitochondries commence à être bien compris, l'effet des co-activateurs nucléaires impliqués dans la biogenèse mitochondriale reste pour le moment moins étudié. Les co-activateurs de la famille PGC-1 (PPAR γ coactivator) jouent un rôle important dans le contrôle de la biogenèse et du fonctionnement de la mitochondrie en intégrant des signaux externes et internes et en augmentant de manière coordonnée l'activité des facteurs de transcription impliqués dans la biogenèse mitochondriale. Cette famille est composée de 3 co-facteurs: PGC-1 α , PGC-1 β et PRC qui ont des domaines fonctionnels similaires et ont un rôle important pour la régulation des gènes impliqués dans le fonctionnement mitochondrial (Figure 11).

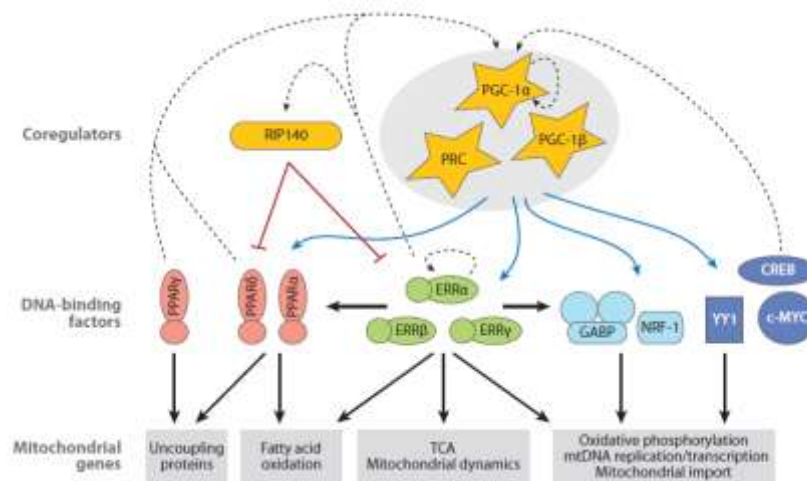


Figure 11: fonctions mitochondriales régulées par les coactivateurs transcriptionnels PGC-1

Les facteurs de transcription PPAR, ERRs, NRFs, CREB, YY1 et cMYC régulent de nombreux gènes codant pour des protéines mitochondriales. Ils forment différents complexes permettant une régulation plus fine des fonctions mitochondriales. Les cofacteurs PGC-1 interagissent avec de multiples facteurs pour coordonner la régulation de nombreuses classes de gènes mitochondriaux [Hock et Kralli, 2009].

a. PGC-1 α et PGC-1 β

PGC-1 α est impliqué dans la régulation de la thermogénèse adaptative. En effet, son expression est considérablement augmentée dans la graisse brune après exposition au froid [Puigserver et al., 1998]. La thermogénèse adaptative est rendue possible par la forte induction par PGC-1 α de la biogenèse mitochondriale. Ce co-activateur induit l'expression des facteurs de transcription mitochondriaux par l'intermédiaire de NRF-1 et NRF-2 en augmentant ainsi la transcription des gènes mitochondriaux [Gleyzer et al., 2005]. De plus, l'expression de PGC-1 α est augmentée en présence de monoxyde d'azote.

PGC-1 β , dont la séquence est très proche de PGC-1 α , est également exprimé dans les mêmes tissus. En revanche, son expression ne varie pas lors de changements de température, comme cela est observé pour PGC-1 α . PGC-1 β induit une augmentation massive des mitochondries dans le muscle squelettique, augmentant ainsi la consommation d'oxygène dans ce tissu. Ces 2 co-facteurs voient donc leur expression induite par différents stress.

b. PRC

PRC, comme les 2 autres facteurs de la famille PGC-1, interagit avec NRF-1 et permet l'activation des gènes cibles de ce facteur [Andersson et R C Scarpulla, 2001]. PRC permet également, par l'intermédiaire de NRF-1 et NRF-2, d'activer l'expression des facteurs de transcription mitochondriaux. Cependant, aucune différence d'expression de PRC n'a été observée entre la graisse brune et la graisse blanche, et son expression est seulement légèrement augmentée dans la graisse brune lors d'une exposition au froid [Andersson et R C Scarpulla, 2001]. PRC, dont l'expression est ubiquitaire, ne semble donc pas avoir de rôle dans la thermogenèse adaptative, contrairement à ses 2 homologues. En revanche, l'expression de PRC est rapidement induite dans les cellules en prolifération, ainsi que par des traitements au sérum de fibroblastes quiescents [Andersson et R C Scarpulla, 2001; Raharijaona et al., 2009; Vercauteren et al., 2006]. Cette induction entraîne une augmentation de l'expression de tous les facteurs de transcription mitochondriaux, ainsi que des gènes nucléaires et mitochondriaux codant pour des sous-unités de la chaîne respiratoire [Gleyzer et al., 2005]. De plus, il a été récemment montré que l'interaction de PRC avec CREB et NRF-1 augmente la réponse à l'activation par le sérum [Vercauteren et al., 2006]. Il a également été montré que l'expression du cytochrome c, dont le promoteur contient les sites de liaisons à CREB et NRF-1, est induite par l'apport de sérum durant la transition entre la phase G0 et G1 du cycle cellulaire dans les fibroblastes quiescents, ce qui a pour effet d'augmenter la respiration mitochondriale. Toutes ces propriétés de PRC sous-tendent une régulation conjointe des fonctions mitochondriales et du cycle cellulaire.

Objectif du travail de thèse

Mon travail de thèse a eu pour objectif d'étudier les mécanismes de régulation transcriptionnelle en conditions physiologiques, et dérégulées en conditions pathologiques. Cela nécessite l'analyse de différents types de données tels que le niveau d'expression des gènes, la fixation de facteurs de transcription, et les motifs correspondant aux sites de fixation de ces facteurs de transcription.

Au cours de ma thèse, j'ai adapté ces stratégies pour étudier les mécanismes de régulation transcriptionnelle du coactivateur PRC, afin de comprendre son rôle dans la coordination transcriptionnelle de la prolifération mitochondriale et cellulaire. Nous avons pour cela utilisé la lignée cellulaire XTC.UC1, cellules dérivant d'un variant oncocytaire de carcinome folliculaire thyroïdien humain. Cette lignée a été utilisée aussi bien pour les analyses de transcriptomique que de ChIP-chip, dans les mêmes conditions d'induction par le sérum. Ceci nous a permis de rechercher les voies de régulation activées à un temps donné et de corréler ces voies au profil transcriptionnel obtenu dans les mêmes conditions. L'exploration par ChIP de plusieurs facteurs de transcription en même temps a eu pour but d'apprécier la complexité des combinaisons de facteurs pour une fonction donnée dans la cellule ainsi que d'appréhender les régulations directes et indirectes exercées par ces facteurs.

Méthodes

L'objectif de notre étude est d'étudier l'impact de PRC sur la coordination de la prolifération mitochondriale et cellulaire. Afin de pouvoir étudier spécifiquement PRC, et non les deux autres membres de la famille PGC-1, nous avons choisi d'étudier une tumeur riche en mitochondries, l'oncocytome thyroïdien, où seul PRC est exprimé. La lignée XTC.UC1 est le seul modèle cellulaire de ce type d'oncocytome. Nous avons donc choisi cette lignée pour étudier les régulations impliquées dans le dialogue nucléo-mitochondrial sous l'effet de PRC.

I- Le transcriptome

Pour appréhender l'effet de PRC sur l'ensemble du transcriptome des cellules XTC.UC1, nous avons choisi d'inhiber l'expression de PRC par l'intermédiaire d'un siRNA. Les siRNAs sont de petits ARN double brin qui reconnaissent des séquences d'ARNm spécifique et entraînent leurs dégradations, empêchant ainsi la traduction de la protéine.

L'objectif de cette étude de transcriptome était de suivre sur 48h l'effet de PRC sur l'ensemble du génome. Pour cela, une étude cinétique a été menée avec des mesures d'expression effectuées à des temps différents : 0h, 12h, 24h et 48h après l'ajout du siRNA PRC. Toutes les cellules utilisées pour cette étude ont été préalablement bloquées en phase G0 du cycle cellulaire par suppression du sérum de veau pendant 24h. Le plan expérimental de cette étude est présenté dans la Figure 12.

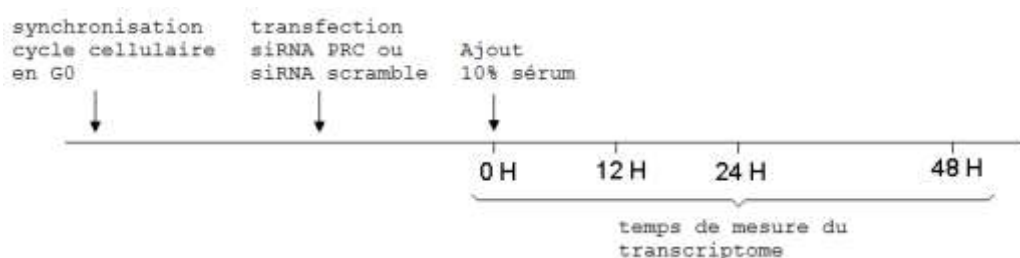


Figure 12: plan expérimental

Schéma expérimental décrivant les étapes appliquées à la lignée cellulaire XTC.UC1 pour réaliser les mesures d'expression cinétique de l'effet du siRNA PRC.

Afin de ne pouvoir identifier l'effet de la transfection du siRNA sur le transcriptome des cellules, nous utilisons un siRNA contrôle, appelé siRNA scramble, que nous transfectons dans des cellules de la même lignée. Celui-ci est transfecté de la même manière que le siRNA PRC, mais il n'a aucun effet d'inhibition puisque ce siRNA

n'a pas de cible spécifique. Les cellules subissent donc exactement le même traitement qu'il s'agisse des contrôles ou des cellules d'intérêt.

Plusieurs points de mesure sont effectués à chaque temps pour chacun des 2 types cellulaires. L'analyse de ces données d'expression permet d'identifier l'ensemble des gènes dont l'expression varie en présence ou non du coactivateur transcriptionnel PRC.

1) Normalisation et clustering des données

La première étape de l'analyse des données de puces à ADN, suite à la quantification du signal, est une étape de normalisation des données. En effet, plusieurs biais expérimentaux peuvent apparaître au cours des étapes de marquage, d'amplification et d'hybridation des différentes puces à ADN. Il est notamment connu que le taux d'incorporation des 2 fluorochromes utilisés pour les puces à ADN est différent. Il en résulte un signal plus important pour un fluorochrome que pour l'autre. De plus, la mesure de l'intensité de l'émission de fluorescence d'une lame à l'autre peut varier indépendamment des échantillons utilisés. Afin de pouvoir comparer les échantillons entre eux, il est indispensable d'éliminer les biais expérimentaux par une étape de normalisation. Ces méthodes reposent sur deux hypothèses principales: la majorité des gènes n'ont pas d'expression différentielle entre les deux conditions étudiées, et les effets systématiques observés ne sont pas des effets biologiques.

Il existe de nombreuses méthodes de normalisation des puces à ADN, plus ou moins spécifiques à la plateforme ou la technologie utilisée. Néanmoins, la méthode du lowess reste une des méthodes les plus classiques dans ce domaine. Le lowess (**l**ocally **w**eighted **s**catterplot **s**moother) est une méthode de régression pondérée locale permettant d'éliminer les biais non linéaires. Chaque échantillon va être normalisé par rapport à un profil de référence, comme par exemple le profil médian, à l'aide d'une fenêtre glissante. Une fonction de régression est calculée pour chaque fenêtre, ce qui va permettre de construire une courbe d'ajustement global pour normaliser les données.

Une fois que les échantillons sont normalisés, on visualise les données grâce à la méthode de classification hiérarchique qui permet de regrouper entre eux les gènes et les échantillons qui se ressemblent le plus. Le clustering hiérarchique est une méthode de classification introduite pour l'analyse des puces à ADN par Eisen en 1998 [Eisen *et al.*, 1998]. Le principe de cette méthode est de calculer une distance, ou une corrélation, entre toutes les paires de gènes et d'échantillons possibles afin de regrouper de manière ascendante les gènes et les échantillons les plus similaires entre eux. L'objectif du

clustering est d'avoir des distances inter-groupes fortes, et des distances intra-groupes faibles. Le clustering hiérarchique permet donc d'identifier des gènes co-régulés, dont l'expression est corrélée, ainsi que des signatures spécifiques d'un type d'échantillons (Figure 13).

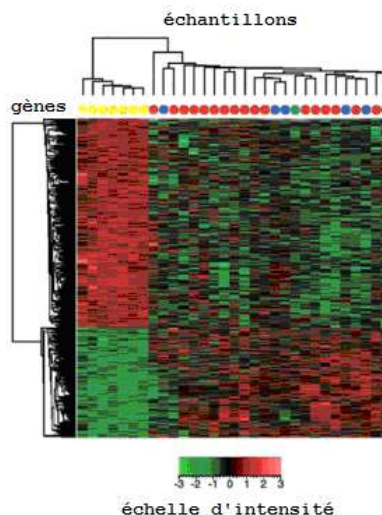


Figure 13: exemple de clustering hiérarchique

Les échantillons (en colonne) et les gènes (en ligne) sont regroupés par clustering hiérarchique. On observe un groupe d'échantillon (en jaune) avec une signature très spécifique par rapport aux échantillons rouges et bleus.

La corrélation de l'expression d'un certain nombre de gènes dans un type d'échantillons donné sous-tend une implication commune de tous ces gènes dans une fonction biologique particulière, ainsi qu'une régulation commune par différents facteurs de transcription [Tavazoie *et al.*, 1999].

2) Annotation fonctionnelle

L'annotation fonctionnelle permet d'identifier la fonction biologique dans laquelle sont impliqués les gènes d'un cluster. Celle-ci se base sur la terminologie GO (Gene Ontology), un vocabulaire structuré et contrôlé pour décrire les gènes et les processus biologiques dans lesquels ils sont impliqués [Ashburner *et al.*, 2000]. Plusieurs outils d'annotation ont été développés à partir des termes GO pour identifier des fonctions biologiques enrichies dans un groupe de gènes. Le principe de ces outils est d'identifier des termes GO statistiquement sur- ou sous-représentés dans le groupe de gènes étudié par rapport à l'ensemble des gènes considérés dans l'étude. L'outil que nous utilisons pour les annotations de groupes de gènes est l'outil GoMiner [Zeeberg *et al.*, 2003]. A partir de la liste des gènes d'un groupe, et de la liste complète des gènes obtenus après

filtrage sur la puce à ADN, cet outil fournit une liste d'annotations fonctionnelles, chacune associée à une p-valeur. En plus de cette p-valeur, nous effectuons un calcul d'enrichissement, afin de tenir compte du nombre de gènes ayant l'annotation considérée. En effet, selon la fonction, le nombre de gènes concernés peut être très variable, ainsi pour une annotation fonctionnelle très répandue, une petite différence peut être très significative, malgré un faible enrichissement, alors que pour une annotation très spécifique, et donc rare, une différence faible peut être peu significative alors que l'enrichissement sera très important. Le calcul de l'enrichissement permet en plus de savoir si l'annotation fonctionnelle est sur- ou sous-représentée dans le cluster, ce que n'indique pas la p-valeur obtenue par un test exact de Fisher avec le logiciel GoMiner.

II- Le ChIP-chip

Afin d'identifier les gènes cibles de PRC et le mode d'action de PRC dans la coordination de la prolifération mitochondriale et cellulaire, nous avons effectué le ChIP-chip de PRC, ainsi que de plusieurs de ses partenaires. Pour limiter les coûts, et pouvoir ainsi étudier plusieurs facteurs de transcription, nous avons tout d'abord cherché à identifier la localisation générale des sites de fixation des facteurs de transcription que nous souhaitons étudier. Pour cela, nous avons mis en place sur la plateforme une puce « hybride », contenant des sondes aléatoirement localisées dans les îlots CpG, dans les promoteurs proximaux, et dans le génome complet. Cette puce a été synthétisée par la société Agilent sur une sélection de sondes que nous avons effectuée. L'hybridation d'un ChIP de chaque facteur sur cette puce permet de mettre en évidence des zones privilégiées par le facteur de transcription étudié. De cette manière, nous avons déterminé que les puces à promoteurs 2x244K d'Agilent étaient les plus adaptées pour pouvoir faire une étude assez complète des cibles de nos facteurs d'intérêts (résultats présentés dans la section suivante).

Les méthodes que j'ai développées au cours de ma thèse sont donc plus particulièrement adaptées à l'analyse de puces à promoteurs, ou de tout autre type de puces dont l'espacement moyen entre deux sondes consécutives sur le génome est supérieure à quelques dizaines de paires de bases (contrairement aux puces « tiling »).

J'ai également contribué, en collaboration avec Audrey Donnart, un ingénieur de la plateforme, au choix stratégique à mettre en place pour la mise au point du ChIP-chip sur la plateforme pour les différents facteurs que nous souhaitons étudier. Nous avons

choisi de tester trois anticorps par facteurs de transcription, et avons sélectionné le plus spécifique à partir des résultats de PRC quantitative. Nous avons également choisi la stratégie de crosslink qui nous semblait la plus appropriée au regard des résultats d'enrichissement, et les temps de sonication nécessaire à l'obtention de fragment d'ADN d'environ 500pb. Tout ce travail a été effectué sur des culots cellulaires de la lignée XTC.UC1 de même passage contenant un million de cellules 48h après induction par le sérum de veau. J'ai initié la biologiste aux bases de linux afin de lui permettre de lancer des outils d'analyses que j'ai développés, et qu'elle puisse observer directement les enrichissements obtenus lors de ces expérimentations. Ceci lui a permis de pouvoir progresser dans ces mises aux points grâce à ce dernier contrôle de qualité.

1) Analyse des données de ChIP-chip

Pour chaque étude de ChIP-chip réalisée, nous hybridons 2 échantillons sur la puce à promoteurs (Figure 14). Le premier échantillon, marqué avec le fluorochrome vert (cyanine 3), correspond à l'ensemble des séquences immunoprécipitées avec l'anticorps spécifique du facteur de transcription étudié. Par la suite, j'appellerai cet échantillon IP. Le deuxième échantillon hybridé, marqué avec le fluorochrome rouge (cyanine 5), correspond à l'échantillon contrôle, contenant donc l'ensemble de l'ADN génomique. Cet échantillon, qui n'a pas subi d'immunoprécipitation, sera appelé par la suite input.

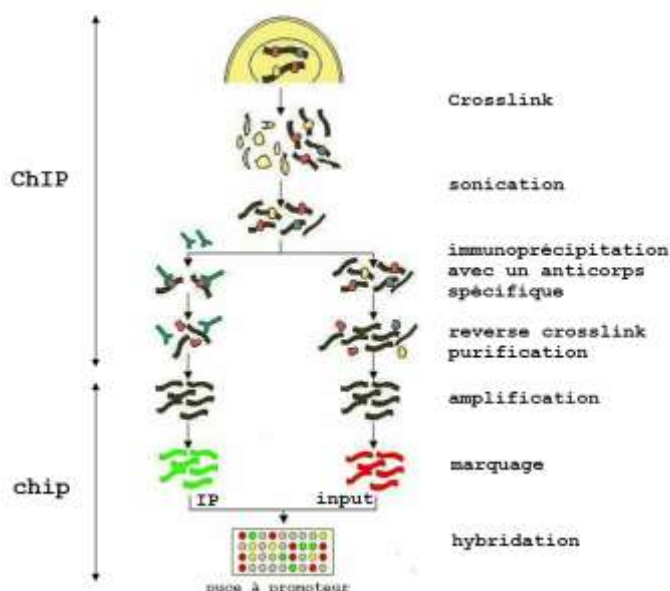


Figure 14: étapes du ChIP-chip

Suite au crosslink permettant de lier de manière covalente les protéines à l'ADN, les cellules sont lysées et l'ADN est fragmenté par sonication. Le pool d'ADN est ensuite séparé en deux tubes. Le premier, appelé IP, est immunoprécipité avec l'anticorps spécifique du facteur de transcription étudié, alors que le second (input) ne subit pas d'immunoprécipitation. Les liaisons covalentes ADN-protéines sont ensuite inversées, puis l'ADN est purifié, amplifié et marqué. Les deux échantillons sont alors hybridés sur une puce à ADN.

a. Normalisation des données

Comme pour l'analyse du transcriptome, la première étape de l'analyse des données de ChIP-chip est une étape de normalisation. Nous avons choisi d'utiliser la même méthode de normalisation que pour le transcriptome, l'ajustement local de type lowess. Cependant, nous avons dû adapter cette méthode car, pour le ChIP-chip, nous utilisons deux lames contenant des sondes différentes pour un même échantillon. De ce fait, il faut normaliser les deux lames indépendamment sur un même profil pour tenir compte par exemple d'un calibrage différent du laser pour la lecture des deux lames, ou d'une quantité de dépôt différente pour les deux lames. Pour cela, j'ai construit un profil moyen à partir des profils médians ordonnés des 2 lames (Figure 15). J'ai ensuite normalisé chacune des deux lames sur ce profil moyen, par la méthode Lowess, et j'ai enfin regroupé l'ensemble des données.

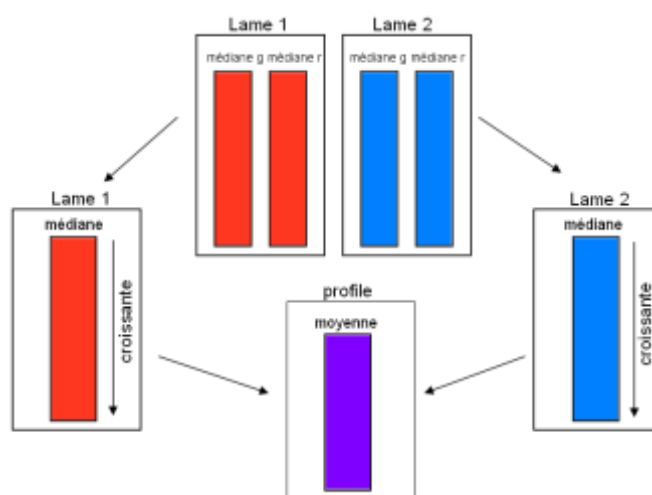


Figure 15: construction du profil moyen pour le Lowess

Un profil médian ordonné est calculé pour chacune des deux lames (rouge et bleue), puis un profil moyen (violet) est construit à partir de ces deux profils. Le profil moyen est utilisé pour normaliser les deux paires de lames indépendamment.

b. Sélection des sondes positives en ChIP

Les sondes positives en ChIP sont les sondes dont l'intensité en vert (IP) est significativement supérieure à l'intensité en rouge (input). Notre méthode de sélection est basée sur l'hypothèse qu'il ne devrait pas exister de sondes négatives en ChIP (plus intense pour l'input que l'IP), et que le bruit de fond est symétrique autour de l'axe de pente 1 centré sur 0 suite à la normalisation. Nous traçons donc dans un premier temps le graphe de l'intensité normalisée des sondes avec en abscisse l'input et en ordonnée l'IP (Figure 16).

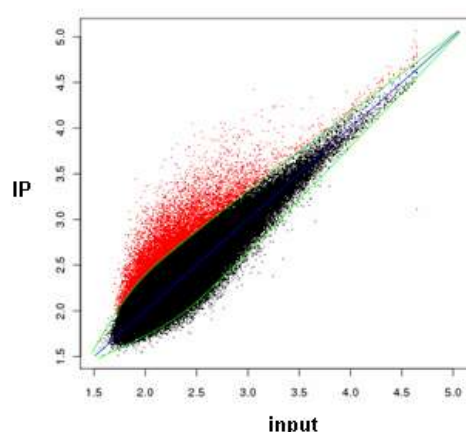


Figure 16: sélection des sondes positives en ChIP

Les données en abscisse correspondent au logarithme de l'intensité normalisée de l'input, et en ordonnée de l'IP. La courbe de régression polynomiale (verte) permet de sélectionner, par symétrie axiale autour de l'axe $y=x$ en bleu, les sondes positives en ChIP (en rouge).

Nous calculons ensuite une courbe de régression polynomiale (en vert) dans la partie inférieure du graphe, correspondant au bruit de fond (les points qui s'éloignent le plus de la droite de pente 1, relativement à son voisinage). Une approximation de cette courbe est obtenue par des splines, i.e. une fonction continue par morceaux par des polynômes de degré 4, ce qui permet d'obtenir une valeur pour toutes les abscisses possibles. Cette courbe isole les 5% des points les plus éloignés que nous considérons comme le bruit de fond normal de l'expérience. Nous reportons ensuite cette courbe par symétrie axiale dans la partie supérieure du graphe, ce qui nous permet de sélectionner les sondes dites positives en ChIP, en rouge sur le graphe.

Nous calculons enfin une p-valeur pour chacune des sondes positives. Pour ce calcul, nous faisons l'hypothèse que la distance des sondes positives obtenue à partir de la régression à la droite $y=x$ suivent une distribution normale dont la variance est déterminée à l'aide de la courbe verte (qui correspond au seuil de 95% de la loi normale). Cette p-valeur est finalement calculée en fonction de la distance de la sonde par rapport au seuil de significativité précédemment défini.

c. Traitement des réplicats et annotation fonctionnelle

Pour chaque facteur de transcription étudié, nous avons réalisé 2 réplicats complètement indépendants, afin de s'assurer d'une bonne reproductibilité des résultats. Pour vérifier cette reproductibilité, j'ai calculé la corrélation entre toutes les sondes de chacun des 2 réplicats, ainsi que pour la meilleure sonde de chaque gène des deux réplicats. En effet, étant donné que la sonication a lieu aléatoirement sur le génome, il est

possible que d'un réplicat à l'autre ce ne soit pas la même sonde qui s'hybride le plus fortement, mais par exemple la sonde voisine. Choisir la meilleure sonde par gène pour chaque réplicat permet de s'assurer que nous obtenons bien des résultats cohérents en terme de gènes dans les deux expérimentations.

Après validation des deux réplicats, nous avons choisi une méthode permettant de combiner deux p-valeurs. Les méthodes classiques pour effectuer cette combinaison sont la sélection du minimum (ou du maximum), la moyenne arithmétique ou la moyenne géométrique. Notre but étant de prendre en compte les deux p-valeurs, nous avons éliminé le premier choix. Ensuite, étant donné que nous travaillons sur des petites valeurs, très inférieures à 1, nous avons choisi d'utiliser la moyenne géométrique plutôt qu'une moyenne arithmétique. Cela permet de ne pas donner de poids trop important à une valeur peu significative. En effet, la moyenne géométrique consistant en une moyenne des puissances, le résultat sera plus représentatif des deux valeurs qu'une moyenne arithmétique, qui tend à favoriser la valeur la plus proche de 1. Par exemple, la moyenne arithmétique de 10^{-2} et 10^{-6} est quasiment égale à 10^{-2} (la plus grande des deux valeurs) et ne reflète pas la très bonne p-valeur à 10^{-6} . Par contre, la moyenne géométrique de ces valeurs est égale à 10^{-4} . J'ai également éliminé les sondes dont la p-valeur était supérieure à 0,01. Pour qu'une sonde soit considérée comme positive en ChIP, il faut qu'elle remplisse la condition:

$$\sqrt{pVal1 \times pVal2} < 10^{-4} \ \& \ pVal1 < 10^{-2} \ \& \ pVal2 < 10^{-2}$$

A partir des annotations des sondes fournies par Agilent, j'ai pu constituer la liste des gènes positifs en ChIP pour chaque facteur de transcription. Les annotations fonctionnelles de chaque liste de gènes positifs en ChIP sont réalisées de la même manière que pour le transcriptome.

2) Découverte de motifs

La découverte de motifs dans les séquences positives en ChIP permet non seulement de localiser précisément le site de fixation du facteur de transcription étudié en ChIP, et par la même occasion de valider la réussite du ChIP, mais aussi d'identifier des sites de fixation de cofacteurs à proximité du facteur étudié. Pour effectuer cette

recherche, j'ai utilisé la séquence de chaque sonde positive en ChIP, ainsi que les 200 bases en amont et en aval de celle-ci (Figure 17).

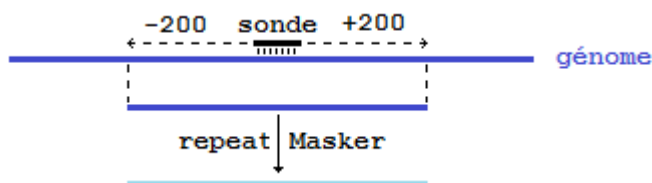


Figure 17: préparation des séquences pour la découverte de motifs

Pour chaque sonde, on récupère une portion du génome encadrant la sonde qui est ensuite masquée pour éliminer les zones de faibles complexités et les transposons.

J'ai choisi de travailler avec des séquences d'environ 450 paires de bases seulement car nous recherchons des motifs de taille assez petite, à partir de 6 bases. Travailler sur de trop grandes séquences rendrait la découverte d'un motif de taille $n=6$ non significative, étant donné qu'un motif de cette taille peut apparaître au hasard toutes les 6^4 bases sur le génome. J'ai ensuite masqué les régions de faible complexité ainsi que les transposons de type SINEs, LINEs, à l'aide de l'outil repeatMasker (<http://repeatmasker.org>).

Ce masquage des séquences est indispensable pour la plupart des outils de découverte de motifs, qui identifieraient dans le cas contraire majoritairement des motifs dans ces régions très répétées et de faible complexité.

a. Les méthodes de découverte de motifs

J'ai choisi un outil de découverte de motifs basé sur un algorithme de comptage de mots, et qui tient compte des résultats de ChIP-chip. J'ai fait ce choix car il est bien adapté à nos études de ChIP, et car le nombre et la taille des séquences à analyser sont trop grands pour utiliser convenablement des méthodes basées sur de l'échantillonnage de Gibbs. De plus, les résultats ne sont pas dépendants des paramètres initiaux comme pour les méthodes locales.

L'outil que j'ai utilisé pour traiter les données est MDmodule, issu du pipeline motifRegressor [Conlon et al., 2003] ; [Schena et al., 1995]. Cet outil effectue la découverte de motifs en favorisant ceux qu'il découvre dans les séquences les plus significatives en ChIP, et qui sont enrichis dans l'ensemble du jeu de séquences. Ceci nécessite de fournir en entrée du programme un jeu de séquences ordonnées en fonction de leurs significativités. Étant donné que nous ne connaissons pas *à priori* la taille des

motifs que nous recherchons, j'ai choisi de sélectionner les 10 meilleurs motifs pour chaque taille dans une fourchette allant de 6 à 18 bases. J'obtiens donc un maximum de 130 motifs, nombre rarement atteint car il n'existe généralement pas 10 motifs de grande taille significatifs dans nos jeux de séquences. Cette démarche de découverte de motifs entraîne une certaine redondance dans les résultats car les motifs de petite taille peuvent être imbriqués, dans une forme éventuellement légèrement dégénérée, dans des motifs de taille supérieure. De plus, certains motifs de même taille peuvent être chevauchants, et il est dans ce cas intéressant de choisir le motif le plus grand englobant ces 2 sous motifs. Enfin, il est souvent possible de découvrir un motif et son complémentaire inverse de façon indépendante. Toutes ces raisons rendent indispensable une étape de comparaison et de regroupement des motifs afin d'obtenir un nombre plus restreint de motifs, et correspondant chacun à un plus grand nombre de gènes.

b. Comparaison de motifs

Afin de pouvoir comparer et regrouper les motifs similaires entre eux, j'ai développé au cours de ma thèse l'outil motifsComparator [Carat et al., 2010]. Cet outil disponible publiquement sur le web (<http://cardioserve.nantes.inserm.fr/motifsComparator/>) permet de comparer les motifs fournis par l'utilisateur entre eux, ainsi qu'avec ceux des deux bases de données de motifs JASPAR et Transfac. J'ai pour cela utilisé la dernière version de JASPAR (JASPAR CORE non redondant) et la dernière version publique de Transfac (Transfac Rel.3.2).

Cette méthode compare tous les motifs fournis par l'utilisateur deux à deux, ainsi qu'avec tous les motifs des deux bases de données. Elle permet de comparer deux motifs de taille identique ou non, avec tous les décalages possibles de ces deux motifs entre eux, en conservant une superposition minimale des deux motifs d'au moins six bases. La comparaison est également effectuée avec le complémentaire inverse de chacun des motifs. Enfin, le motif est comparé à son propre complémentaire inverse, ce qui permet de détecter des palindromes, séquences caractéristiques de certains sites de fixation de facteurs de transcription. Notre méthode globale de comparaison prend en compte uniquement les bases à chaque position dont la fréquence dépasse 5% pour au moins un des deux motifs. (Figure 18). Bien qu'elle effectue un très grand nombre de comparaisons, notre méthode reste rapide car elle profite d'une algorithmique poussée (utilisation au maximum de méthodes de programmation dynamique pour éviter des

calculs inutiles) et d'une implémentation en parallèle pour tirer profit de la capacité de nos serveurs.

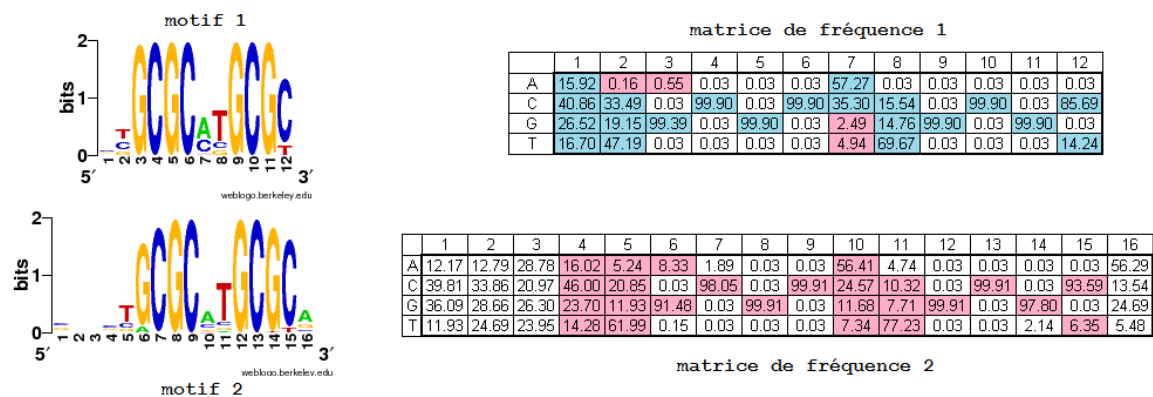


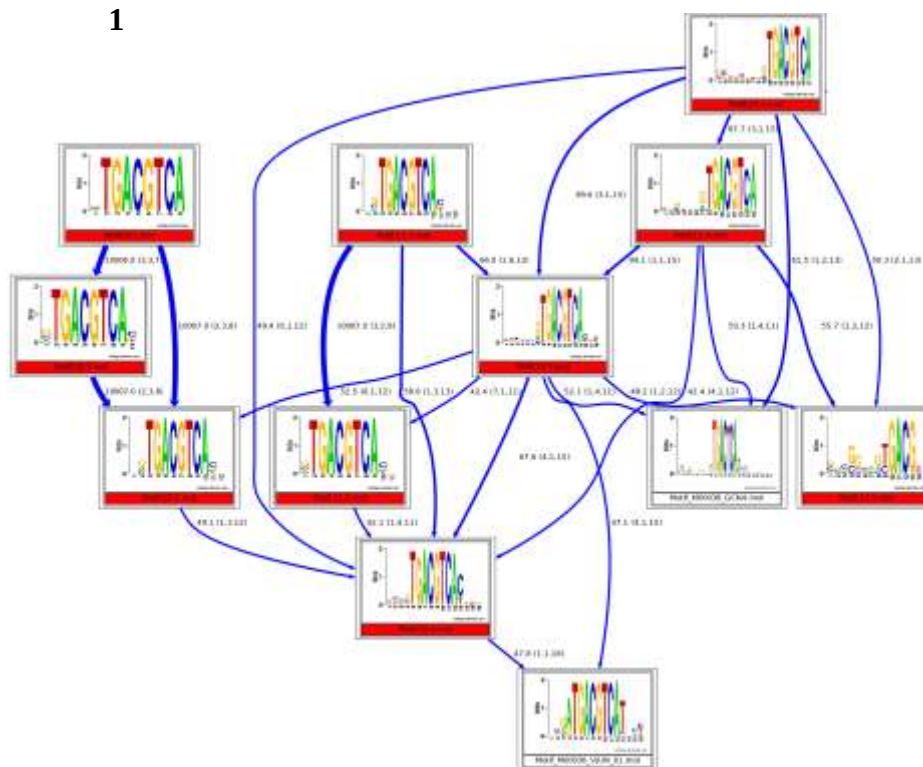
Figure 18: principe de comparaison de 2 motifs

Pour chaque décalage possible des deux motifs, on construit deux distributions correspondant à la fréquence de chaque base à chaque position dépassant un seuil fixé (ici 5%). La distribution correspondant à la matrice de fréquence 1 est composée des fréquences bleues, ainsi que des fréquences roses correspondant à des bases dont la fréquence dépasse 5% dans la matrice de fréquence 2.

Cette méthode permet d'éviter de fréquents biais rencontrés par d'autres méthodes de comparaison, qui considèrent comme différents les motifs s'il existe de petites variations entre deux fréquences faibles. De plus, la comparaison globale, contrairement aux méthodes de comparaison colonne par colonne, permet de ne pas biaiser le calcul de la p-valeur. En effet, dans le cas d'une comparaison par colonne, il suffit qu'une des colonnes soit identique dans les deux motifs pour que la p-valeur globale de comparaison des deux motifs soit égale à zéro. Notre méthode globale n'est pas biaisée par ce phénomène puisque nous n'effectuons qu'un seul calcul de p-valeur sur la distribution globale de toutes les fréquences supérieures à 5% des deux motifs. Un test de χ^2 d'homogénéité permet alors de comparer les 2 distributions, et d'attribuer une p-valeur à la comparaison des 2 motifs.

Une distance est calculée pour chaque paire de motifs, et les motifs significativement comparables sont alors regroupés en graphe (Figure 19-1). Dans ce graphe, les motifs fournis par l'utilisateur sont soulignés en rouge, et ceux issus des bases de données publiques en blanc. Une autre représentation des résultats en dendrogramme est également possible (Figure 19-2). Cette représentation concerne uniquement les motifs fournis par l'utilisateur.

1



2

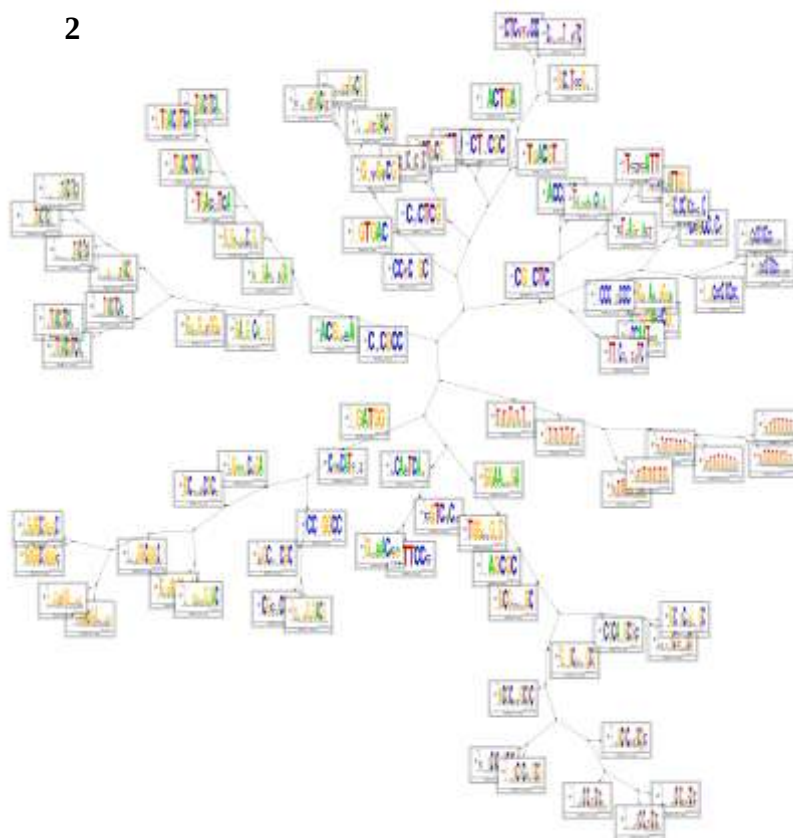


Figure 19: exemple de regroupement de motifs

Résultats de motifsComparator pour le regroupement de motifs. Le graphe 1 permet de regrouper 10 motifs (en rouge) découverts dans un jeu de séquences positives en ChIP et 2 motifs issus des bases de données publiques (en blanc). Le dendrogramme 2 permet de visualiser l'ensemble des motifs fournis par l'utilisateur en fonction de leurs ressemblances.

c. Validation des motifs

Une fois les motifs découverts et regroupés, une dernière étape de l'analyse consiste à vérifier la spécificité de ces motifs dans le jeu de séquences positives en ChIP, vis-à-vis de motifs enrichis dans le génome complet. Pour cela, nous construisons un jeu de séquences contrôles, similaire au jeu de séquences positives en ChIP, à partir d'un tirage au hasard dans le jeu de séquences non positives en ChIP. Ensuite, pour chaque motif nous calculons, à partir de la matrice PWM du motif, le meilleur score par séquence pour les 2 jeux de séquences. Ensuite, nous traçons la distribution des scores pour le jeu de séquences positives en ChIP et pour le jeu de séquences contrôles, et nous réalisons enfin un test de χ^2 d'homogénéité pour comparer les 2 distributions (Figure 20).

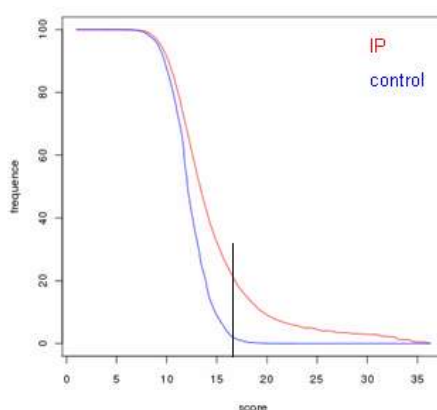


Figure 20: Méthode de validation d'un motif

Distribution de la fréquence des scores d'un motif pour les séquences positives en ChIP en rouge et pour les séquences contrôles en bleu. Le seuil choisi permet de sélectionner le score qui sépare le mieux les deux courbes.

Si le test de χ^2 est significatif, nous fixons comme seuil de sélection le score qui correspond au plus grand écart entre les 2 courbes (IP *versus* contrôle). Nous sélectionnons alors toutes les séquences positives en ChIP ayant obtenu un score qui dépasse le seuil défini précédemment. Si le test n'est pas significatif, nous éliminons le motif pour nos analyses.

III- Intégration de données hétérogènes

Afin d'avoir une vision globale des mécanismes de régulation du coactivateur transcriptionnel PRC et de son impact sur la régulation de l'ensemble des gènes du

génomique, il est nécessaire de combiner les données issues du transcriptome et des différents ChIP réalisés.

1) ChIP-chip + ChIP-chip

L'intégration de données issues de ChIP-chip de différents facteurs de transcription permet d'identifier des modules de facteurs de transcription. Cette étude est indispensable pour identifier les facteurs interagissant avec le cofacteur transcriptionnel PRC, étant donné que celui-ci n'a pas de domaine de liaison à l'ADN. Ceci peut également permettre de mettre en évidence différents complexes de facteurs de transcription, pouvant agir de manière plus spécifique sur une fonction biologique particulière ou sur un groupe de gènes précis. Nous avons mis en place plusieurs stratégies pour analyser les résultats des différents ChIP que nous avons réalisés.

Pour étudier les gènes cibles de deux facteurs de transcription, nous calculons la corrélation de la meilleure sonde de chaque gène entre les deux facteurs. Nous pouvons ensuite sélectionner les gènes dont au moins une sonde est positive pour chacun des deux facteurs. Il peut s'agir de la même sonde, ou de deux sondes plus distales sur le promoteur.

Pour analyser les gènes cibles de trois facteurs de transcription, nous avons utilisé les diagrammes de Venn, qui permettent d'observer les intersections entre trois listes différentes de gènes. Ceci permet non seulement d'analyser les gènes cibles communs à trois facteurs, mais également ceux spécifiques à chacun d'entre eux.

Enfin, pour observer simultanément les résultats de tous les facteurs de transcription, nous avons construit une matrice que nous traitons de manière similaire aux données de transcriptome, avec une carte thermique. Chaque colonne correspond à un facteur de transcription, et chaque ligne à la p-valeur moyenne des deux réplicats de chaque facteur. Ensuite, les p-valeurs ont été codées afin d'obtenir un code couleur clairement analysable comme décrit dans le Tableau 2:

Tableau 2: codage des p-valeurs pour le clustering

p-valeur (x)	$x \geq 10^{-2}$	$10^{-2} < x \leq 10^{-3}$	$10^{-3} < x \leq 10^{-4}$	$x < 10^{-4}$
Code de saturation en rouge	0	0.3	0.6	1

Les données sont ensuite traitées par classification hiérarchique afin d'identifier des groupes de gènes régulés par un complexe particulier de facteur de transcription. Ces groupes de gènes sont annotés à l'aide de GoMiner afin d'identifier des fonctions biologiques particulières associées à ces complexes.

2) ChIP-chip + transcriptome

L'intégration des données de ChIP-chip et de transcriptome a pour objectif d'identifier les modifications d'expression des gènes induites par les complexes de facteurs de transcription étudiés. Cette analyse est rendue possible car nous avons travaillé avec les mêmes cellules (mêmes passages) et traitées par le sérum de veau sur une même durée pour le transcriptome et le ChIP-chip. La première solution envisagée pour analyser conjointement ces résultats est de rechercher des surreprésentations de gènes enrichis en ChIP dans les clusters obtenus dans l'analyse du transcriptome. Une autre possibilité est d'identifier des fonctions biologiques communes entre un cluster et un groupe de gènes régulés par un complexe de facteurs de transcription.

3) Construction de réseaux de régulation transcriptionnelle

La construction de réseaux de régulation transcriptionnelle est l'étape ultime de l'intégration de données. Elle nécessite une compréhension complète des mécanismes de régulation afin d'identifier tous les gènes impliqués, ainsi que le rôle des différents facteurs de transcription intervenant dans la régulation. Une fois établi, ce réseau permettra d'identifier des gènes clés à l'origine des événements de dérégulation sous-jacents à la prolifération cellulaire et mitochondriale. Cette compréhension permettra d'envisager des solutions pour rétablir des processus physiologiques.

Résultats

I- motifsComparator

L'objectif de la mise en place de motifsComparator était de pouvoir comparer des motifs de taille variable afin de pouvoir les identifier grâce aux bases de données publiques, éliminer la redondance et regrouper les motifs imbriqués ou chevauchants. Comme le montre la Figure 21, ma méthode de comparaison permet de regrouper tous les motifs de la famille HSF (Heat Shock Factors) présents dans la base de données JASPAR (Figure 21-1), des motifs chevauchants de taille variable en fonction du membre étudié. De même, l'outil permet de regrouper les motifs communs de JASPAR et de Transfac, comme c'est le cas par exemple des motifs du facteur de transcription RORA (Figure 21-2).

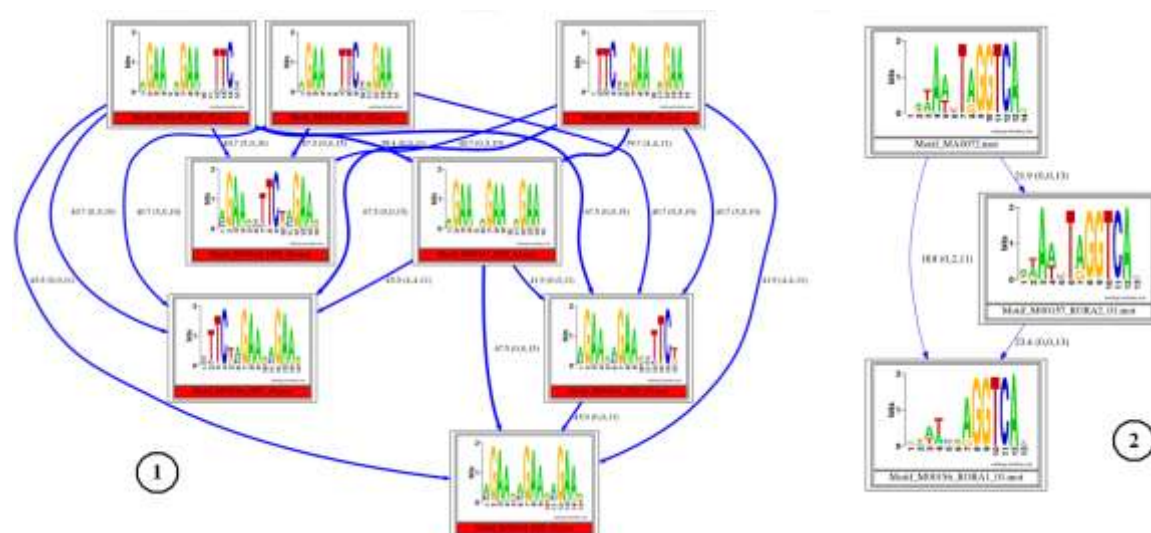


Figure 21: regroupement de motifs similaires avec motifsComparator

Le graphe 1 montre le regroupement des 8 motifs de la famille HSF (Heat Shock Factors) présents dans la base de données JASPAR. Le graphe 2 permet de regrouper les motifs du facteur de transcription RORA (Retinoid related Orphan Receptor Alpha) présents dans les 2 bases de données utilisées dans motifsComparator, JASPAR et Transfac.

Mon outil permet également d'identifier des motifs correspondant à des sites de fixation de facteur de transcription connus grâce à la comparaison des motifs découverts dans les séquences positives en ChIP avec les bases de données publiques. La Figure 22 présente le cas de plusieurs motifs découverts dans un jeu de test de séquences positives en ChIP ESR1, identifié comme similaire au motif PPARGgamma (MA0066) et ESR1

(MA0112) dans la base de données Jaspar. Le motif PPARgamma est similaire à celui d'ESR1 de part son appartenance à la même famille de facteur que ESR1, la famille NHR (Nuclear **H**ormone **R**eceptor).

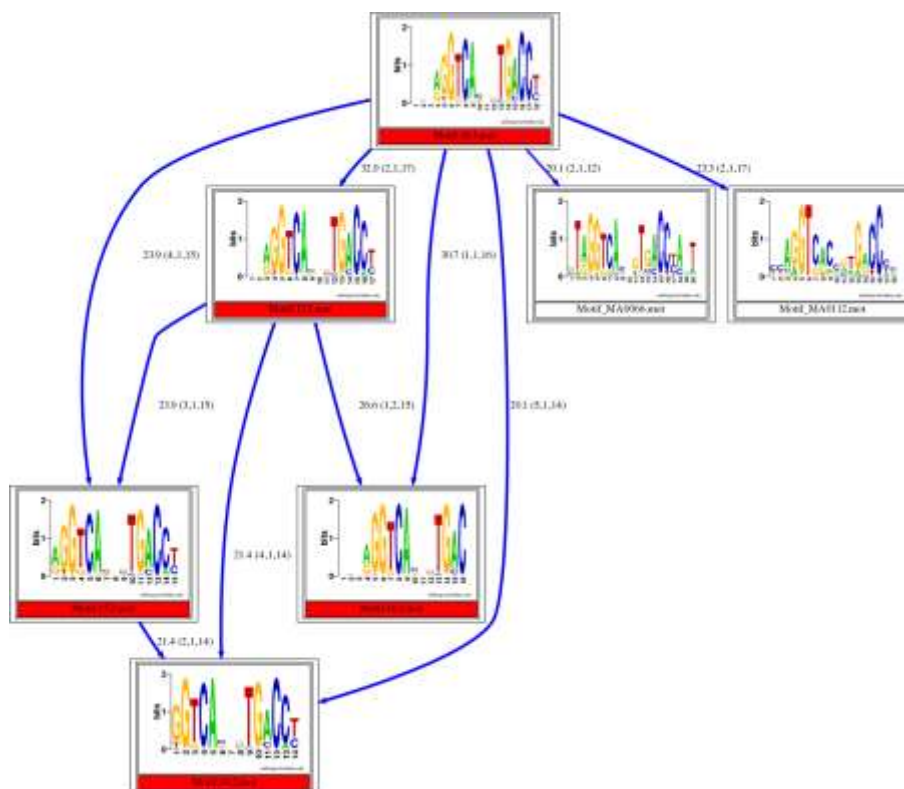


Figure 22: identification du motif ESR1

Les motifs en rouge correspondent aux motifs fournis par l'utilisateur, et ceux en blanc correspondent aux motifs provenant des bases de données JASPAR et TRANSFAC et identifiés comme ressemblant à ceux de l'utilisateur par motifsComparator. Ici, les motifs de l'utilisateur correspondent aux sites de fixation des facteurs de transcription ESR1 et PPARγ.

Enfin, mon outil propose un ensemble de motifs à regrouper, avec les décalages nécessaires à effectuer afin de pouvoir construire un motif consensus à partir de tous ces motifs communs. L'intérêt de la construction du motif consensus est que celui-ci permet d'identifier un nombre de gènes cibles beaucoup plus important (Figure 23).

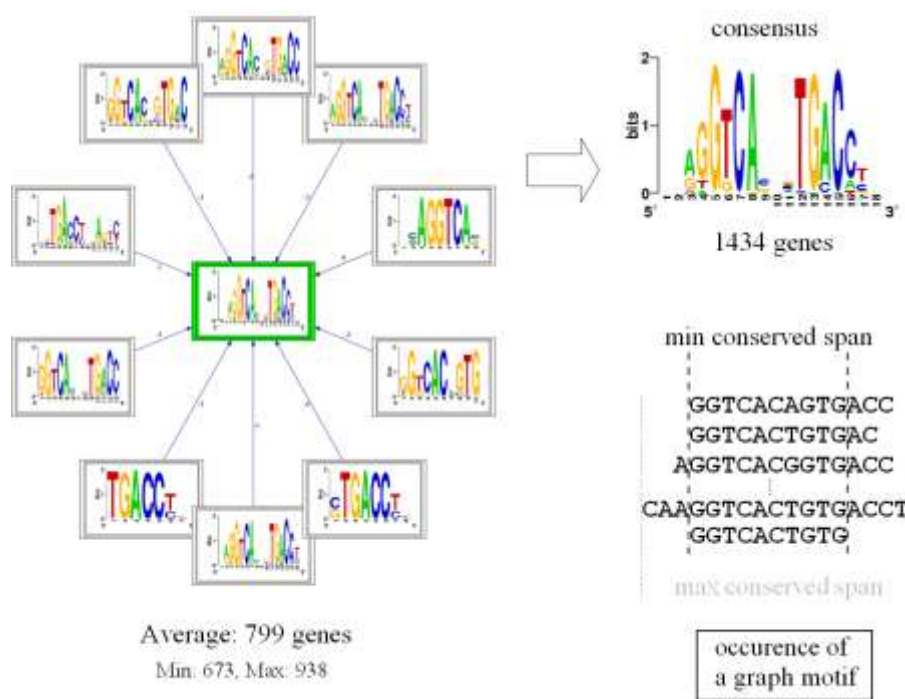


Figure 23: construction d'un motif consensus

La construction d'un motif consensus à partir de l'ensemble des séquences positives en ChIP contenant l'un des 11 motifs du graphe permet d'obtenir un motif présent dans la région régulatrice de 1434 gènes, alors que chacun des motifs pris séparément concerne seulement en moyenne 800 gènes.

Dans le cas présent, le motif consensus généré à partir de l'ensemble des séquences contenant l'un des 11 motifs découverts dans le jeu de séquence test correspond à 1434 gènes différents. Le nombre de gènes concernés est donc presque doublé par rapport au nombre moyen initial de gènes identifiés par les motifs isolés, qui était alors de 799. De plus, ce regroupement permet d'éliminer une forte redondance des motifs due à la découverte de mots allant de la taille $n=6$ à $n=18$.

L'outil motifsComparator, présenté au congrès international MCCMB'09 (**M**oscow **C**onference on **C**omputational **M**olecular **B**iology 2009), a ensuite été publié en version étendue en janvier 2010 dans la revue JBCB (**J**ournal of **B**ioinformatics and **C**omputational **B**iology).

A PARALLEL SCHEME FOR COMPARING TRANSCRIPTION FACTOR BINDING SITES MATRICES

SOLENNE CARAT* and RÉMI HOULGATTE†

Institut du thorax, INSERM U 915

Université de Nantes, France

**Solenne.Carat@univ-nantes.fr*

†Remi.Houlgatte@univ-nantes.fr

JÉRÉMIE BOURDON

LINA CNRS UMR 6241, Université de Nantes, France

INRIA Rennes-Bretagne-Atlantique, France

Jeremie.Bourdon@univ-nantes.fr

Received 15 October 2009

Revised 16 December 2009

Accepted 15 January 2010

Gene regulation implies many mechanisms. Their identification is a crucial task to construct regulatory networks, and is necessary to understand the pathology in many cases. This requires the identification of transcription factors that play a role in regulation. Numerous motif discovery tools are now available. Combining efficiently their results appears useful for comparing and clustering these motifs in order to reduce redundancies and to identify the corresponding transcription factor. We develop a method that produces, compares and clusters a set of motifs and identifies some close motifs in databases like JASPAR and the public version of Transfac. Unlike previous comparison methods, where each matrix column is compared independently, we have developed a global method to compare motifs that also helps to reduce the number of false positives. We also propose an original graph motif model that generalizes the classical position specific pattern matrices. Finally, we present an application of our method to study ChIP-chip data sets in the context of an eukaryotic organism.

Keywords: Transcription factor binding sites identification; PWM clustering; ChIP-chip analysis; Regulatory Networks.

1. Introduction

One of the most challenging problems in genomics is the understanding of mechanisms involved in gene regulation. Identifying all transcription factors having a role in a living system is thus crucial to construct gene regulatory networks. Transcription factors usually recognize specific transcription factor binding site (TFBS) on DNA. These TFBS are commonly represented by position specific matrices, like

Position Weight Matrices (PWM)¹ or Position Frequency Matrices (PFM)² that allows recognition of a slightly conserved part of the DNA sequence. Given a set of DNA sequences where a transcription factor has bound, one first has to extract conserved parts in the sequences. Such parts are putative binding sites. Many tools based on several methods, such as Expectation-Maximisation, Gibbs sampling and word counting, have developed over the last decade to identify these binding sites. Expectation maximisation (EM) is a local optimization procedure to maximize a likelihood function with hidden variables, but it is sensitive to its initialization point. One tool of the most used based on EM algorithm is MEME.³ Gibbs sampling is a general technique to perform probabilistic inference. Unlike EM, gibbs sampling algorithm is based on global search upon a parameterized distribution. However, reestimation of parameters based on randomly generated samples is time consuming because of large number of iterations. AlignACE⁴ is an example of Gibbs sampling algorithm. Finally, several word counting methods integrate supplementary information to sequences such as phylogenetical conservation, ChIP-chip results, to detect more significant patterns. For instance, MDscan⁵ or MotifRegressor⁶ use ChIP-chip result to refine analysis. All these tools produce significant but different results and the use of a combination of several tools provides an exhaustive analysis of TFBS.⁷ The following step is the comparison of the different PWM. The most used approach to quantify PWM similarities is based on a column-by-column comparison with various distance measures.^{8,9} We have developed a method that compares and clusters a set of motif to build a subset of pertinent and non-redundant motifs. Our method uses an improved PWM comparison approach based on a global similarity search with filtering of non-informative positions. Such a global distance ensures that several biases (see Sec. 2.1.4) induced by a column-by-column comparison are removed. Then, a threshold clustering of PWM is performed. A PWM consensus is built from all the threshold clusters. All these PWM consensus are further compared to motif databases (e.g. JASPAR¹⁰ and the public release of Transfac¹¹), to characterize their similarities with known motifs, and searched singularities such as palindromes, tandem repeats, simple repeats. . . Our complete method has been implemented with some computational tricks (parallelization, dynamic programming) that allow to deal with some large sets of patterns. Version 1 of our tool, motifsComparator, is limited to PWM clustering and further versions will include the whole process.

2. Methods

In this section, we describe precisely our method. The first part corresponds to the definition of an appropriate distance between two pattern matrices. Most of them can be adapted to compare both Position Weight Matrices and Position Frequency Matrices. Next, we discuss some clustering methods that can be used in several types of use (visualization, graph motifs, consensus, . . .). Finally, some useful implementation tricks are described.

2.1. Pattern matrices comparison

The capacity to compare pattern matrices corresponding to transcription factor binding sites is essential to avoid redundancies and to identify corresponding transcription factor from known matrices available in public databases. Several methods have already been developed to compare PWM. Most of them consider PWM like a product multinomial distribution in which each column is a set of independent observations. PWM comparison is reduced to a column-by-column comparison.¹² Here, we discuss five main methods based on this principle. Then, we define a new distance that allows us to compare pattern matrices in a more global way.

In the remainder of the section, we will use the following notations:

- $\mathcal{A}$ denotes an alphabet (typically, $\mathcal{A} = \{\text{A, C, G, T}\}$) for DNA sequences, endowed with a probabilistic background model (i.e. the prior probability of letter A is p_A);
- $P = (P_{i,\sigma})_{i \in \{1, \dots, n\}, \sigma \in \mathcal{A}}$ and $Q = (Q_{i,\sigma})_{i \in \{1, \dots, n\}, \sigma \in \mathcal{A}}$ denote two n -length pattern scoring matrices;
- $\bar{P}_i = \sum_{\sigma \in \mathcal{A}} p_\sigma P_{i,\sigma}$ is the expectation of the i th column of P .
- $\hat{P}_i = \sum_{\sigma \in \mathcal{A}} P_{i,\sigma}$ is the sum of all terms in the i th column of P .
- $p(x, d) = \text{Prob}\{X_d > x\}$, where X_d is a d -order χ^2 random variable is the p -value of score x in a χ^2 statistics.

We are interested in defining some scoring measure between two matrices P and Q .

Notice that in the case of Position Frequency Matrices and a uniform probabilistic background model (that is assumed in several studies), $\bar{P}_i = 1$.

2.1.1. Kullback-Leibler score

This distance, defined by Kullback in Ref. 13, is often used for matrix comparisons.^{14,15} The similarity between two motifs is defined by

$$S_{KL}(P, Q) = \frac{1}{2n} \sum_{i=1}^n \sum_{\sigma \in \mathcal{A}} p_\sigma \left[\frac{P_{i,\sigma}}{\bar{P}_i} \log \left(\frac{P_{i,\sigma} \bar{Q}_i}{Q_{i,\sigma} \bar{P}_i} \right) + \frac{Q_{i,\sigma}}{\bar{Q}_i} \log \left(\frac{Q_{i,\sigma} \bar{P}_i}{P_{i,\sigma} \bar{Q}_i} \right) \right].$$

2.1.2. Pearson correlation coefficient score

The Pearson correlation coefficient was introduced in motif comparison by Pietrokovski.¹⁶ The formula to compute similarity between two PWM is as follows:

$$S_{PCC}(P, Q) = \sum_{i=1}^n \frac{\sum_{\sigma \in \mathcal{A}} (P_{i,\sigma} - \bar{P}_i)(Q_{i,\sigma} - \bar{Q}_i)}{\sqrt{\sum_{\sigma \in \mathcal{A}} (P_{i,\sigma} - \bar{P}_i)^2 \sum_{\sigma \in \mathcal{A}} (Q_{i,\sigma} - \bar{Q}_i)^2}}.$$

Notice that the higher the score is, the more similar the matrices are. One of the major disadvantages of this scoring diagram, is that dissimilar columns do not have the same penalization weight in the final score.

2.1.3. Average log-likelihood ratio

This metric, introduced by Wang and Stormo,¹⁷ is a weight sum of two log-likelihood ratio. The measure takes the background into account.

$$S_{ALLR}(P, Q) = \sum_{i=1}^n \frac{\sum_{\sigma \in \mathcal{A}} P_{i,\sigma} \log \left(\frac{P_{i,\sigma}}{p_{\sigma} \hat{P}_i} \right) + Q_{i,\sigma} \log \left(\frac{Q_{i,\sigma}}{p_{\sigma} \hat{Q}_i} \right)}{\hat{P}_i + \hat{Q}_i}.$$

2.1.4. Pearson χ^2 column-by-column score

In Ref. 9, Schones *et al.* noticed that Pearson χ^2 test can be used in the context of motif comparison. This score consists in comparing the distributions of two aligned columns of the matrices. These columns follow a multinomial distribution and can be compared by using a χ^2 homogeneity test. We first define the column-by-column score between $P_i = (P_{i,\sigma})_{\sigma \in \mathcal{A}}$, the i th column of P and $Q_i = (Q_{i,\sigma})_{\sigma \in \mathcal{A}}$, the i th column of Q .

$$C(P_i, Q_i) = (\hat{P}_i + \hat{Q}_i) \left[\sum_{\sigma \in \mathcal{A}} \left(\frac{P_{i,\sigma}^2}{\hat{P}_i} + \frac{Q_{i,\sigma}^2}{\hat{Q}_i} \right) \frac{1}{P_{i,\sigma} + Q_{i,\sigma}} - 1 \right].$$

Under the null hypothesis that the two columns follow the same multinomial law, $C(P_i, Q_i)$ is comparable to a χ^2 distribution of order 3. Then, if one assumes that all the columns are independent, one can define a distance by taking the geometric mean of all p -values.

$$S_{\chi^2}(P, Q) = \left(\prod_{i=1}^n p(C(P_i, Q_i), 3) \right)^{1/n}.$$

Notice that when the marginal frequencies are small, Fisher-Irwin test, that involves multiple hypergeometric distribution, is more suitable.

Notice also that such a method implies several biases.

Bias 1. The same absolute error impacts very differently for small frequencies than for high frequencies. The following table illustrates this fact by providing the p -value obtained when comparing two pairs of columns that look similar pair-wise.

Pairs	A	C	G	T	p -value
1	1%	33%	33%	33%	0.395
	4%	32%	32%	32%	
2	22%	26%	26%	26%	0.03
	25%	25%	25%	25%	

Bias 2. Combining the p -values of different individual column tests may imply some strange artifacts. Indeed, the p -value of a comparison between two equal columns is 0 (notice that in a normal computer precision, this is also true for very close but not necessarily equal columns). As a consequence, as soon as an alignment between two motifs contains two equal columns, the score of a complete alignment is also zero, whatever the other columns are. Figure 1 illustrates these two arguments.

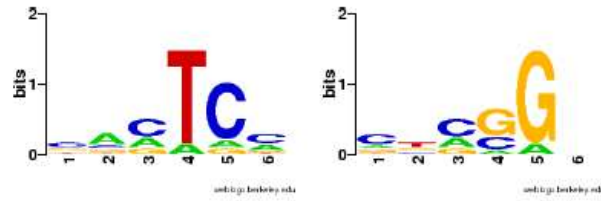


Fig. 1. Illustration of bias 2: column 3 is the same in both sequences. When comparing these two patterns, one obtains a p -value of 0.

2.1.5. Euclidean distance

The Euclidean distance score between two motifs is given by

$$S_{EUCL}(P, Q) = \sum_{i=1}^n \sum_{\sigma \in \mathcal{A}} (P_{i,\sigma} - Q_{i,\sigma})^2.$$

Comparing the Euclidean distances between two columns can appear to be simple and inefficient by a lack of normalization and the difficulty of providing a real significance of this score. Nevertheless, in Ref. 8, it is proved that an approximate significance can be computed by using permutations of the existing data. With this new information, the Euclidean distance score performs well in some real cases.

2.1.6. Global comparison

Here, we aim at defining a new score formula that decreases the scale effects of small frequencies (bias 1) and allows us to compare scores with different length. First, we noticed that small values for $P_{i,\sigma}$ can drastically affect the final score, notably because all the comparisons were done between ratios of values. We decide to suppress these non-representative values by taking solely the values over a given threshold ε (typically frequencies over 5%). By reducing the number of values by column, column-by-column comparison scores are no more appropriate. We thus design a χ^2 test suited to this case. First, let us consider the following contingency table defined by

$$T = \{(P_{i,\sigma}, Q_{i,\sigma}), i \in \{1, \dots, n\}, \sigma \in \mathcal{A} \text{ and } P_{i,\sigma} \geq \varepsilon \text{ or } Q_{i,\sigma} \geq \varepsilon\}.$$

Finally, one has to compare two distributions that may be assumed to be of multinomial type when the threshold is small.

First, one computes a χ^2 statistics related to T ,

$$K(T) = \frac{\widehat{E(T)}}{\widehat{O(T)}^2} \sum_{(E,O) \in T} \frac{O^2}{E} - 1, \quad (1)$$

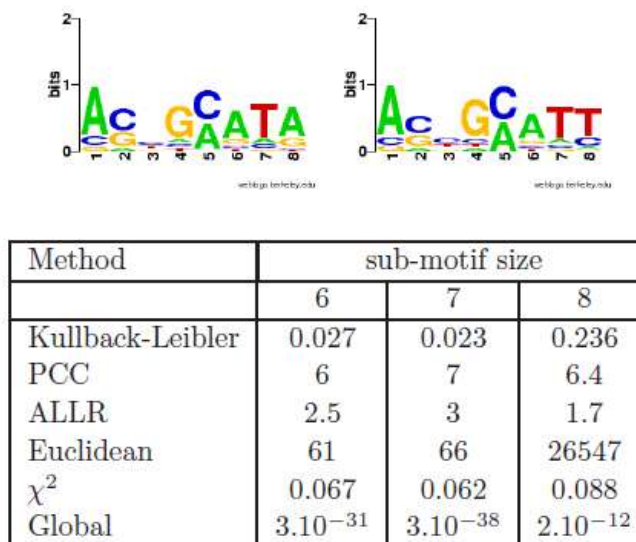


Fig. 2. A comparison of different distance scores for submotifs composed by the 6th, 7th and 8th first positions of two motifs. The motifs coincide on the 7th first positions.

where $\widehat{E}(T) = \sum_{(E,O) \in T} E$ and $\widehat{O}(T) = \sum_{(E,O) \in T} O$.

$$S_{GLOBAL}(P, Q) = p(K(T), |T| - 1),$$

where $|T|$ is the cardinality of set T . Notice at this point that this score is not symmetric. One matrix plays the role of E(xpected values) while the other plays the role of O(berved values). This score can be transformed into a symmetric one by summing the statistics $K(T)$ for both choices or by taking the minimal p -value between the two possibilities.

Figure 2 presents a comparison of different scoring methods. First notice that, except for Euclidean score, all scores provide coherent results. Nevertheless, there is a lack of distribution knowledge for column-by-column tests. Clearly, since they are sums or products of non-normalized variables, their variance, and thus their distribution, depend on the number of columns of the motif alignment. Their interpretation for comparing large sets of heterogenous motif is difficult, making the use of column-by-column tests inappropriate in this context. Furthermore, Euclidean distances always compute a better distance for shorter sub-motifs. It is thus difficult to compare motifs with different sizes. This can be partly solved by considering a p -value computation based on permutations of columns. That costs a lot of computation time. Finally, in this example, the longest coherent submotifs are those that possess the best global score (size 7).

2.2. Pattern clustering

Many clustering methods have been developed. A general principle of clustering is to maximize distance inter-cluster and minimize distance intra-cluster. In the

context of pattern clustering, the most used methods are unsupervised. Two principal methods can be described: hierarchical clustering and partition clustering. Hierarchical clustering is a deterministic and agglomerative method. The output of such method is a dendrogram, and the critical point of this method is the choice of a good threshold to partition it. Partitions clustering methods, like k -means, are stochastic methods, so they are sensitive to initial conditions and can converge to local minimum. However, they are easy to implement and are usable on very large datasets. Furthermore, choosing the parameters, like the number of centroids for k -means, remains difficult. In our work, we use a slightly different method for clustering the patterns. This latter method is particularly well adapted for comparing a set of patterns against a large dataset of patterns. Indeed, with this method, it is not necessary to know the distances between two patterns of the database. This saves a lot of computation time.

2.2.1. Hierarchical clustering

In hierarchical clustering, all motifs and clusters of motifs are compared pairwise using a dissimilarity measure. The most similar pair is grouped together in a new cluster. At the end of clustering, there is only one cluster, containing all the motifs. A dendrogram allows following gathering during clustering. The main interest of hierarchical clustering is that it allows seeing each step of clustering from dendrograms, and it produces a nice graphical output. However, partitioning the resulting dendrogram is a difficult task, requiring heuristics. Here, we apply an improved version of neighbor joining.¹⁸ Figure 3 depicts an example of dendrogram.

2.2.2. Partition clustering

Partition clustering stands for a set of methods to obtain a user defined and fixed number of clusters from the dataset. In the context of PWM clustering, Schones *et al.*⁹ have used a k -medoids method. This latter method aims at finding the best configuration (assign each matrix to one of the k chosen centers such that the squared error of the distances is minimal).

Choosing a good number of clusters is crucial, especially in the context of PWM clustering where obviously, several matrices are close and must be in the same cluster but a majority of other matrices are isolated and define their own cluster. In order to choose an appropriate number of clusters, a silhouette plotting is often used. Figure 4 shows an example of clusters obtained by a partition around medoids method.

2.2.3. Threshold clustering

This is surely one of the simplest clustering method. Nevertheless, we find this method very appropriate for comparing datasets of motifs to databases. Here, one constructs a graph whose vertices correspond to all pattern matrices and there exists

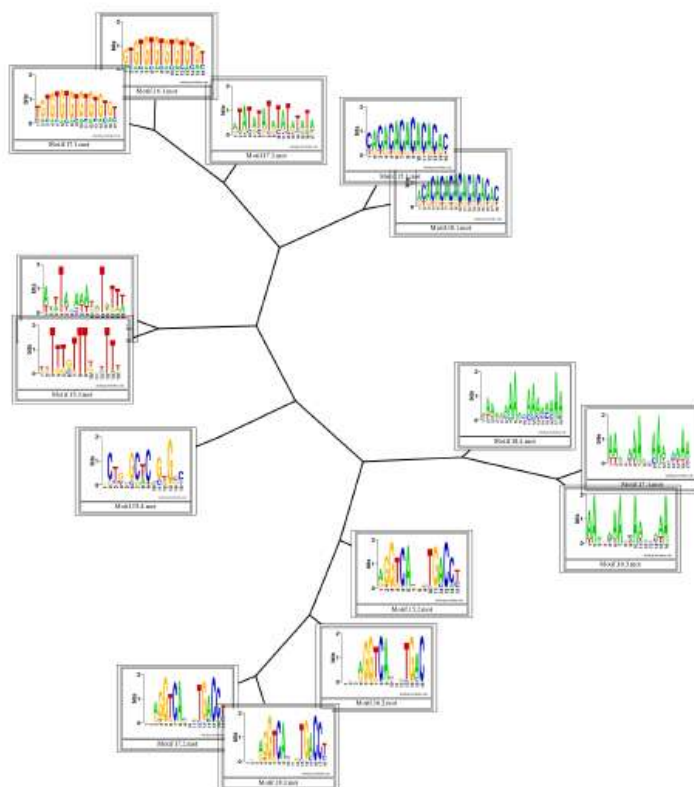


Fig. 3. An example of hierarchical clustering result.

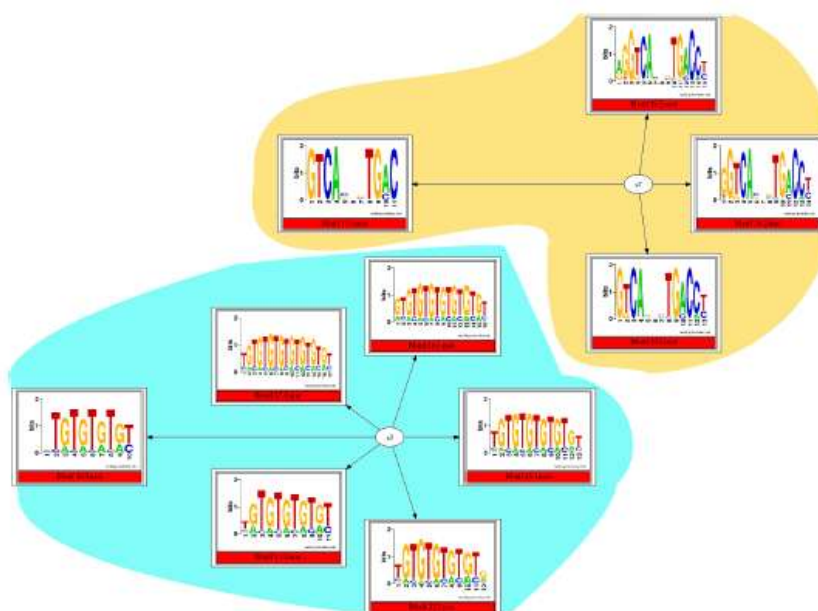


Fig. 4. An example of partition clustering result.

an edge between two vertices if and only if the distance between the associated matrices is below a given threshold. The clusters then correspond to the strongly connected components of the graph. Of course, this simple method will work with any kind of distances between matrices. In addition, this method ensures, by definition, that the closest matrices appear in the same cluster. Nevertheless, determining an appropriate threshold is crucial and it can be a difficult task depending on the scoring distance.

In this context, a good distance must ensure that there exists an important gap between matrices that can be considered as close and the other. It must ensure that matrices will be compared in a homogeneous scale which eliminates every distance (whose variances depend on the length of the aligned matrices) but the global score. In order to compute an appropriate threshold, one can consider as a characteristic parameter the number of edges in the final graph. It is null if the threshold is too restrictive and half of the square of the number of matrices if it is too permissive.

We have used this method in conjunction with the global score. We chose a threshold that guaranteed to use 90% of the edges with a significant distance. This allows defining some clusters similar to Fig. 5.

In addition, unlike hierarchical clustering and partition clustering, threshold clustering can deal with an incomplete distance matrix for constructing clusters. This can drastically reduce the comparison time when comparing a pattern set to a huge database. Indeed, distances between pairs of motifs from the database do

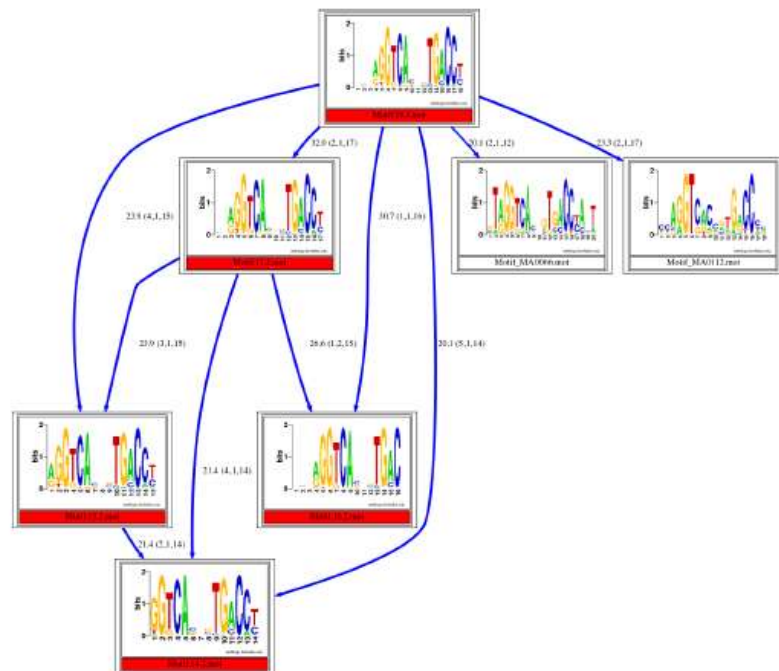


Fig. 5. An example of threshold cluster: Motifs are marked as black (motifs from the dataset) and white (motifs from the database).

not have to be computed in this case. It is sufficient to know the distance between every pairs of motifs in the set and between motifs from the set and every motifs from the database.

2.2.4. Graph motifs

After the clustering phase, it appears natural to deal with a new style of pattern consisting of a composition of several alternative but close pattern matrices. This increases the significance of motifs. In this work, we define the so-called motif graphs and show how any classical operations can be defined for such new motifs. Notice that our graph motif representation is quite close to Hidden Markov Models representations developed in grammatical inference (see Ref. 19 for a short review of HMM classical models as well as a presentation of HMMER3). There, one focuses on chains of highly conserved patterns described by PWM with the opposite objective of increasing the specificity of motifs regarding a set of sequences.

Definition 1. A graph motif is defined by a set $\mathcal{M}$ of pairs (P, o) , where P is a pattern matrix and o is an integer offset.

Here, the offset describes a kind of gap between the current position and the beginning position of the matrix. This can be summarized by Fig. 6.

Occurrences of a graph motif. Let us first recall that computing an appropriate score for pattern matrices is essential to determine if a position specific pattern matrix P of length n occurs in position k of a given sequence $\mathbb{S}$. One assigns a contribution $s_{i,\sigma}$ to each term $P_{i,\sigma}$ of the matrix P (in the case of PWM matrices, $s_{i,\sigma} = -\log_2 \frac{P_{i,\sigma}}{P_{i,p_\sigma}}$). The global score $s(P, \mathbb{S}, k)$ (i.e. the score of a n -length matrix P at position k in $\mathbb{S}$) is then defined as the sum

$$s(P, \mathbb{S}, k) = \sum_{i=1}^n s_{i, \mathbb{S}[k+i-1]},$$

where $\mathbb{S}[j]$ denotes the j th symbol of sequence $\mathbb{S}$.

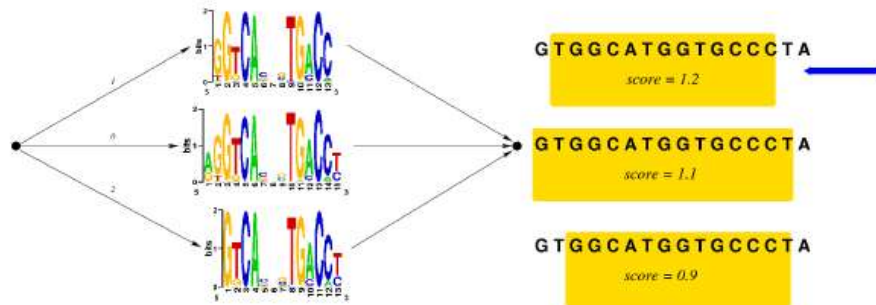


Fig. 6. An example of graph motif and an occurrence of such motif.

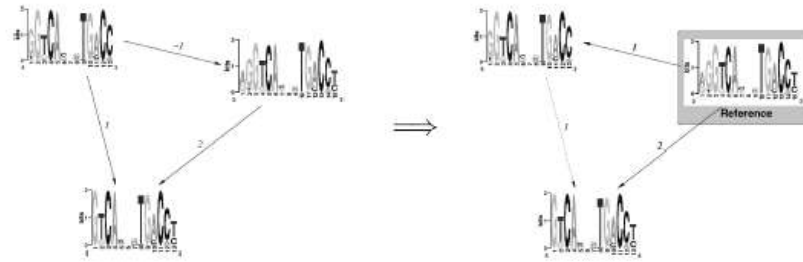


Fig. 7. Transformation from a cluster of pattern to a graph motif.

Such a definition translates in the context of graph motifs by taking the maximum of all possible scores.

Definition 2. The score $s(\mathcal{M}, \mathbb{S}, k)$ of a graph motif $\mathcal{M}$ at position k in $\mathbb{S}$ is defined by

$$s(\mathcal{M}, \mathbb{S}, k) = \max_{(P, o) \in \mathcal{M}} \frac{1}{\ell(P)} s(P, \mathbb{S}, k + o),$$

where $\ell(P)$ stands for the length of motif P .

From graph clustering to graph motifs. In the previous section, one constructs some clusters that group several matrices together. The edges of this cluster are endowed with some relative offset between two matrices. It is easy to construct a graph motif from such a structure by choosing a particular vertex of the cluster as a reference and by computing the offsets of all the remaining matrices relatively to this vertex.

Figure 7 shows an example of transformation from a clustering of pattern to a graph motif. First, one chooses a reference pattern. Then one computes the offset between all the other patterns to the reference.

Consensus for a graph motif. For several applications, it is important to deal with a single but representative pattern matrix. It is possible to construct such consensus matrices from a graph motif (that is supposed to represent the same motif with some minor differences). The consensus matrix consists in an alignment of a sub-window of each occurrence of a graph motif. An appropriate choice for defining a sub-window is to take the one defined by the longest pattern matrix of the graph motif. In this case, it is more convenient to compute all the offset values with reference to this central motif as depicted in Fig. 8.

3. Implementation Tricks

In order to have an exhaustive comparison of all the pattern matrices, one has to compute a distance between any pairs of the pattern set that is a maximum over all possible shifts between the two matrices and any sub-windows of the aligned matrices. Our principle is summarized in Fig. 9.

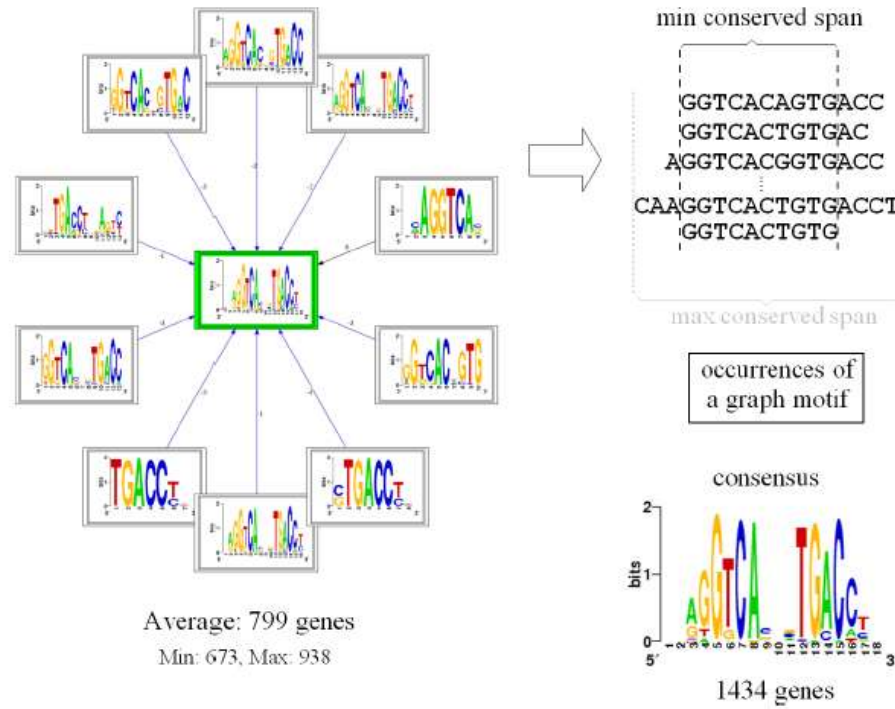
496 *S. Carat, R. Houlgatte & J. Bourdon*


Fig. 8. An example of consensus built from a graph motif.

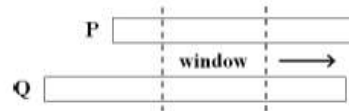


Fig. 9. Principle of the comparison scheme between two matrices. All possible shifts and all possible sub-windows are compared.

Computing the distance between two matrices depends linearly on the length of the matrices. This results in a naive complexity of order n^2k^3 , where n is the number of matrices and k is the length of the matrices.

We have used two implementation tricks that reduce drastically the execution time when comparing large sets of matrices. First, one uses a parallelization technique to reduce the complexity by a factor depending on the capacity of the server that executes our application. Next, one draws an iterative technique that reduces the number of comparisons by a factor k by using a recursive definition of our global distance. Finally, threshold clustering methods do not require to know the complete distance matrix. For instance, it is not necessary to compare the pattern matrices of the databases together. The final complexity of our clustering method is $\frac{n(n+N)k^2}{C}$, where N is the number of motifs in the database, n is the number of motifs in the pattern set, k is the maximal length of a pattern in the set, and C is the number of cores of the server.

3.1. Parallelization

First, notice that the computation of each distance between two matrices is independent. Such a task can be performed in parallel by some independent processes in a multi-core framework. Our implementation assigns the computation of all the distances of a matrix at one particular process. The distances are then written in the shared memory when the process finishes. This avoids the classical bottlenecks of parallelized application that only one process can access the shared memory at a time. Thus, having too many accesses to the shared memory increases the computation time.

By using this trick, we have reduced the computation time of all the distances by a factor seven on an eight-core computer.

3.2. Iterative computation of the global score

Notice now that for a fixed alignment of the two matrices, one computes the maximum global score for several different sizes of windows. We now show that these scores are dependent and they can be computed iteratively. This is mainly involved by the fact that $\widehat{E}(T)$ and $\widehat{O}(T)$ are sums as defined in formula (1) of Sec. 2.1.6. Indeed, let us consider the sum $R(T) = \sum_{(E,O) \in T} \frac{Q^2}{E}$. It is obvious that $R(T \cup (\alpha, \beta))$ can be calculated in constant time when $R(T)$ is known. It is also obvious that $K(T)$ can be computed in constant time when $R(T)$ is known. As a consequence, a dynamical programming algorithm that reports iteratively $K(T)$ can easily be designed. It then allows us to compute the distance between two aligned matrices for increasing sizes of windows in linear time.

4. Results and Discussion

4.1. Software availability

The complete method has been developed as a web application. The core of the implementation is done in PHP 5 language, making the application platform independent. A first online version is available on our servers.^a This version implements a hierarchical clustering that allows us to obtain an informative representation of the cluster set. A threshold clustering method is also implemented. It permits us to compare a set of motifs to JASPAR and TRANSFAC free public releases. All comparisons are based on our global distance measure for the moment. Nevertheless, we plan in the close future to add some other distance scores such as Kullback–Lebler and other used distances.

4.2. Comparing JASPAR and transfac databases

The main transcription factors binding site databases are JASPAR and Transfac. Our analysis is based on the latest free release of Transfac (Transfac Rel. 3.2) and

^a<http://cardioserve.nantes.inserm.fr/motifsComparator>.

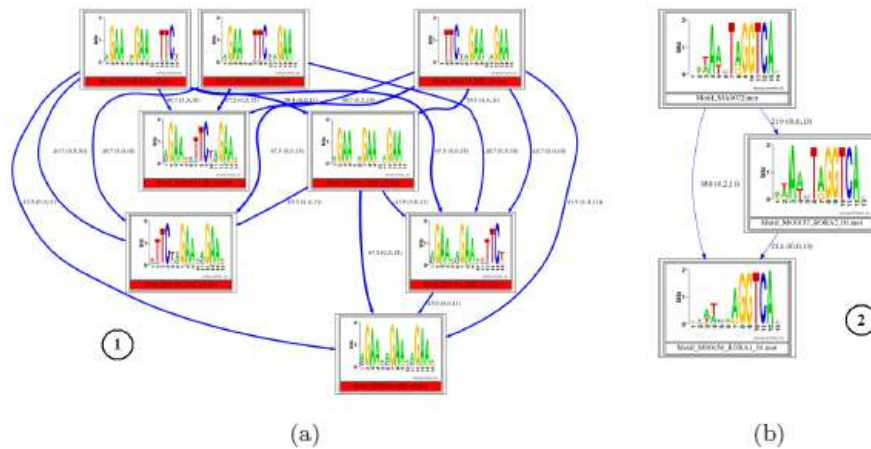


Fig. 10. An example of two clusters extracted from a comparison between JASPAR and Transfac databases.

JASPAR CORE 2008 dataset. Transfac release 3.2 includes 261 matrices, while JASPAR contains only 138 PWM. JASPAR is smaller than Transfac because of a manual curation that eliminates redundancies in JASPAR.

Comparison of all PWMs available in these two databases (Fig. 10) allows us to find not only redundancies between the two databases, but also redundant elements in Transfac. They appear in the same cluster as for instance Fig. 10(b). Here, one presents the transcription factor sub-family ROR (Retinoic Acid receptor related Orphan Receptor). We have found redundancies of RORA between JASPAR (MA0071) and Transfac (M00156, M00157). Besides, several transcription factor families are reconstructed, like the HSF family [Fig. 10(a)]. Indeed, we have reconstructed a graph motif with the entire HSF element available in Transfac. These results allow a biological validation of our global comparison method.

4.3. Comparing several experiments

One of the aims of this method is to facilitate ChIP-chip data analysis. Indeed, motifs discovered in input sequences are compared together to eliminate redundancies, with public databases to identify known transcription factor binding sites and with other ChIP-chip experiments to find some conserved patterns. We have used our method to analyze a study of Carroll *et al.*²⁰ on ChIP-chip data for ESR1 transcription factor. Elimination of redundancies is essential in this study because of the number of motifs to analyze. Furthermore, using a consensus pattern allows the significant increase of the number of sequences concerned by the motif (Fig. 8). Comparison of motif with public databases allows identification of corresponding transcription factor binding sites. In the case of ChIP-chip ESR1, we have compared all the motifs discovered by MDmodule^b to JASPAR and Transfac

^bMDmodule is the successor of MDscan tool. It is a part of MotifRegressor pattern discovery package.⁶

databases (Fig. 5). Several motifs are found similar to the best input motif. They correspond to PPARgamma and T3R. PPARgamma, T3R and ESR1 are part of the nuclear receptor superfamily, and more specifically to nuclear hormone receptors family (NHR). NHR dimers bind to regulatory sequences composed of two half-sites, where half-sites have the consensus sequence AGGTCA as showed in Ref. 10. So gathering of such motif is expected.

4.4. Building motif networks

Comparison of motifs found in several ChIP-chip experiments allows us to construct regulatory network, where known motifs are represented by transcription factor name and unknown by consensus motif (Fig. 11). We have identified various groups of genes, represented in circles containing their number. They are co-regulated by a subset of motifs and are associated to particular functions. This is proven by mapping a significant functional annotation (obtained with GOMiner tool²¹) of each gene cluster. Comparison of several ChIP-chip experiments allows us to validate interaction between transcription factors and a set of deregulated genes and to refine network. Significant functional annotations make sense on each genes cluster and can assist in defining putative therapeutic target.

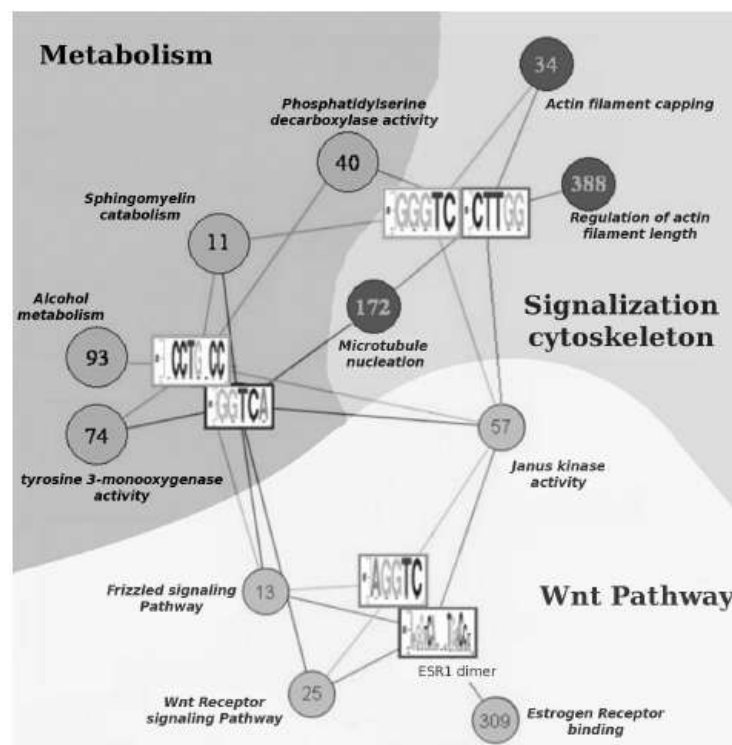


Fig. 11. An example of motif networks for ChIP-chip experiments on ESR1.

5. Conclusion

We have presented a complete method to study some sets of motifs described by pattern matrices like PWM, PSSM or PFM. Eliminating redundancies requires to compare all motifs together, but also with reverse complement of all motifs and to databases. It is thus necessary to use some tricky implementation techniques in order to deal with large sets of motifs, like those arising in the study of ChIP-chip data for instance. Finally, as a complementary result, by comparing a pattern matrix to itself, to its shifts and to its reverse complement, one can determine if the motif is a palindrome, a common characteristic of TFBS, or a periodic motif that usually represents artifacts of the discovery step.

Acknowledgments

The authors deeply appreciate anonymous referees' comments that helped to significantly improve the quality of the paper. We would also like to thank Mireille Régnier for many fruitful discussions around this work. Finally, this work is partially supported by BIL regional research program.

References

1. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A, Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*, *Nucleic Acids Res* **10**:2997–3011, 1982.
2. Staden R, Computer methods to locate signals in nucleic acid sequences, *Nucleic Acids Res* **12**:505–519, 1984.
3. Bailey TL, Elkan C, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc Int Conf Intell Syst Mol Biol* **2**:28–36, 1994.
4. Hughes JD, Estep PW, Tavazoie S, Church GM, Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*, *J Mol Biol* **296**:1205–1214, 2000.
5. Liu XS, Brutlag DL, Liu JS, An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments, *Nat Biotechnol* **20**:835–839, 2002.
6. Conlon EM, Liu XS, Lieb JD, Liu JS, Integrating regulatory motif discovery and genome-wide expression analysis, *Proc Natl Acad Sci USA* **100**:3339–3344, 2003.
7. MacIsaac KD, Fraenkel E, Practical strategies for discovering regulatory DNA sequence motifs, *PLoS Comput Biol* **2**:e36, 2006.
8. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS, Quantifying similarity between motifs, *Genome Biol* **8**:R24, 2007.
9. Schones DE, Sumazin P, Zhang MQ, Similarity of position frequency matrices for transcription factor binding sites, *Bioinformatics* **21**:307–313, 2005.
10. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B, JASPAR: An open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res* **32**:D91–94, 2004.
11. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prss M, Reuter I, Schacherer F, TRANSFAC: An integrated system for gene expression regulation, *Nucleic Acids Res* **28**:316–319, 2000.

12. Liu J, Neuwald A, Lawrence C, Bayesian models for multiple local sequence alignment and gibbs sampling strategies, *Journal of the American Statistical Association*, pp. 90–432, 1995.
13. Kullback S, *Information Theory and Statistics*, Wiley, New York, 1959.
14. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B, Computational detection of cis-regulatory modules, *Bioinformatics* 19:5–14, 2003.
15. Roepcke S, Grossmann S, Rahmann S, Vingron M, T-Reg Comparator: An analysis tool for the comparison of position weight matrices, *Nucleic Acids Res* 33:W438–441, 2005.
16. Pietrokovski S, Searching databases of conserved sequence regions by aligning protein multiple-alignments, *Nucleic Acids Res* 24:3836–3845, 1996.
17. Wang T, Stormo GD, Combining phylogenetic data with co-regulated genes to identify regulatory motifs, *Bioinformatics* 19:2369–2380, 2003.
18. Sheneman L, Evans J, Foster JA, Clearcut: A fast implementation of relaxed neighbor joining, *Bioinformatics* 22:2823–2824, 2006.
19. Eddy SR, A probabilistic model of local sequence alignment that simplifies statistical significance estimation, *PLoS Comput Biol* 4:e1000069, 2008.
20. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoutte J, Brodsky AS *et al.*, Genome-wide analysis of estrogen receptor binding sites, *Nat. Genet* 38:1289–1297, 2006.
21. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M *et al.*, GoMiner: A resource for biological interpretation of genomic and proteomic data, *Genome Biol* 4:R28, 2003.



Solenne Carat's research topics cover genomic data analysis and bioinformatics. In 2003, she started bioinformatics studies and graduated from Auvergne University. Then, she moved to Paris VII University where she obtained a M.Sc degree. Finally, she integrated INSERM U915 Bioinformatics team and obtained a Ph.D. grant. Her Ph.D. thesis is co-advised by researchers in biology from INSERM U915 and by computer scientists from the computer science unit of Nantes University. She currently focuses on the integration of transcriptomic and ChIP-chip data in order to construct regulatory networks.



Rémi Houlgatte is the team leader of INSERM U915 “Genomics and biotherapies” project, Nantes University. He received his Ph.D. in Biology from Paris VI university in 1988 and obtained a junior researcher position in 1990, working with INSERM U298, Angers University. In 1991, he moved to CNRS UPR420 Villejuif where he headed a bioinformatics team. From 1998 to 2005, he worked in Marseille University with CIML INSERM U36 and INSERM ERM206 where he was successively co-director of TAGC team and director of Oncogenomic team. Finally, he obtained an INSERM senior researcher position in Nantes University in 2004. His research

areas cover systems biology, integrative genomics, bioinformatics and therapeutic target discoveries.



Jérémie Bourdon has conducted a broad area of research from mathematics, computer science, and bioinformatics. He graduated from the Caen University (Computer Science), and conducted dynamical systems researches (theoretical) at Caen university (Ph.D. obtained in 2002). In 2003, he obtained an assistant professorship position in the computer science unit of Nantes University. There, he changed his focus to probabilities (theoretical and applications) and bioinformatics (design of algorithms dedicated to pattern discovery). Till now, he began a strong collaboration with INSERM U915 and deeply studied genomic data. In 2008, he joined the INRIA project-team SYMBIOSE (Rennes, France) and began researches on systems biology and integrative genomics (collaboration with INSERM U915).

L'objectif à plus long terme est de compléter cet outil, afin de proposer un processus global qui permet, à partir de séquences fournies par l'utilisateur, d'effectuer le masquage des séquences, la découverte des motifs ainsi que leurs comparaisons et leurs regroupement, et enfin de construire les motifs consensus. Toutes ces étapes sont actuellement disponibles en interne dans notre équipe, mais nécessitent encore d'être adaptées pour un utilisateur externe. En effet, l'objectif final de l'outil est d'être le plus modulable possible, afin de laisser l'utilisateur libre de choisir les étapes qu'il souhaite effectuer lui-même, et celles qu'il souhaite réaliser avec motifsComparator. Ceci nécessite une bonne adaptabilité des programmes à différents types d'entrées.

II- Etude du coactivateur transcriptionnel PRC

1) Etude du transcriptome

La première étude menée dans l'équipe pour comprendre les mécanismes d'action de PRC dans la coordination de la prolifération mitochondriale et cellulaire a consisté en une étude de transcriptome. Cette étude, visant à étudier la réponse au siRNA PRC de la lignée cellulaire XTC.UC1 après induction par le sérum, a été initiée dans l'équipe par un précédent étudiant en thèse. Elle a donné lieu à une première publication des résultats [Raharijaona *et al.*, 2009]. J'ai repris ces travaux, et effectué de nouvelles analyses afin de pouvoir étudier conjointement les résultats de transcriptome et de ChIP-chip.

La classification hiérarchique des données de transcriptome est présentée dans la Figure 24. Les échantillons ont été triés par ordre chronologique d'induction au sérum (0, 12, 24 et 48h après l'ajout de sérum de veau à 10%), ainsi que par classe d'échantillons : cellules transfectées par un siRNA PRC ou cellules transfectées par un siRNA scramble (contrôle). La classification hiérarchique des gènes a permis d'identifier 8 clusters, avec des fonctions biologiques particulières. Cependant, l'effet prédominant observé dans ces clusters correspond à l'induction de l'expression des gènes par le sérum, et non à celui du siRNA PRC. En effet, dans le clustering global, il n'y a pas de différence apparente entre les échantillons transfectés avec le siRNA PRC ou le siRNA scramble. En revanche, une analyse comparative plus fine de chaque cluster, en éliminant l'effet principal du sérum, permet de mettre en évidence la différence d'expression des gènes de chaque cluster sous l'effet du siRNA PRC. L'élimination de l'effet du sérum est réalisée en recentrant toutes les données à chaque temps en fonction des données du siRNA scramble.

L'étude de chaque cluster indépendamment permet d'observer des effets inhibiteurs (gènes sur-exprimés en présence du siRNA PRC, donc rouge dans le clustering) et des effets activateurs de PRC (gènes vert dans le clustering sous-exprimés avec le siRNA PRC). En effet, le siRNA PRC étant un inhibiteur de PRC, une surexpression d'un gène sous l'effet du siRNA PRC traduit le fait que l'absence de PRC entraîne une augmentation de l'expression du gène, donc qu'à l'état sauvage PRC inhibe ce gène. On observe également des effets très précoces de PRC, visibles seulement 12h après la reprise du cycle cellulaire, et des effets plus tardifs, visibles seulement à 24 voire à 48h après l'induction du cycle cellulaire. Certains effets tardifs de PRC peuvent être expliqués par un effet indirect de PRC sur l'expression de certains gènes.

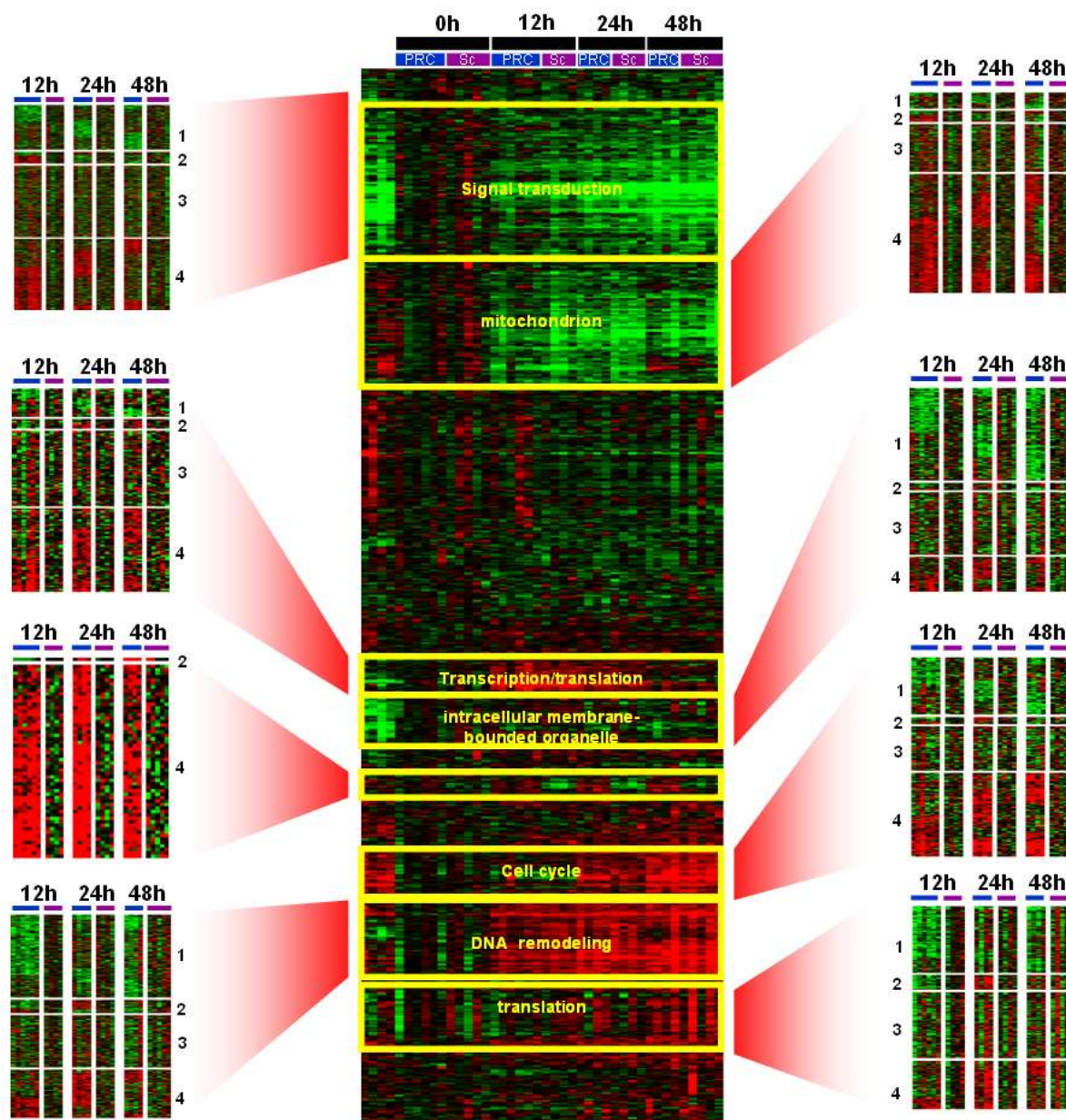


Figure 24: clustering hiérarchique des gènes sensibles au siRNA PRC

L'image centrale du clustering, dont les échantillons sont ordonnés chronologiquement et par type (en bleu les cellules transfectées par le siRNA PRC, en violet les cellules transfectées par le siRNA scramble) permet d'identifier 8 clusters (en jaune) spécifiques dont les annotations fonctionnelles significatives sont précisées. L'effet majoritaire observé correspond à l'induction des gènes par le sérum. Chaque cluster a été analysé spécifiquement (zoom sur les côtés) pour observer les effets du siRNA PRC en fonction de la cinétique de traitement.

L'annotation fonctionnelle des 8 clusters de gènes a permis de mettre en évidence des fonctions particulières et notamment pour les clusters 2 et 6 (de haut en bas) correspondant à des fonctions mitochondriales et au cycle cellulaire. Ces deux clusters nous ont particulièrement intéressés pour la suite de notre étude visant à comprendre le rôle de PRC dans la coordination de la prolifération mitochondriale et cellulaire.

Nous avons recherché au sein de ces deux clusters s'il était possible d'identifier des facteurs de transcription, ou encore des motifs de facteurs de transcription enrichis dans les promoteurs des gènes de ces clusters. En effet, la régulation corrélée d'un groupe de gènes sous-tend une régulation commune par un ou plusieurs facteurs de transcription. Etant donné que PRC est un coactivateur transcriptionnel, nous avons cherché à identifier les facteurs pouvant interagir avec lui pour la régulation de ces gènes. Cette recherche a permis d'identifier cinq facteurs de transcription, également connus dans la littérature pour intervenir dans la régulation de gènes mitochondriaux et du cycle cellulaire. Il s'agit des facteurs NRF1, GABP, $ERR\alpha$, CREB et YY1. Les sites de fixation de ces facteurs de transcription sont présentés dans la Figure 25. Les sites de fixation les mieux décrits pour ces facteurs de transcription sont ceux issus de Genomatix. Leur fréquence a été retrouvée significativement enrichie dans les deux clusters de gènes d'intérêt.

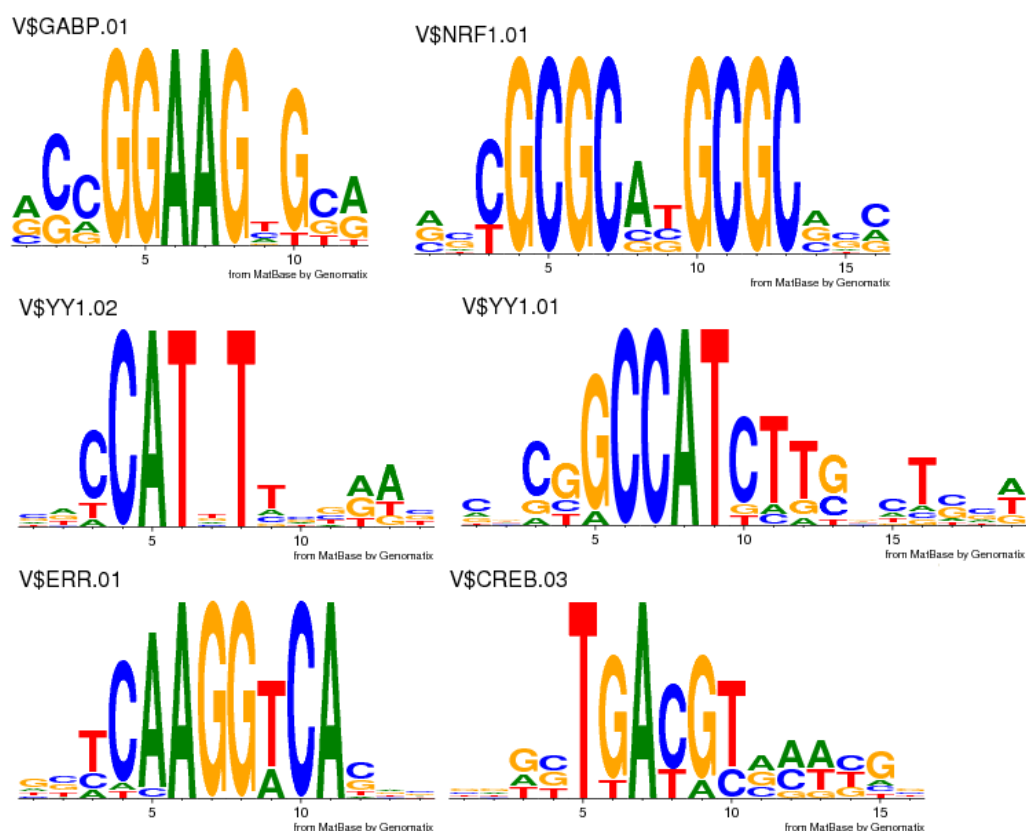


Figure 25: motifs des 5 facteurs de transcription étudiés en ChIP

Motifs des sites de fixation des facteurs de transcription issus de Genomatix® correspondant de haut en bas et de gauche à droite à : GABP, NRF1, YY1, YY1, $ERR\alpha$ et CREB.

Au vu de ces résultats, nous avons décidé d'étudier l'ensemble de ces facteurs de transcription par ChIP-chip, en plus de l'étude de PRC.

2) Etude des ChIP-chip

Afin de savoir quel type de lame choisir pour réaliser nos expériences de ChIP-chip, nous avons utilisé la puce hybride d'environ 44000 sondes que nous avons conçue au laboratoire. Cette lame contient des sondes localisées dans les promoteurs proximaux des gènes, d'autres dans les îlots CpG et d'autres enfin qui n'ont pas de localisation particulière dans le génome. La lame contient, en plus, des sondes contrôles d'Agilent. Nous avons hybridé une lame par facteur de transcription, afin de pouvoir adapter notre stratégie aux particularités de reconnaissance des différents facteurs de transcription. Les résultats des différentes hybridations ont été relativement similaires, ce qui nous a permis d'utiliser le même type de puce pour toutes les analyses ultérieures. Les résultats présentés dans la Figure 26 correspondent au facteur de transcription GABP.

a. Choix du type de puces

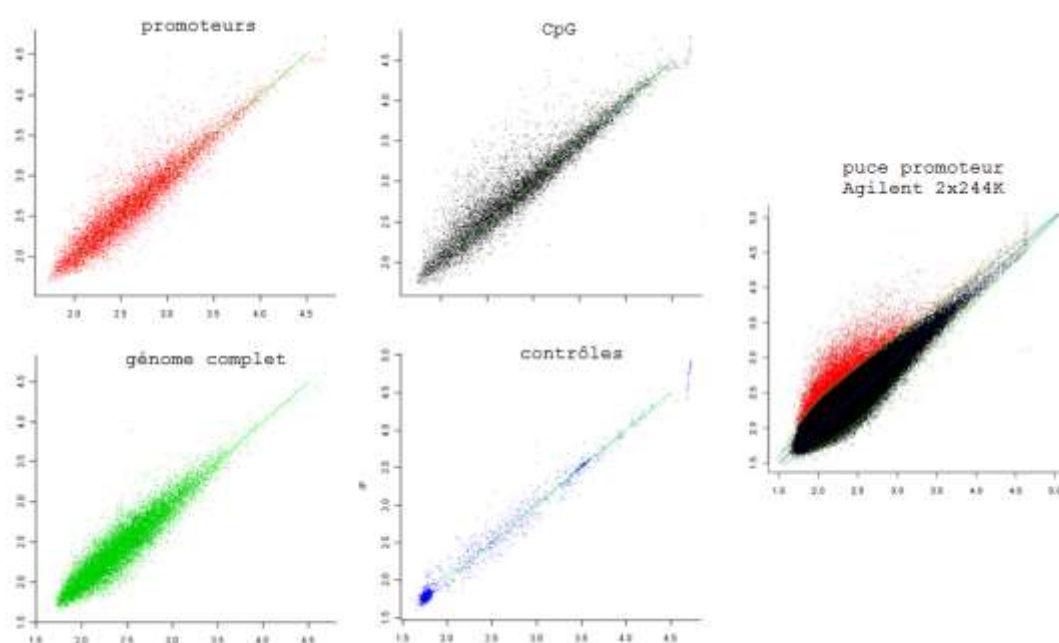


Figure 26: Localisation des sites de fixation de GABP

Résultats de l'hybridation du ChIP GABP sur la puce hybride développée sur notre plateforme (quatre images de gauche) et sur la puce à promoteur Agilent 2x244k (image de droite). Pour la puce hybride, les sondes rouges sont localisées dans les promoteurs proximaux, les noires dans les îlots CpG, les vertes dans le génome complet et les bleues sont des contrôles. Les gènes positifs en ChIP sont identifiés sur la puce de promoteurs Agilent.

Tout d'abord, les résultats présentés dans le graphe contrôle de la puce hybride, en bleu, montrent bien qu'il n'y a pas de différence entre les deux échantillons en ce qui concerne les hybridations non spécifiques et le bruit de fond. Pour les trois autres graphes

de la puce hybride nous observons un enrichissement, c'est-à-dire un signal en IP plus important qu'en input, concernant les sondes localisées dans les promoteurs proximaux et dans les îlots CpG, ce qui n'est pas le cas des sondes localisées dans tout le génome. Ces résultats nous ont permis d'éliminer de notre liste de choix les puces pangénomiques, étant donné que les cibles privilégiées par nos facteurs d'intérêt semblent majoritairement être localisées à proximité des gènes. Ceci nous permet également de réduire considérablement les coûts des expériences, étant donné que nous avons opté pour une puce qui ne couvre qu'une portion du génome. Nous avons donc choisi d'utiliser pour la suite des analyses les puces à promoteur Agilent 2x244k, contenant environ 490 000 sondes localisées dans les promoteurs proximaux sur une distance de 10kb en moyenne, chevauchant le départ de transcription (-7,5 kb à + 2,5 kb).

b. Sélection des sondes positives en ChIP

Il existe de nombreuses méthodes pour sélectionner les sondes positives en ChIP. Cependant, ces méthodes sont adaptées aux puces qui contiennent nécessairement plusieurs sondes positives consécutives (une approximation par une loi normale du signal positif facilite la sélection des sondes et l'élimination du bruit de fond). Ces méthodes ne s'appliquent pas au type de puces que nous utilisons car les sondes sont trop éloignées et il est rare que plusieurs sondes positives se suivent. J'ai donc dû développer une nouvelle méthode adaptée à ce cas.

Celle-ci permet de sélectionner les sondes positives en ChIP en se basant sur l'hypothèse qu'il n'existe pas de sondes négatives en ChIP (enrichies dans l'input). Ceci permet de définir un seuil de bruit de fond dans la partie inférieure de la courbe, puis de reporter ce seuil dans la partie supérieure afin de sélectionner ensuite les sondes positives en ChIP.

J'ai utilisé cette méthode pour analyser les six facteurs de transcription que nous avons choisi d'étudier. Les résultats, pour un réplicat de chaque facteur, sont présentés dans la Figure 27. Les sondes positives en ChIP sont représentées en rouge sur chacun des graphes. En fonction du facteur de transcription étudié, le nombre de sondes sélectionnées est très variable. Les facteurs qui ciblent le plus de sondes sont les facteurs CREB, et NRF1. Ces résultats concernant le facteur de transcription CREB ne sont pas surprenants, étant donné qu'il s'agit d'un facteur de transcription ubiquitaire, connu pour réguler l'expression d'un grand nombre de gènes. En revanche, ce résultat est plus étonnant pour

le facteur de transcription NRF1, dont le nombre de gènes cibles connu est plus réduit dans les modèles cellulaires précédemment explorés dans la littérature.

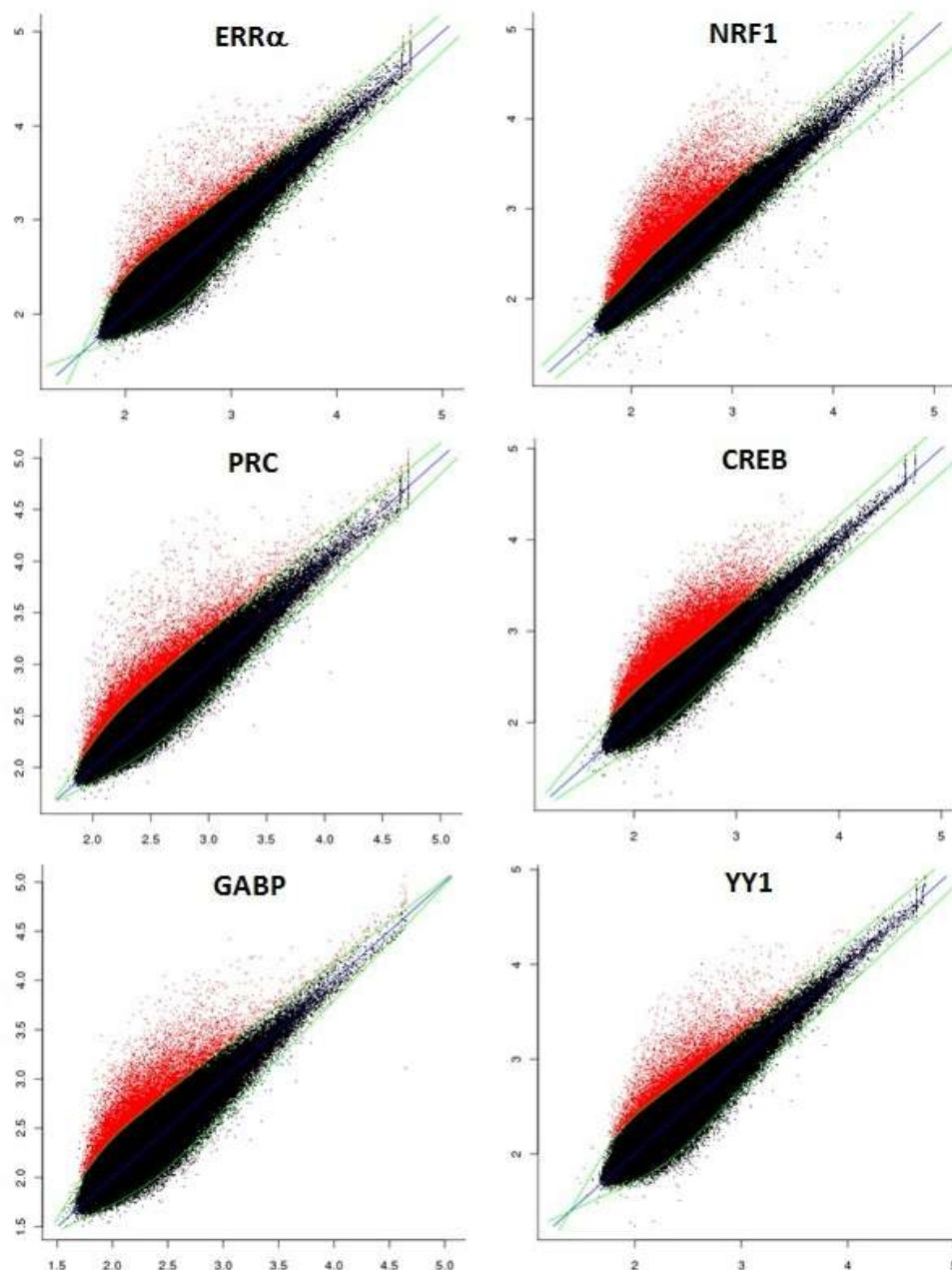


Figure 27: sélection des sondes positives en ChIP des 6 facteurs

Les données en abscisse correspondent au log de l'intensité normalisée de l'input, et en ordonnée du ChIP pour chacun des facteurs étudié. Les sondes en rouge correspondent aux sondes positives en ChIP.

c. Reproductibilité des résultats

Afin de s'assurer d'une bonne qualité des résultats, nous avons réalisé deux réplicats par facteur de transcription. Pour vérifier la reproductibilité des expériences, nous avons calculé la corrélation des p-valeurs de chaque sonde pour les deux réplicats.

Etant donné que la sonication coupe l'ADN à des positions aléatoires, il est possible que deux fragments d'ADN contenant un même TFBS se fixent sur deux sondes différentes (Figure 28), correspondant pourtant au même gène. Nous avons donc choisi de calculer également la corrélation des p-valeurs entre la meilleure sonde de chaque gène de l'un et l'autre des répliquats.

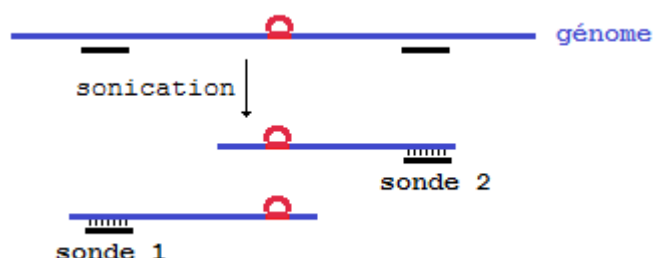


Figure 28: hybridations multiples dues à la coupure aléatoire de l'ADN

En fonction du lieu de coupure de l'ADN par sonication, deux fragments d'ADN contenant le même site de fixation (en rouge), et fixant donc le même facteur peuvent s'hybrider sur deux sondes différentes.

Les résultats de corrélation pour toutes les sondes des deux répliquats du facteur de transcription CREB ainsi que pour la meilleure sonde de chaque gène du même facteur de transcription sont exposés dans la Figure 29.

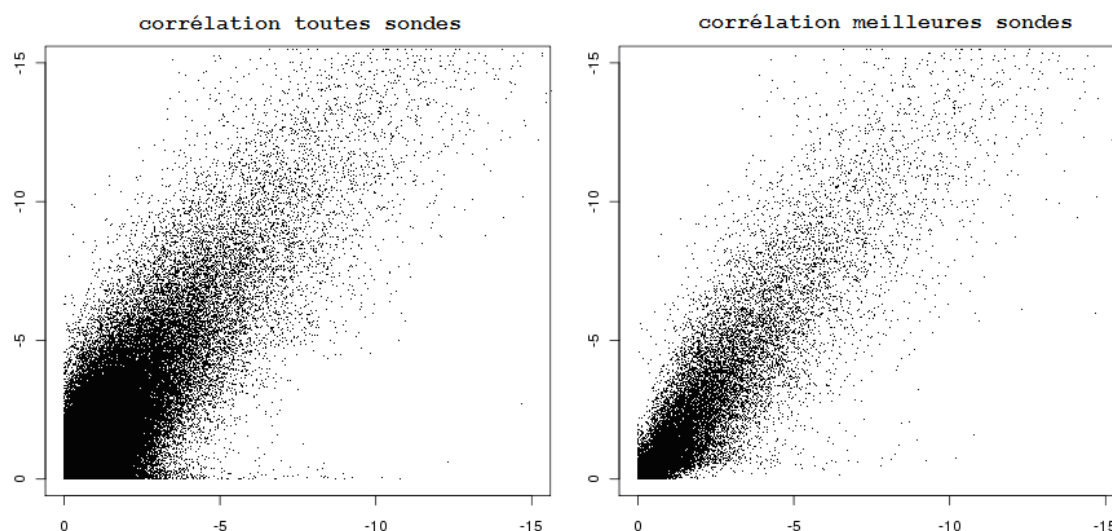


Figure 29: corrélation des répliquats du facteur de transcription CREB

Les données en abscisse correspondent aux log de base 10 des p-Valeurs du réplikat 1 et celles en ordonnée du réplikat 2 de CREB. Le graphe de gauche correspond aux p-Valeurs en log de base 10 de l'ensemble des sondes de la puce, alors que le graphe de droite correspond à la meilleure sonde de chaque gène pour chacun des répliquats.

Pour la corrélation de toutes les sondes des deux répliquats de CREB, le coefficient de corrélation de Pearson est égal à 0.52, avec une p-valeur $< 2.2 \times 10^{-16}$. Concernant la

corrélation calculée pour la meilleure sonde de chaque gène pour chaque réplicat, le coefficient de corrélation est de 0.78 avec la même p-valeur. Nous obtenons donc une bonne corrélation, qui est très significative au regard du nombre de sondes contenues dans chaque vecteur. Comme nous nous y attendions, la corrélation est fortement améliorée en sélectionnant la meilleure sonde de chaque gène.

Les corrélations des réplicats des six facteurs étudiés basées sur la meilleure sonde de chaque gène sont présentées dans la Figure 30 et le Tableau 3.

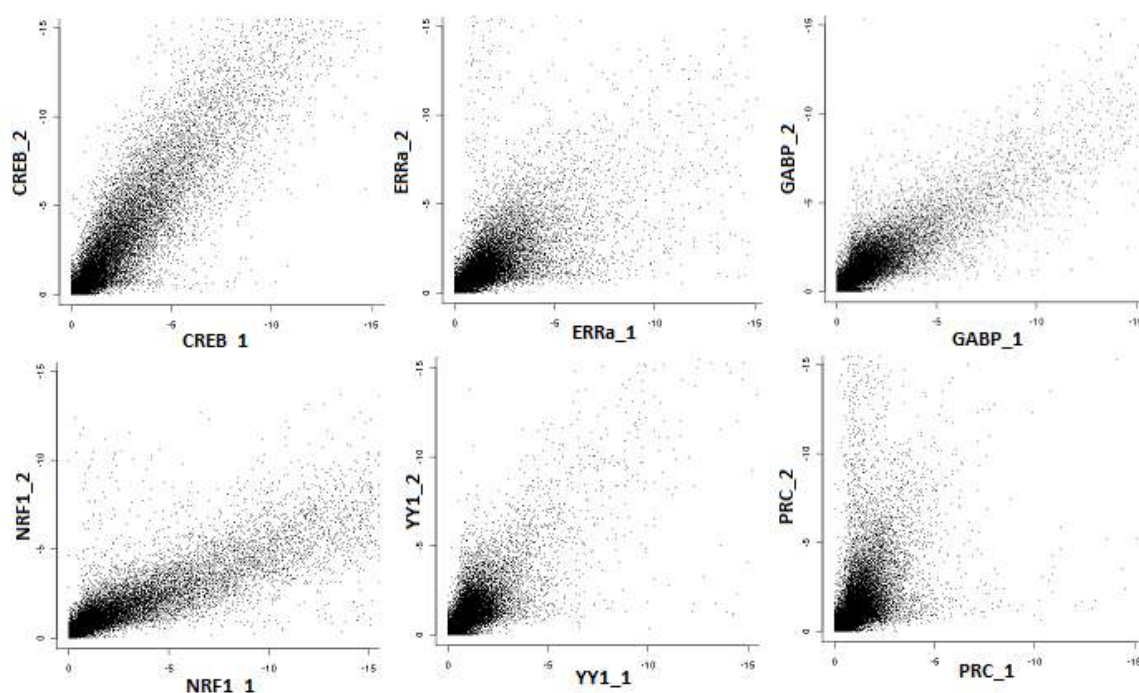


Figure 30: corrélation des réplicats des 6 facteurs étudiés en ChIP

Corrélations des réplicats de (de gauche à droite et de bas en haut) CREB, $ERR\alpha$, GABP, NRF1, YY1 et PRC, pour la meilleure sonde de chaque gène en log de base 10.

Tableau 3: corrélation des réplicats des 6 facteurs de transcription

facteur	CREB	$ERR\alpha$	GABP	NRF1	YY1	PRC
Coefficient de Pearson	0,78	0,75	0,76	0,84	0,71	0,7
t	214	192	197	262	169	166
p-valeur	$< 2,2 \times 10^{-16}$					

Globalement, il y a une très bonne reproductibilité entre les deux réplicats de chaque facteur de transcription. Ceci traduit une bonne reproductibilité des différentes étapes techniques, et donc une bonne fiabilité des expériences de ChIP-chip. On peut noter que le facteur le mieux corrélé est le facteur NRF1, avec un coefficient de corrélation supérieur à 0.8. La corrélation entre les deux réplicats de PRC un peu moins

bonne peut être expliquée par le fait que nous utilisons un anticorps maison pour le ChIP, qui est peut être un peu moins efficace que les anticorps commerciaux utilisés pour les autres facteurs. De plus, PRC étant un coactivateur, il n'a pas de domaine de liaison à l'ADN. L'étape de liaison des protéines à l'ADN lors du ChIP est donc plus importante car nous devons fixer tout le complexe. Après différents tests de crosslink, notamment à base de DTBP, nous avons finalement choisi de garder la même stratégie que pour les autres facteurs de transcription. En effet, même si les liaisons covalentes étaient plus fortes avec le DTBP, elles étaient également plus difficiles à rompre lors du reverse crosslink, et l'enrichissement final était moins bon.

d. Sélection des gènes positifs en ChIP

Etant donné que les résultats sont bien reproductibles pour tous les facteurs de transcription, j'ai sélectionné les gènes positifs en ChIP sur la base de la moyenne géométrique des p-valeurs des sondes des deux réplicats. Un gène est considéré comme positif en ChIP si au moins une sonde correspondant au promoteur de ce gène a une p-valeur moyenne inférieure à 10^{-3} . Le Tableau 4 montre les effectifs des gènes positifs pour chacun des facteurs de transcription étudié.

Tableau 4: effectif des gènes positifs pour les différents facteurs

FT	Nb Gènes sélectionnés 10^{-3}
PRCe	1 252
ERRa	2 348
NRF1	5 679
GABP	3 091
CREB	6 377
YY1	1 079

Comme pour les sondes, les facteurs de transcription ciblant le plus grand nombre de gènes sont les facteurs NRF1, avec 5 679 gènes positifs, et CREB concernant 6 377 gènes. Les autres facteurs de transcription ciblent en moyenne trois fois moins de gènes.

En étudiant les résultats des six ChIP pour les six facteurs de transcription, on peut construire un réseau d'interaction entre les différents facteurs de transcription (Figure 31). Ce réseau montre le rôle central que semble jouer CREB, en interaction avec PRC, pour moduler l'expression des différents facteurs de transcription.

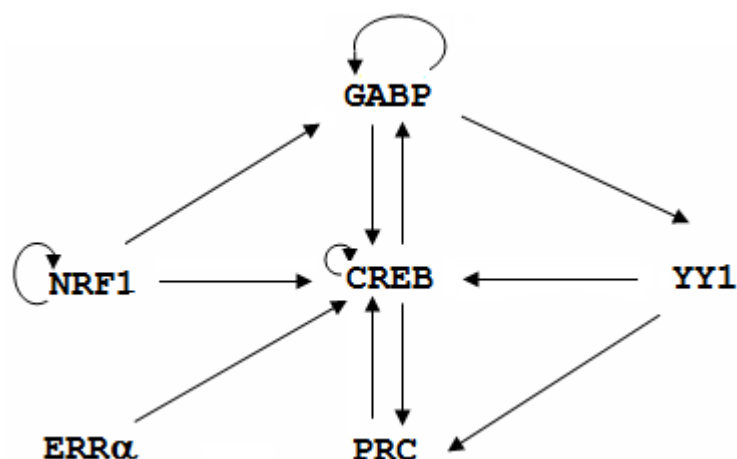


Figure 31: réseaux de régulation des 6 facteurs de transcription

Le réseau contient les 6 facteurs de transcription qui ont été étudiés en ChIP. Le facteur de transcription ciblé par une flèche a été retrouvé positif en ChIP pour le facteur de transcription à la source de la flèche.

Le Tableau 5 indique la p-valeur moyenne de chaque facteur de transcription pour chacun des jeux de gènes positifs en ChIP. Les cases rouges correspondent aux p-valeurs significatives, utilisées pour la construction du réseau précédent. Les cases oranges correspondent à des p-valeurs légèrement sous le seuil de significativité choisi, qu'il pourrait être quand même intéressant de considérer.

Tableau 5: moyenne géométrique des p-valeurs des 6 facteurs de transcription

	CREB	ERRa	GABP	NRF1	PRC	YY1
gene pos CREB	4E-07	0.076	1E-04	0.014	9E-07	0.0046
gene pos ERRa	1E-05	0.042	0.003	0.025	0.036	0.0405
gene pos GABP	2E-04	0.002	7E-07	0.059	0.061	5E-05
gene pos NRF1	0	0.014	0	1E-06	0.007	0.0085
gene pos PRC	6E-06	0.026	0.007	0.012	0.025	0.0439
gene pos YY1	9E-04	0.004	0.011	0.002	4E-05	0.0775

e. Annotation fonctionnelle des gènes positifs en ChIP

L'annotation fonctionnelle des gènes positifs pour chaque facteur de transcription a pour objectif d'identifier des fonctions spécifiques régulées par chacun des facteurs, et également de valider les résultats de ChIP-chip. En effet, trouver des annotations fonctionnelles statistiquement significatives dans un groupe de gènes particulier indique que ce groupe n'est pas construit de manière aléatoire, et qu'il a donc un sens biologique. De plus, ces résultats peuvent être comparés à la littérature afin de vérifier que les fonctions trouvées correspondent bien à des fonctions connues pour être régulées par les

facteurs considérés. Les annotations fonctionnelles présentées dans le Tableau 6 ont été sélectionnées et synthétisées à partir des critères suivants : p-valeur $< 10^{-3}$, enrichissement > 1.5 et au moins 5 gènes porteurs de l'annotation. La liste complète des annotations est fournie en Annexe.

Tableau 6: annotation fonctionnelle des gènes positifs en ChIP

NRF1		pVal		CREB		pVal	
cyclin-dependent protein kinase regulator activity		1.00E-05		replication fork		1.00E-05	
mitotic cell cycle checkpoint		1.00E-05		mitotic cell cycle checkpoint		1.00E-05	
OXPPOS		1.00E-05		M phase of mitotic cell cycle		1.00E-05	
posttranscriptional regulation of gene expression		1.00E-05		DNA repair		1.00E-05	
protein ubiquitination		1.00E-05		proteasome core complex		1.00E-05	
NRF2(GABP)		pVal		YY1		pVal	
replication fork		1.00E-05		mitochondrial ribosome		1.00E-05	
nuclear replisome		1.00E-05		mitochondrial matrix		1.00E-05	
OXPPOS		1.00E-05		translation		1.00E-05	
ribonucleoprotein complex		1.00E-05		RNA metabolic process		1.00E-05	
intracellular protein transport		1.00E-05		OXPPOS		1.00E-04	
ERRalpha		pVal		PRC		pVal	
MAP kinase phosphatase activity		1.00E-05		Positive and negative regulation of cell proliferation		1.00E-05	
regulation of transcription from RNA polymerase II promoter		1.00E-05		chromatin assembly or disassembly		1.00E-05	
positive regulation of gene expression		1.00E-05		positive regulation of signal transduction		1.00E-05	
transcription repressor activity		1.00E-05		negative regulation of cellular metabolic process		1.00E-05	
lipid transport		1.50E-03		response to hypoxia		1.00E-05	

Les facteurs de transcription NRF1, GABP et YY1, connus pour être impliqués dans la régulation de la fonction énergétique mitochondriale possèdent bien des annotations significativement associées à cette fonction. $ERR\alpha$, également connu pour réguler la fonction mitochondriale, semble dans notre modèle plutôt lié à une régulation post-traductionnelle de cette fonction, voire à un rôle important dans l'efficacité transcriptionnelle. Comme attendu, CREB et NRF1 sont impliqués dans la régulation du cycle cellulaire, plutôt dans la phase G1/S concernant NRF1, et dans la phase M pour CREB. Pour PRC, les ontologies montrent bien que ce coactivateur intervient dans des régulations positives liées au cycle et au métabolisme cellulaire mais aussi dans des régulations négatives de ces mêmes fonctions.

f. Découverte de motifs dans les régions positives en ChIP

La découverte de motifs dans les séquences positives en ChIP a pour objectif non seulement de valider le ChIP-chip en identifiant le site de fixation du facteur de

transcription étudié, mais également de déterminer la présence des cofacteurs au voisinage de ce site de fixation.

J'ai utilisé le logiciel MDmodule pour découvrir des motifs dans les séquences encadrant les sondes positives en ChIP de chaque facteur de transcription. J'ai ensuite utilisé motifsComparator pour comparer tous les motifs que j'ai découverts entre eux et avec les bases de données Jaspar et Transfac. Malheureusement, nous ne pouvons pas utiliser la base de données Genomatix® pour effectuer nos comparaisons car cette base nécessite une licence, ce qui rendrait impossible l'utilisation libre de motifsComparator. Les PWM correspondantes aux motifs présentés précédemment (Figure 25) ne sont donc pas accessibles pour effectuer les comparaisons. Dans les dendrogrammes suivants correspondant à la découverte de motifs effectuée pour chaque ChIP, les motifs ont donc été sélectionnés en fonction de leur ressemblance visuelle avec les logos disponibles sur Genomatix®.

Le dendrogramme présenté Figure 32 correspond aux motifs découverts dans les séquences positives pour CREB. Parmi l'ensemble de ces motifs, j'ai pu identifier de nombreux variants du motif caractérisant le site de fixation de CREB (en orange). Ceci permet de valider la spécificité du ChIP-chip, qui a bien permis de sélectionner des séquences reconnues par ce facteur de transcription. De plus, j'ai pu identifier dans ces séquences plusieurs motifs correspondant à NRF1 (en rouge) et un motif correspondant à GABP (en bleu). Il semblerait donc qu'un certain nombre de gènes régulés par le facteur de transcription CREB soit également régulés par les deux facteurs de la famille NRF. Enfin, d'autres motifs qui ont été découverts n'ont pas pu être associés à un facteur particulier, car ils ne sont pas répertoriés dans les bases de données. Certains motifs sont détectés comme étant des motifs répétés en tandem par motifsComparator, et peuvent être éliminés de l'analyse. Il s'agit par exemple des motifs répétés TGTG.

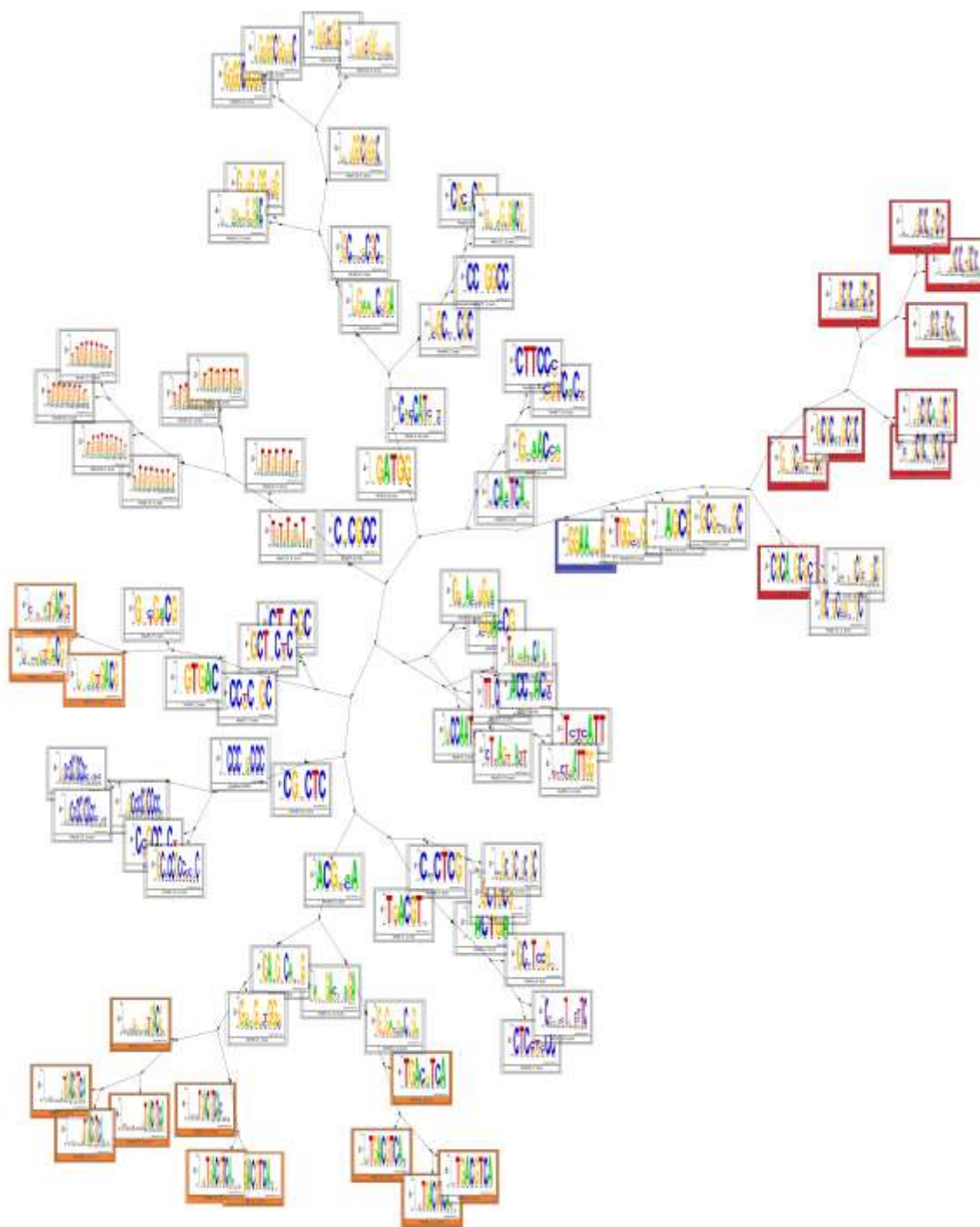


Figure 32: dendrogramme des motifs découverts pour les gènes positifs en CREB

Dendrogramme regroupant l'ensemble des motifs découverts dans les séquences encadrant les sondes positives en ChIP pour le facteur de transcription CREB. Les motifs en orange ressemblent aux motifs de CREB, les motifs en rouge à NRF1 et les motifs en bleu à GABP.

Les motifs découverts dans les séquences positives en $ERR\alpha$ sont présentés dans la Figure 33. La première chose que l'on constate en observant le dendrogramme est la forte présence de motifs très riche en A et en T. En revanche, aucun motif attendu n'est retrouvé dans ces séquences, ni celui de $ERR\alpha$, ni ceux des autres facteurs de transcription étudiés.

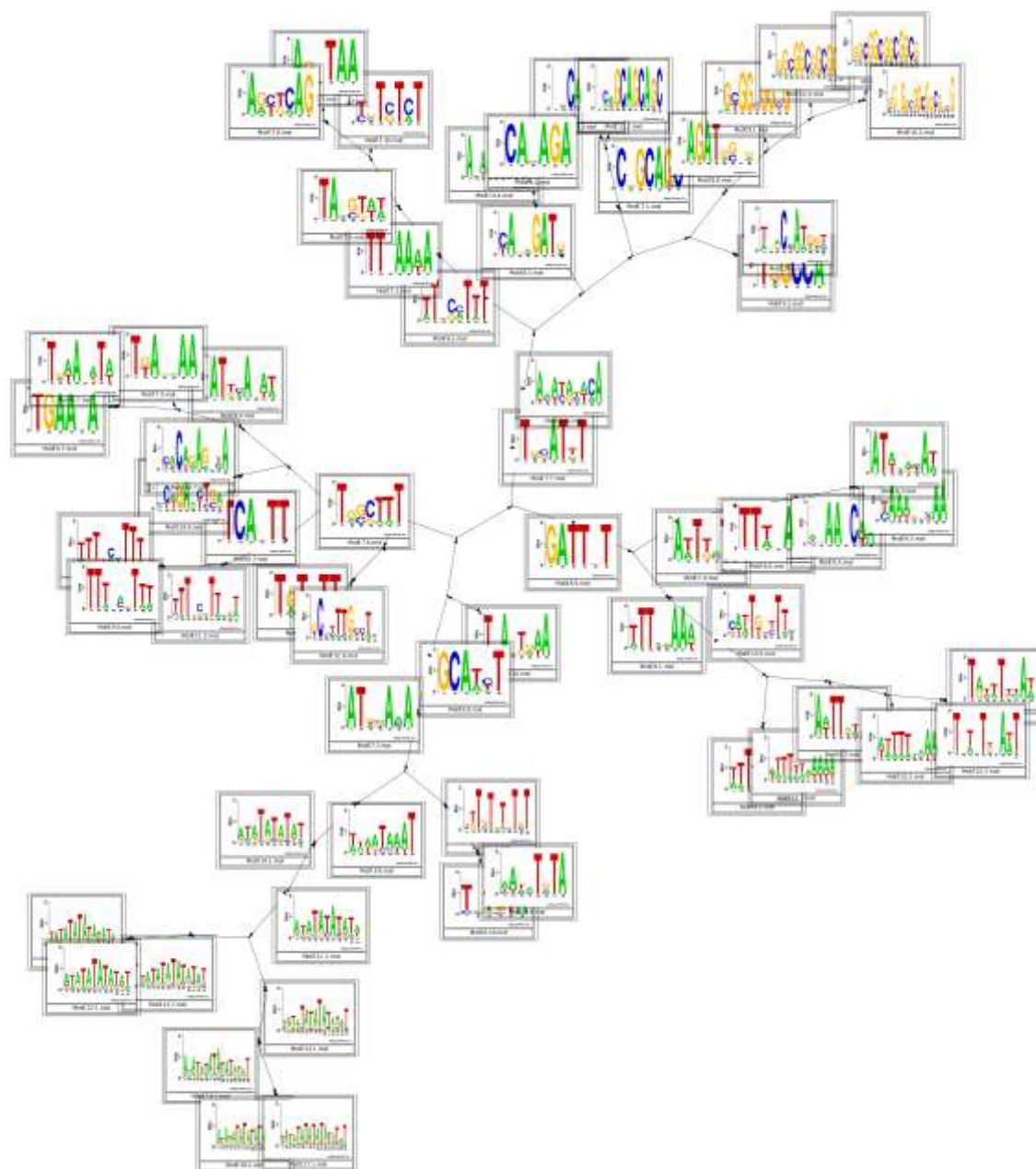


Figure 33: dendrogramme des motifs découverts pour les gènes positifs en $ERR\alpha$

Dendrogramme regroupant l'ensemble des motifs découverts dans les séquences encadrant les sondes positives en ChIP pour le facteur de transcription $ERR\alpha$.

La Figure 34 illustre les motifs découverts dans les séquences positives en ChIP du facteur de transcription GABP. Comme on peut le voir sur le dendrogramme, les motifs de GABP (en bleu) et de NRF1 (en rouge) sont retrouvés dans ces séquences. Ceci correspond bien aux résultats trouvés pour le ChIP de CREB, ainsi qu'à la littérature. En effet, NRF1 et NRF2 sont connus pour réguler ensemble des gènes codant pour des sous-unités des complexes de la chaîne respiratoire notamment. De nombreux motifs répétés en tandem sont également découverts dans ces séquences comme les motifs ACAC et leurs complémentaires inverses GTGT ainsi que les motifs ATAT.

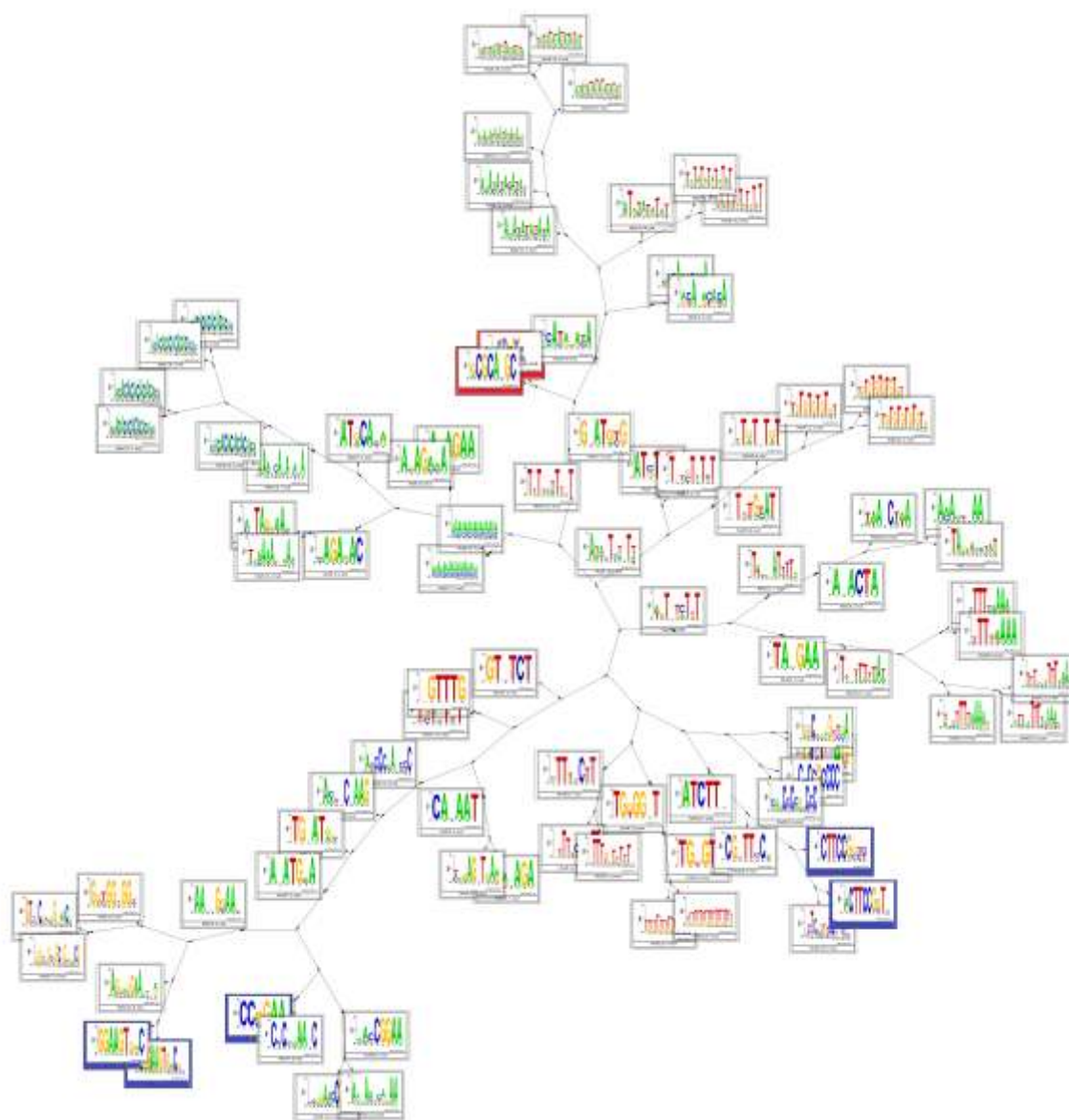


Figure 34: dendrogramme des motifs découverts pour les gènes positifs en GABP

Dendrogramme regroupant l'ensemble des motifs découverts dans les séquences encadrant les sondes positives en ChIP pour le facteur de transcription GABP. Les motifs en bleu ressemblent aux motifs de GABP et les motifs en rouge à NRF1.

Pour le ChIP-chip du facteur de transcription NRF1, le dendrogramme (Figure 35) permet de mettre en évidence de nombreuses copies du motif correspondant à NRF1, ainsi que des motifs de GABP. Ces résultats corroborent bien les résultats présentés pour le facteur GABP. De nombreux autres motifs ne sont pour le moment par encore identifiés.

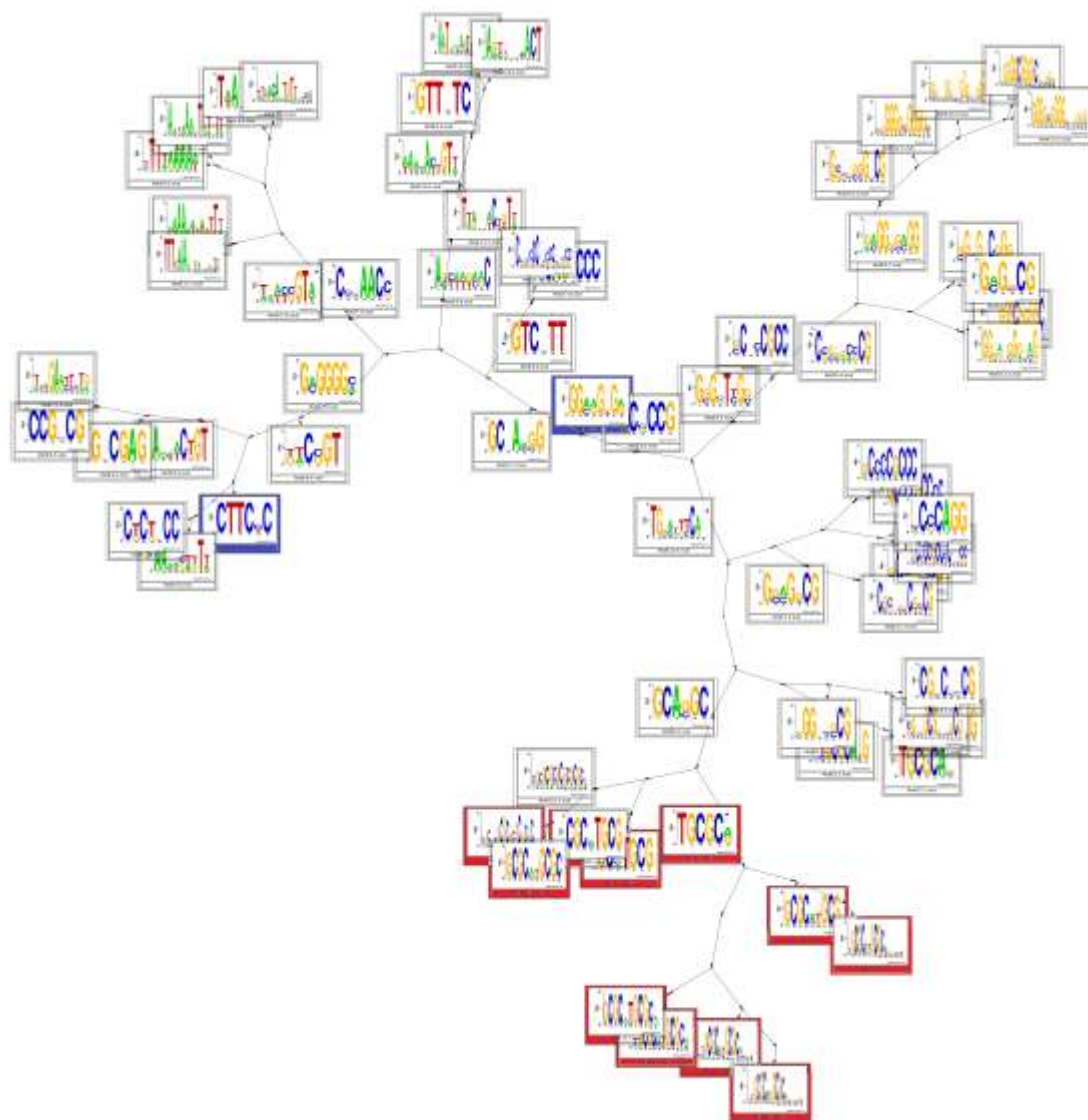


Figure 35: dendrogramme des motifs découverts pour les gènes positifs en NRF1

Dendrogramme regroupant l'ensemble des motifs découverts dans les séquences encadrant les sondes positives en ChIP pour le facteur de transcription NRF1. Les motifs en rouge ressemblent aux motifs de NRF1 et les motifs en bleu à GABP.

La découverte de motifs dans les séquences positives en ChIP PRC a pour objectif d'identifier les cofacteurs avec lesquels PRC interagit. En effet, PRC n'a pas de domaine de liaison à l'ADN, il n'est donc pas possible d'identifier un motif caractéristique le

concernant. La Figure 36 représente le dendrogramme des motifs découverts dans les séquences positives en PRC. Il ne permet pas de mettre en évidence des sites de fixation des autres facteurs de transcription que nous souhaitons étudier. Cependant, ce résultat peut s'expliquer par le nombre important de cofacteurs avec lesquels PRC interagit. En effet, si PRC interagit avec un grand nombre de partenaires, les motifs de chacun d'eux ne sont pas forcément significatifs car ils ne concernent qu'une petite fraction des gènes positifs en PRC. Hormis les motifs répétés en tandem découverts dans ces séquences, la majorité des motifs découverts sont très riches en G et C.

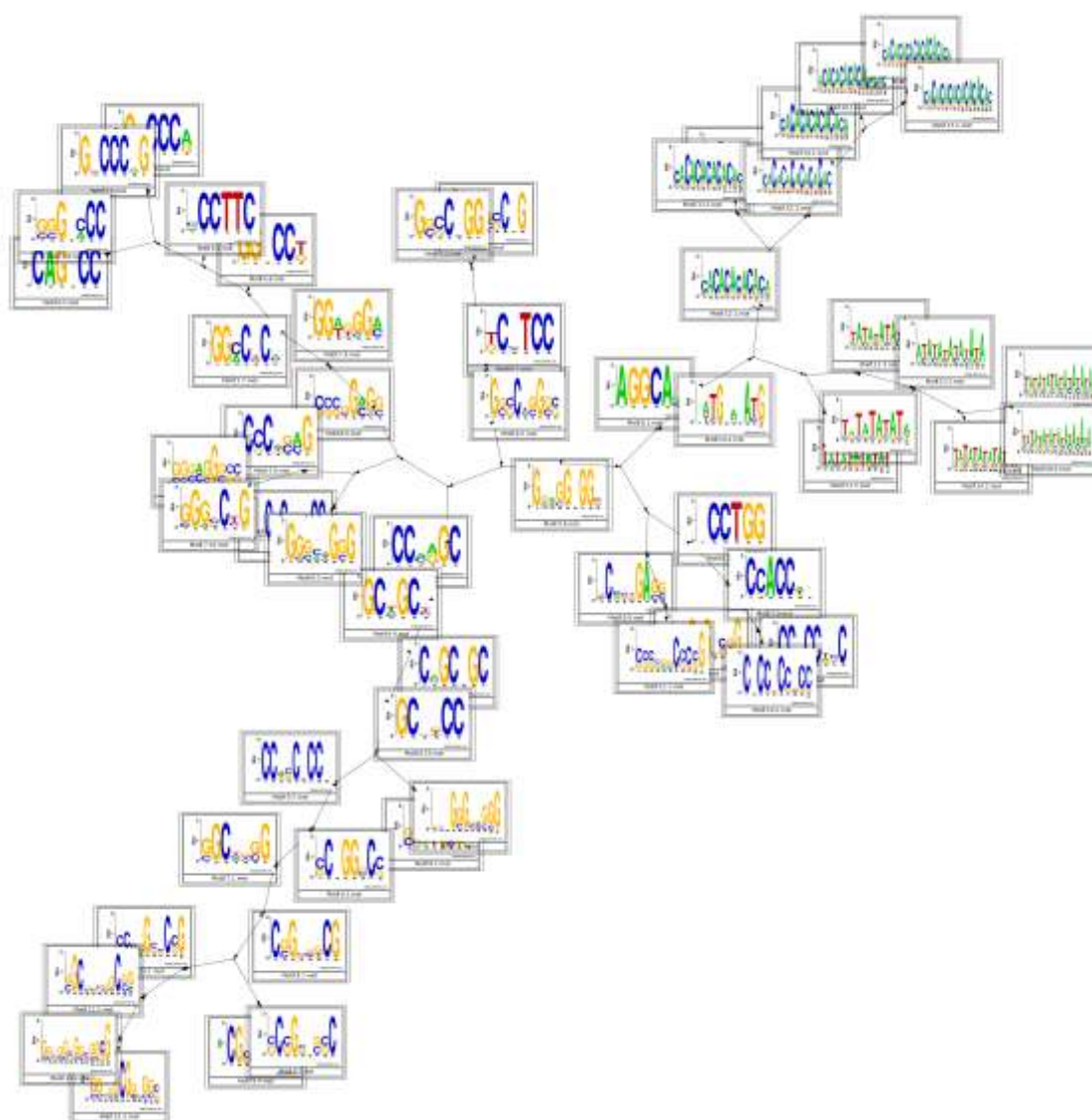


Figure 36: dendrogramme des motifs découverts pour les gènes positifs en PRC

Dendrogramme regroupant l'ensemble des motifs découverts dans les séquences encadrant les sondes positives en ChIP pour le facteur de transcription PRC.

Enfin, les motifs découverts dans les séquences positives en YY1 sont présentés dans la Figure 37. Les motifs en vert dans le dendrogramme correspondent au site de fixation du facteur de transcription de YY1. On trouve de très nombreuses copies du motif ainsi que de son complémentaire inverse. Les motifs correspondant à NRF1 (en rouge) et à GABP (en bleu) sont également découverts. Ceci permet de penser que ces trois facteurs de transcription agissent ensemble pour la régulation de certains gènes. Ce résultat est en accord avec la littérature, puisque les gènes impliqués dans la phosphorylation oxydative, l'import mitochondrial et la transcription de l'ADN mitochondrial sont connus pour être régulés par ces trois facteurs.

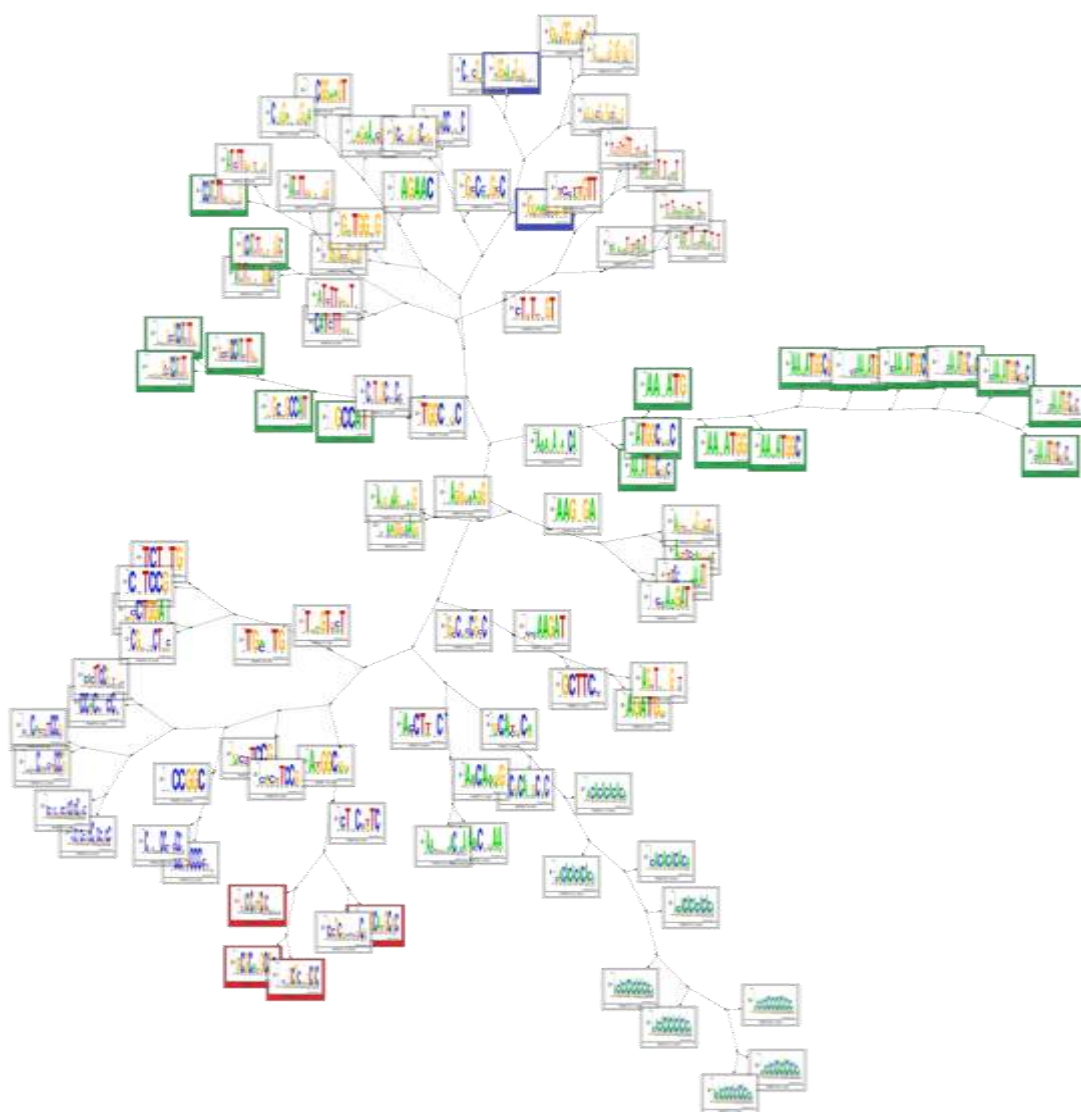


Figure 37: dendrogramme des motifs découverts pour les gènes positifs en YY1

Dendrogramme regroupant l'ensemble des motifs découverts dans les séquences encadrant les sondes positives en ChIP pour le facteur de transcription YY1. Les motifs en vert ressemblent aux motifs de YY1, les motifs en rouge à NRF1 et les motifs en bleu à GABP.

D'après l'ensemble de résultats de découverte de motifs dans les différents jeux de séquences positives en ChIP, on peut construire le réseau de motifs suivant (Figure 38) concernant les motifs des 6 facteurs de transcription étudiés.

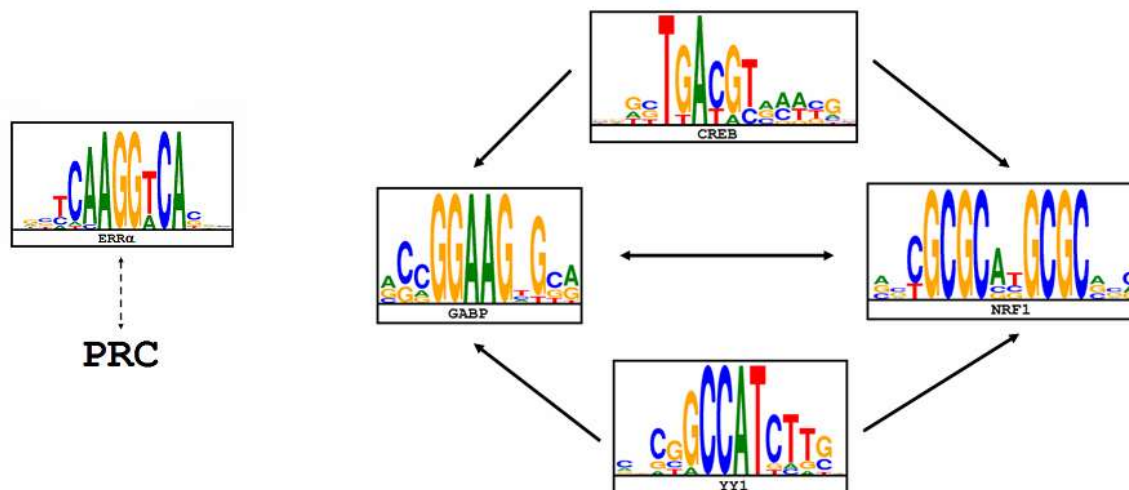


Figure 38: identification des motifs dans les jeux de séquences positives en ChIP

Les 5 logos correspondant aux facteurs de transcription étudiés en ChIP sont représentés dans ce réseau. Le motif à la source de la flèche a été retrouvé dans le jeu de séquences positives pour le facteur ciblé par la flèche. La flèche en pointillés traduit une très forte corrélation entre les facteurs, malgré une absence de découverte de motifs.

Ce réseau montre qu'il y a une forte interaction entre les quatre facteurs de transcription NRF1, GABP, CREB et YY1. Les deux autres facteurs de transcription peuvent être associés entre eux grâce aux résultats de corrélation entre les sondes positives en ChIP de chaque de ces deux facteurs. On peut noter qu'il y a une réciprocity entre les motifs des facteurs NRF1 et GABP, la présence de l'un entraînant la présence de l'autre. De plus, les facteurs CREB et YY1 sont tous les deux associés à ces deux facteurs.

g. Validation des motifs

La validation de motifs permet de s'assurer que les motifs découverts dans les séquences positives en ChIP sont bien caractéristiques du jeu de séquences positives, et qu'il ne s'agit pas de motifs enrichis dans le génome complet. Pour cela, j'ai mis au point une méthode pour comparer la distribution du score obtenu avec la matrice PWM du motif considéré dans le jeu de séquences positives en ChIP, et celle obtenue dans le jeu de séquences contrôles.

La Figure 39 montre un exemple de résultat obtenu pour les motifs correspondant au facteur de transcription YY1 découverts dans les séquences positives en ChIP pour YY1. J'ai tout d'abord construit le motif consensus regroupant toutes les occurrences du motif YY1 découvertes, puis j'ai calculé le score obtenu par la matrice PWM du motif consensus pour toutes les séquences positives en ChIP YY1 (courbe rouge) et pour un jeu de séquences contrôle (courbe bleue).

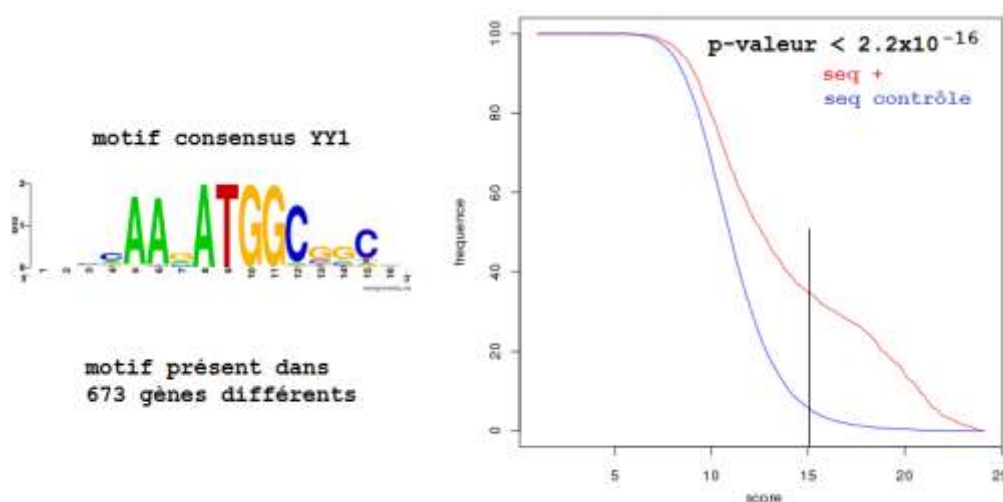
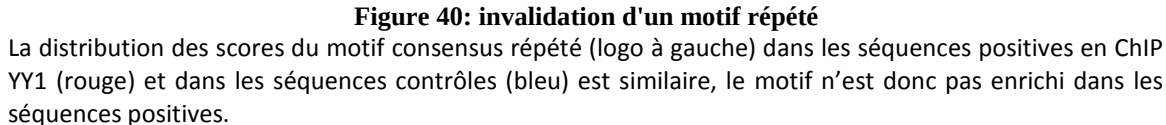


Figure 39: validation du motif correspondant à YY1

Le motif consensus (logo à gauche) construit à partir de tous les motifs communs de YY1 découvert dans les séquences positives en YY1 est localisé dans 673 gènes différents. Ce motif consensus est statistiquement surreprésenté dans les séquences positives ($p\text{-valeur} < 2.2 \times 10^{-16}$) par rapport aux séquences contrôles (courbe de droite, séquences positives en rouge et contrôles en bleu).

La construction du motif consensus permet dans un premier temps d'éliminer la redondance, ainsi que les motifs imbriqués et les complémentaires inverses, et aussi d'augmenter le nombre de gènes à considérer. Le motif consensus de YY1 est ainsi retrouvé dans 673 gènes positifs en ChIP. La distribution des scores obtenus pour le jeu de séquences positives en ChIP est très significativement différente de la courbe obtenue sur le jeu de séquences contrôles, composé du même nombre de séquences tirées aléatoirement parmi l'ensemble des sondes non positives de la puce. Le motif du facteur YY1 est donc bien enrichi très significativement dans notre jeu de séquence. Le seuil fixé au score 15 est le seuil qui discrimine le mieux les deux distributions.

Le même test statistique a été effectué sur un autre groupe de motifs, également découvert dans le jeu de séquences positives en YY1, correspondant à un motif répété en tandem. Les résultats sont présentés dans la Figure 40.



La distribution des scores du motif consensus répété (logo à gauche) dans les séquences positives en ChIP YY1 (rouge) et dans les séquences contrôles (bleu) est similaire, le motif n'est donc pas enrichi dans les séquences positives.

Pour ce motif consensus, il n'y a aucune différence entre les deux distributions de scores. Même si le motif est bien enrichi dans notre jeu de séquences, puisqu'il a été identifié comme significatif lors de la découverte de motifs par MDmodule, il n'est cependant pas caractéristique de notre jeu de ChIP puisqu'il est représenté de la même manière dans un jeu de séquences aléatoires. Ce motif doit donc être éliminé de l'analyse puisqu'il ne correspond pas à une caractéristique de notre jeu de séquences.

Cette stratégie de validation des motifs, réalisée pour tous les motifs découverts dans les différents facteurs de transcription, permet de pouvoir sélectionner les motifs spécifiques des jeux de séquences étudiés. La découverte de motifs répétés en tandem et non significatifs est un biais connu de la plupart des outils de découverte de motifs, qui préconisent le plus souvent de travailler sur des séquences masquées pour limiter ces problèmes. Cependant, des motifs de ce genre sont quand même généralement retrouvés. La méthode que j'ai développée semble donc adaptée pour éliminer *a posteriori* ces motifs, et ainsi éviter ce biais. De plus, cette méthode va permettre de mettre en évidence des motifs pour le moment non identifiés, mais significativement enrichis dans le jeu de séquences positives par rapport au jeu de séquences contrôles.

h. Intégration des résultats de ChIP des différents facteurs de transcription

La découverte de motifs dans les séquences positives en ChIP a permis de mettre en évidence des interactions entre les différents facteurs de transcription. Afin d'analyser de manière plus globale ces interactions, j'ai calculé les corrélations entre la meilleure

sonde pour chaque gène de toutes les paires de facteurs de transcription. Les résultats sont présentés dans la Figure 41.

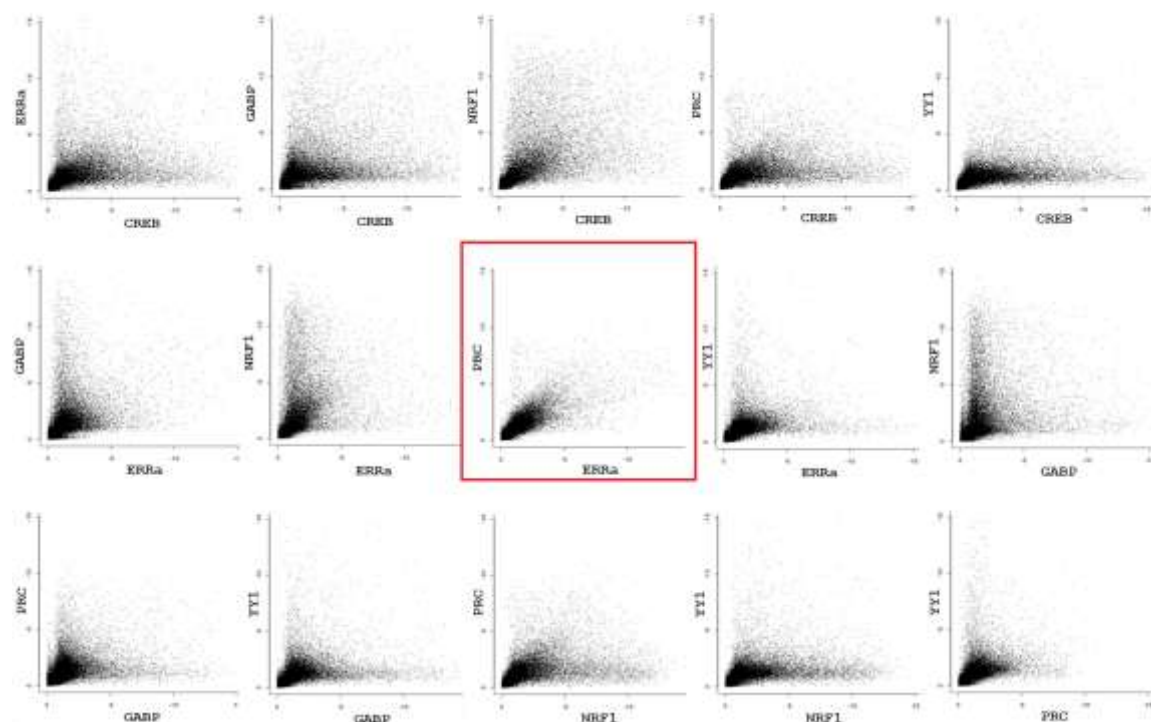


Figure 41: corrélation des meilleures sondes par gènes entre facteurs

Chaque graphe représente la corrélation de la moyenne géométrique en log base 10 de la meilleure sonde de chaque gène entre deux facteurs de transcription. Le cadre rouge correspond à la corrélation des facteurs de transcription $ERR\alpha$ et PRC.

Dans la majorité des cas, une petite proportion de gènes seulement est significative pour les deux facteurs ($\log_{10}(\text{p-valeur}) < 3$ pour les 2 facteurs). En revanche, pour les données concernant les facteurs de transcription PRC et $ERR\alpha$ (encadré rouge), on obtient une très bonne corrélation (Figure 42), avec un coefficient de corrélation de Pearson égale à 0.82 et une p-valeur inférieure à 2.2×10^{-16} .

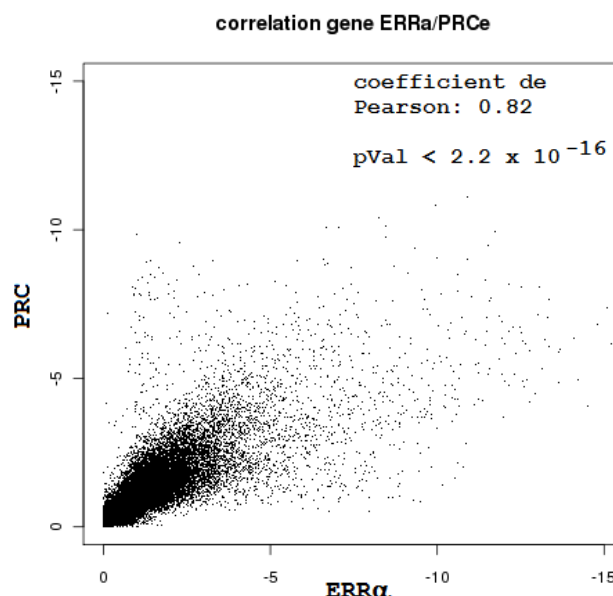


Figure 42: corrélation des gènes positifs en PRC et ERR α

Les données en abscisse correspondent aux p-Valeurs des gènes positifs en ERR α et en ordonnée aux gènes positifs en PRC.

Ces résultats montrent que la majorité des gènes positifs en PRC le sont également en ERR α . Ces deux facteurs semblent donc intervenir de manière concomitante pour la régulation de l'expression des gènes positifs en ChIP.

Afin de pouvoir étudier les résultats de ChIP-chip des six facteurs de transcription simultanément, j'ai construit une matrice de score, pour tous les gènes de la puce, basée sur la p-valeur moyenne de la meilleure sonde de chaque gène des deux répliquats de chaque facteur de transcription. J'ai ensuite effectué une classification hiérarchique de cette matrice afin d'identifier des groupes de gènes coréglés par un complexe de facteurs de transcription. Le résultat de ce clustering est montré dans la Figure 43.

Dans un premier temps, on peut voir que la classification hiérarchique des échantillons regroupe les facteurs PRC et ERR α dans une même branche. Ce résultat confirme bien la corrélation précédente, les facteurs PRC et ERR α ont un profil très similaire. On constate également que les facteurs de transcription CREB et NRF1, regroupés dans une autre branche, concernent également un grand nombre de gènes en commun.

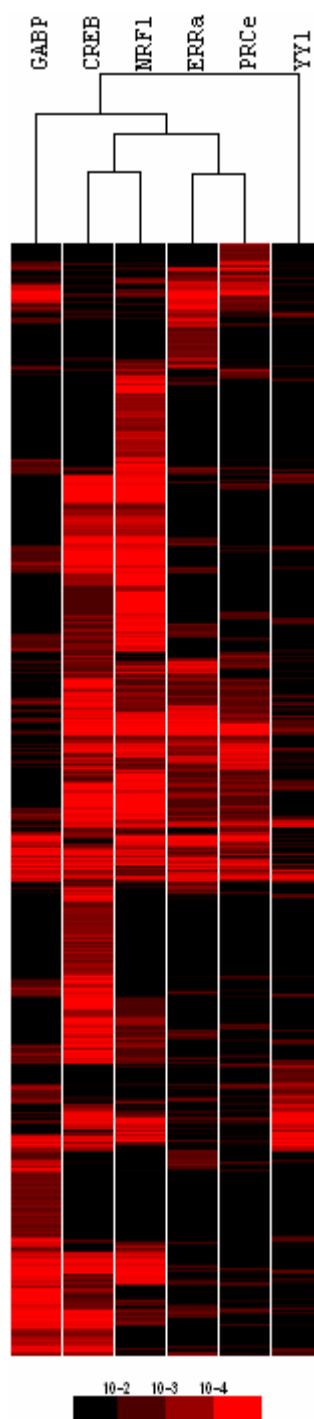


Figure 43: clustering des gènes positifs en ChIP

Chaque colonne correspond à un facteur de transcription et chaque ligne à un gène positif pour au moins un facteur. La gamme d'intensité permet de visualiser l'ordre de grandeur de la moyenne géométrique de la meilleure sonde des réplicats de chaque facteur de transcription.

Cette classification permet d'identifier des complexes de facteurs de transcription particuliers. Globalement, le complexe PRC/ERR α est associé soit au dimère CREB/NRF1, soit à GABP. Un seul groupe de gènes associe ces 5 facteurs de

transcription. Au temps 48h après reprise du cycle cellulaire, seulement 1/3 des gènes positifs en ChIP sont positifs en PRC.

Une dernière représentation possible de l'intégration des différents résultats de ChIP-chip est la construction d'un réseau à partir d'intersections des listes de gènes positifs en ChIP. Un tel réseau est présenté en Figure 44.

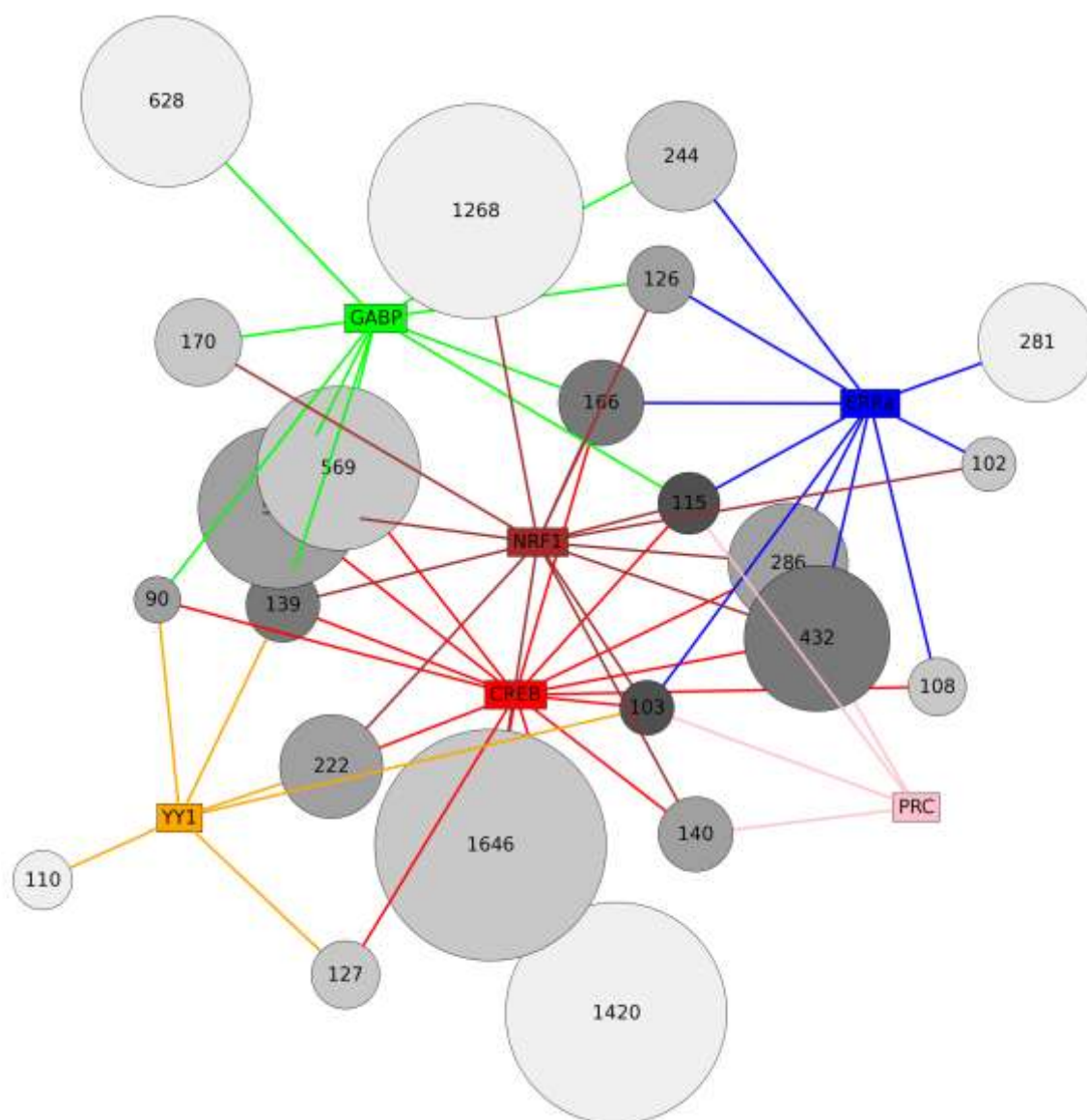


Figure 44: réseau d'intersection de listes de gènes positifs en CHIP

Chaque rectangle de couleur correspond à un facteur de transcription étudié en ChIP-chip. Chaque cercle correspond à une liste de gènes positifs pour l'ensemble des facteurs de transcription avec lesquels il est connecté. La taille du cercle est proportionnelle au nombre de gènes positifs pour ces facteurs, et son intensité de gris augmente avec la significativité de l'intersection.

Le réseau présenté ne montre que les intersections de listes dont l'effectif est supérieur à 40. Chaque cercle représente une liste de gènes, dont l'effectif est indiqué à l'intérieur du cercle, régulé par un jeu de facteurs de transcription précis. Plus l'intensité

de gris du cercle est importante, et plus la probabilité d'obtenir au hasard cet effectif pour cette intersection est faible. Malgré cette présélection sur les effectifs, la densité du réseau le rend difficilement analysable.

i. Localisation des sites de fixation sur le génome

Enfin, j'ai cherché à identifier la localisation des sites de fixation des différents facteurs de transcription les uns par rapport aux autres, et par rapport au départ de transcription. J'ai pour cela calculé les distances entre les sondes positives en ChIP et le départ de transcription du gène considéré. Etant donné que la distribution des sondes n'est pas homogène sur l'ensemble des promoteurs, j'ai calculé pour chaque position la différence de proportion entre le nombre de sondes positives et le nombre total de sondes à chaque position. La Figure 45 présente les résultats pour les 6 facteurs de transcription étudiés.

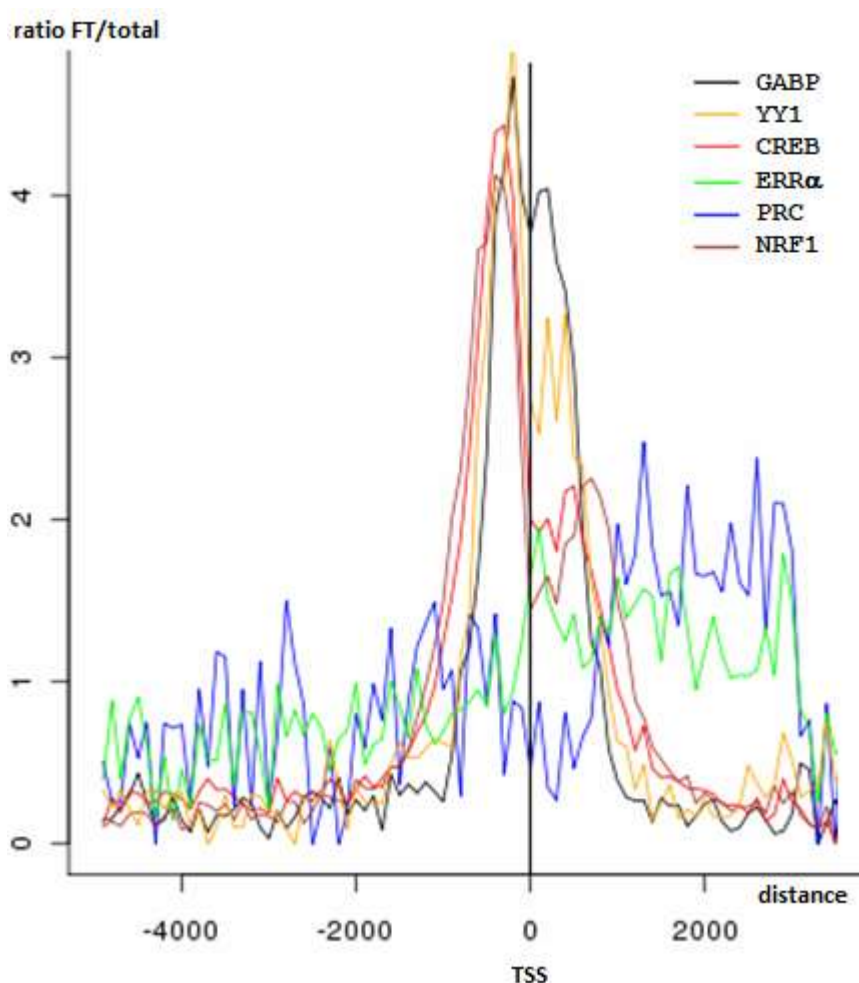


Figure 45: distribution des distances des sondes positives en ChIP aux départs de transcription
Chaque courbe correspond au ratio du nombre de sondes positives en ChIP pour un facteur donné par rapport au nombre total de sondes à une distance donnée du TSS (Transcription Start Site).

Les ratios supérieurs à 1 correspondent à des régions où le nombre de sondes positives pour le facteur considéré est proportionnellement plus important que pour l'ensemble des sondes de la puce. On observe deux profils principaux, un profil correspondant aux facteurs GABP, YY1, CREB et NRF1, avec un pic en amont du départ de transcription et un profil correspondant à PRC et $ERR\alpha$ avec une zone de fixation plus large et localisée après le départ de transcription.

Les quatre facteurs de transcription GABP, YY1, CREB et NRF1 présentent en fait une courbe avec 2 pics rapprochés et entourant le départ de transcription, un aux alentours de -700 paires de bases, et un dans les 200 premières bases après le départ de transcription. La co-localisation de ces quatre facteurs de transcription correspond bien aux résultats trouvés lors de la découverte de motifs. En effet, pour ces facteurs, il était fréquent de retrouver les motifs correspondant à leur site de fixation dans les séquences positives en ChIP de chacun d'entre eux.

Les deux facteurs PRC et $ERR\alpha$ sont majoritairement trouvés ensemble après le départ de transcription, sur une zone allant de 500 à 3000 paires de bases environ. Le fait de retrouver ces deux facteurs ensemble corrobore les résultats précédents concernant la forte corrélation obtenue entre les sondes positives de ces deux facteurs. Leur localisation à l'intérieur du transcrit explique peut être l'absence de découverte de motifs, leur site de fixation sur l'ADN étant localisé ailleurs.

Discussion et Perspectives

L'objectif de mon travail de thèse était d'étudier les stratégies de génomique intégrative à mettre en place pour comprendre les mécanismes de régulation du coactivateur transcriptionnel PRC. Ceci nécessite l'intégration de nombreuses données d'origines et de types divers. Tout d'abord des données de transcriptomique, afin de connaître le niveau d'expression de chaque gène. Ces données quantitatives, correspondant aux intensités de fluorescence émises proportionnellement à la quantité d'ARN messenger correspondant à chaque gène, sont normalisées et regroupées par classification hiérarchique pour l'analyse. Dans notre cas, ces données sont également dépendantes de l'utilisation de siRNA qui inhibe progressivement l'expression de PRC permettant ainsi de mesurer la différence d'expression entre des gènes soumis aux régulations transcriptionnelles de PRC. L'expression de ce facteur étant elle-même sous la dépendance du cycle cellulaire, il était nécessaire d'induire l'expression de PRC par du sérum avant de pouvoir l'inhiber spécifiquement. La complexité du traitement des données est alors liée à l'identification des signaux différentiels dus au sérum, de ceux dus au siRNA PRC. De plus, la mesure de cinétique a fourni des données différentes à chaque temps de mesures (0, 12, 24 et 48h) ce qui a compliqué l'analyse puisqu'il était alors possible de rechercher des effets directs et des effets indirects de PRC, ainsi que des effets précoces ou tardifs sur le niveau d'expression des gènes. Les effets indirects de ce facteur n'avaient encore jamais pu être mis en évidence puisque l'étude précédente portant sur l'inhibition de PRC avait été réalisée après Sh-RNA [Vercauteren *et al.*, 2009]. De ce fait, seuls des effets tardifs avaient pu être observés. Or, ce facteur étant au carrefour de voies métaboliques permettant l'adaptation rapide de la cellule à un stress ou aux différentes phases du cycle cellulaire, les régulations précoces paraissent essentielles à identifier. Il est en revanche très difficile de différencier les effets directs et indirects. En effet, s'agissant le plus souvent de boucles de régulations, il est difficile d'identifier le facteur initiateur. Il serait pour cela intéressant de combiner au siRNA PRC les données correspondant aux siRNA d'autres facteurs de transcription paraissant plus particulièrement impliqués dans la régulation transcriptionnelle PRC-dépendante.

A ces données s'ajoutent également des données d'annotations fonctionnelles, auxquelles sont associées une p-valeur et un enrichissement. Chaque annotation significative est associée à un groupe de gènes, qui peut correspondre à un cluster complet, ou seulement à une sous signature plus spécifique du siRNA PRC. L'étude

cinétique nous a aussi permis d'identifier un rôle plus particulier de PRC au niveau de la phase M du cycle cellulaire alors que les études précédentes avaient plutôt souligné un rôle en phase G1/S.

L'autre grande catégorie de données que nous avons étudiée pour cette analyse concerne les données de ChIP-chip. Nous possédons les données correspondant à six facteurs de transcription que sont PRC, $ERR\alpha$, NRF1, GABP, CREB, et YY1. Pour chaque jeu de données, nous avons une liste de sondes auxquelles sont associés une p-valeur pour la significativité en ChIP, une séquence, un gène et une position sur le génome. Ces informations permettent de déterminer la liste de gènes positifs pour chaque facteur, d'établir la distance au TSS préférentielle pour chacun d'entre eux et d'effectuer la découverte de motifs dans les séquences. La découverte de motifs permet non seulement d'identifier les facteurs de transcription de l'étude, mais également de découvrir de nouveaux cofacteurs potentiels, qui pourront également être étudié en ChIP-chip dans une prochaine étude. Etudier plusieurs facteurs de transcription sur la même lignée cellulaire et dans les mêmes conditions permet d'analyser les différents complexes qui peuvent se former pour la régulation spécifique de fonctions particulières, d'identifier les partenaires privilégiés de PRC, et aussi de détecter des boucles de régulation entre ces différents facteurs. Plus le nombre de facteurs étudiés est important, et plus il sera possible de construire un réseau le plus exhaustif possible de la régulation transcriptionnelle PRC-dépendante. Cependant, la difficulté et les incohérences entre les différents résultats augmentent d'autant que le nombre de facteurs étudiés est important.

La diversité et la complexité de l'ensemble de ces données fournissent des angles de vue différents sur un même système biologique, et peuvent parfois sembler contradictoires. Ceci rend l'intégration de ces données très difficile, notamment quand la notion de cinétique est intégrée. En effet, nous avons cherché à corréliser les effets observés sur le transcriptome à 12, 24 et 48h de traitement par le siRNA PRC aux études de ChIP-chip concernant l'interaction des facteurs de transcription à 48 heures. A ce temps tardif, il est alors difficile d'identifier les effets directs de PRC des effets indirects. Il est donc probable que certains gènes dont l'expression est différentielle soient régulés par d'autres facteurs de transcription, eux mêmes régulés par PRC et ses cofacteurs. Ainsi, la recherche de signatures spécifiques du siRNA PRC pour chaque groupe de gènes (transcriptome) en fonction des gènes positifs en tel ou tel facteur de transcription (ChIP-chip) n'a pas permis d'identifier d'enrichissement significatif. Il serait donc peut être utile d'effectuer de nouvelles expériences de ChIP-chip sur des cellules ayant repris leur cycle

cellulaire depuis 12 ou 24 heures seulement. Cependant, nous avons fait dans un premier temps le choix des cellules à 48h car c'est à ce temps que le siRNA était le plus efficace pour l'inhibition de PRC, et qu'il était également plus facile d'avoir un nombre suffisant de cellules (1 million par ChIP) à un nombre de passage relativement bas (P68) pour effectuer les manipulations de ChIP-chip. Si l'on considère le temps 12 heures, le niveau d'inhibition de PRC n'est peut-être pas suffisant mais le signal lié au sérum est maximal ce qui peut permettre de différencier les deux effets dans la cinétique. De plus, en utilisant des cellules seulement après 12 ou 24h de reprise du cycle cellulaire, les effets observés ne correspondront qu'aux réponses les plus rapides à PRC.

L'hypothèse d'effet indirect de PRC semble confirmée par le réseau d'interactions que j'ai pu construire (Figure 31) en regardant quels facteurs sont présents dans les listes de gènes positifs pour chaque facteur étudié par ChIP-chip. Ce réseau montre que PRC régule l'expression de CREB, qui lui-même intervient dans la régulation de plusieurs autres facteurs de transcription, dont PRC. Les niveaux d'expression de $ERR\alpha$ et de NRF1 ne semblent pas être régulés, directement ou non, par PRC dans notre modèle. Peut-être qu'un autre facteur régulé par PRC, non étudié pour le moment, joue un rôle important pour la régulation de ces deux facteurs. Afin de compléter notre analyse, nous avons choisi d'étudier également le facteur de transcription cMYC en ChIP-chip, un facteur de transcription également retrouvé dans un cluster du transcriptome. La boucle de régulation entre CREB et PRC rend compréhensible le mécanisme de coordination de la prolifération mitochondriale, notamment ici par l'intermédiaire de PRC et de GABP, et cellulaire par le rôle central de CREB dans ce modèle.

Un autre point difficile à intégrer correspond aux résultats de découverte de motifs dans les gènes positifs en ChIP pour les facteurs NRF1 et GABP. En effet, en accord avec la littérature, on retrouve bien des motifs de NRF1 et de GABP dans les deux jeux de séquences positives en NRF1 et GABP (Figure 38). En revanche, on ne trouve pas de corrélation particulière entre les sondes positives pour ces deux facteurs. Les gènes cibles, même s'ils ont la possibilité de fixer les deux facteurs de transcription, ne sont régulés que par un seul des facteurs de transcription dans notre étude. Peut-être que ce phénomène peut s'expliquer par une non synchronisation entre ces deux événements, que nous ne pouvons pas observer en étudiant uniquement le temps 48h. Un encombrement stérique entre les deux facteurs dont les sites de fixation seraient trop proches sur les gènes peut aussi expliquer cette absence de corrélation. Il serait peut-être intéressant de faire

également une étude cinétique en ChIP-chip afin d'étudier l'ordre d'arrivée des différents facteurs sur les promoteurs des gènes.

De plus, même si l'on observe une très bonne corrélation entre les sondes positives en ChIP des facteurs de transcription PRC et $ERR\alpha$, ainsi qu'une localisation commune dans les 1500 premières bases des transcrits des gènes positifs en ChIP, on ne retrouve pas le motif de $ERR\alpha$, ni dans les séquences positives en PRC, ni dans les séquences positives en $ERR\alpha$. On ne trouve pas non plus de motifs pour PRC, mais ceci est normal puisque PRC ne se fixe pas directement à l'ADN, mais qu'il interagit en tant que coactivateur transcriptionnel avec d'autres facteurs de transcription. Une hypothèse concernant ces deux phénomènes (absence de motifs de $ERR\alpha$, et localisation des sondes après le départ de transcription sur une zone assez large) serait que le complexe PRC/ $ERR\alpha$ soit emporté avec le complexe d'initiation de la transcription sur la chromatine pour la transcription des gènes. Pour vérifier cette hypothèse, nous avons réalisé deux expériences de ChIP-chip concernant le facteur de transcription RNAPol II, afin de comparer la position des sondes positives en ChIP pour RNAPol II avec la position des sondes positives pour PRC et $ERR\alpha$. Si nous trouvons une localisation commune pour les sondes positives pour ces 3 facteurs, il faudra essayer de localiser les sites de fixation de $ERR\alpha$ initiaux, et valider biologiquement ce phénomène. Ce type de mécanisme a été précédemment décrit sur une étude de ChIP-chip de l'ARNpolII [Guenther et al., 2007]. Il a été montré que les gènes activement transcrits ne présentaient plus l'ARNpol II au niveau du promoteur.

Bibliographie

- Andersson, U., et R. C. Scarpulla (2001), Pgc-1-related coactivator, a novel, serum-inducible coactivator of nuclear respiratory factor 1-dependent transcription in mammalian cells, *Mol. Cell. Biol*, 21(11), 3738-3749, doi:10.1128/MCB.21.11.3738-3749.2001.
- Ashburner, M. et al. (2000), Gene Ontology: tool for the unification of biology, *Nat Genet*, 25(1), 25-29, doi:10.1038/75556.
- Bailey, T. L., et C. Elkan (1994), Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc Int Conf Intell Syst Mol Biol*, 2, 28-36.
- Birney, E. et al. (2007), Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, 447(7146), 799-816, doi:10.1038/nature05874.
- Carat, S., R. Houlgatte, et J. Bourdon (2010), A parallel scheme for comparing transcription factor binding sites matrices, *J Bioinform Comput Biol*, 8(3), 485-502.
- Carroll, J. S. et al. (2006), Genome-wide analysis of estrogen receptor binding sites, *Nat. Genet*, 38(11), 1289-1297, doi:10.1038/ng1901.
- Conlon, E. M., X. S. Liu, J. D. Lieb, et J. S. Liu (2003), Integrating regulatory motif discovery and genome-wide expression analysis, *Proc. Natl. Acad. Sci. U.S.A*, 100(6), 3339-3344, doi:10.1073/pnas.0630591100.
- Eisen, M. B., P. T. Spellman, P. O. Brown, et D. Botstein (1998), Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U.S.A*, 95(25), 14863-14868.
- ENCODE Project Consortium (2004), The ENCODE (ENCyclopedia Of DNA Elements) Project, *Science*, 306(5696), 636-640, doi:10.1126/science.1105136.
- Evans, M. J., et R. C. Scarpulla (1990), NRF-1: a trans-activator of nuclear-encoded respiratory genes in animal cells, *Genes Dev*, 4(6), 1023-1034.
- Fiers, W., R. Contreras, R. De Wachter, G. Haegeman, J. Merregaert, W. M. Jou, et A. Vandenberghe (1971), Recent progress in the sequence determination of bacteriophage MS2 RNA, *Biochimie*, 53(4), 495-506.
- Gerstein, M. B., C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korb, O. Emanuelsson, Z. D. Zhang, S. Weissman, et M. Snyder (2007), What is a gene, post-ENCODE? History and updated definition, *Genome Research*, 17(6), 669-681, doi:10.1101/gr.6339607.
- Giguère, V. (2008), Transcriptional control of energy homeostasis by the estrogen-related

- receptors, *Endocr. Rev*, 29(6), 677-696, doi:10.1210/er.2008-0017.
- Gilchrist, D. A., D. C. Fargo, et K. Adelman (2009), Using ChIP-chip and ChIP-seq to study the regulation of gene expression: Genome-wide localization studies reveal widespread regulation of transcription elongation, *Methods*, 48(4), 398-408, doi:10.1016/j.ymeth.2009.02.024.
- Gleyzer, N., K. Vercauteren, et R. C. Scarpulla (2005), Control of mitochondrial transcription specificity factors (TFB1M and TFB2M) by nuclear respiratory factors (NRF-1 and NRF-2) and PGC-1 family coactivators, *Mol. Cell. Biol*, 25(4), 1354-1366, doi:10.1128/MCB.25.4.1354-1366.2005.
- Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, et R. A. Young (2007), A chromatin landmark and transcription initiation at most promoters in human cells, *Cell*, 130(1), 77-88, doi:10.1016/j.cell.2007.05.042.
- Gugneja, S., et R. C. Scarpulla (1997), Serine phosphorylation within a concise amino-terminal domain in nuclear respiratory factor 1 enhances DNA binding, *J. Biol. Chem*, 272(30), 18732-18739.
- Hock, M. B., et A. Kralli (2009), Transcriptional control of mitochondrial biogenesis and function, *Annu. Rev. Physiol*, 71, 177-203, doi:10.1146/annurev.physiol.010908.163119.
- Hughes, J. D., P. W. Estep, S. Tavazoie, et G. M. Church (2000), Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*, *J. Mol. Biol*, 296(5), 1205-1214, doi:10.1006/jmbi.2000.3519.
- Johnson, D. S., A. Mortazavi, R. M. Myers, et B. Wold (2007), Genome-wide mapping of in vivo protein-DNA interactions, *Science*, 316(5830), 1497-1502, doi:10.1126/science.1141319.
- Liu, X. S., D. L. Brutlag, et J. S. Liu (2002), An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments, *Nat. Biotechnol*, 20(8), 835-839, doi:10.1038/nbt717.
- Mardis, E. R. (2008), The impact of next-generation sequencing technology on genetics, *Trends in Genetics*, 24(3), 133-141, doi:10.1016/j.tig.2007.12.007.
- Margulies, M. et al. (2005), Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, 437(7057), 376-380, doi:10.1038/nature03959.
- Miller, J. R., S. Koren, et G. Sutton (2010), Assembly algorithms for next-generation sequencing data, *Genomics*, 95(6), 315-327, doi:10.1016/j.ygeno.2010.03.001.
- Nirenberg, M., P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, et C. O'Neal (1965), RNA codewords and protein synthesis, VII. On the general nature of the RNA code., *Proc Natl Acad Sci U S A*, 53(5), 1161-1168.

- Ongwijitwat, S., H. L. Liang, E. M. Graboyes, et M. T. T. Wong-Riley (2006), Nuclear respiratory factor 2 senses changing cellular energy demands and its silencing down-regulates cytochrome oxidase and other target gene mRNAs, *Gene*, 374, 39-49, doi:10.1016/j.gene.2006.01.009.
- Park, P. J. (2009), ChIP-seq: advantages and challenges of a maturing technology, *Nat. Rev. Genet.*, 10(10), 669-680, doi:10.1038/nrg2641.
- Puigserver, P., Z. Wu, C. W. Park, R. Graves, M. Wright, et B. M. Spiegelman (1998), A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis, *Cell*, 92(6), 829-839.
- Raharijaona, M., S. Le Pennec, J. Poirier, D. Mirebeau-Prunier, C. Rouxel, C. Jacques, J. Fontaine, Y. Malthiery, R. Houlgatte, et F. Savagner (2009), PGC-1-related coactivator modulates mitochondrial-nuclear crosstalk through endogenous nitric oxide in a cellular model of oncocytic thyroid tumours, *PLoS ONE*, 4(11), e7964, doi:10.1371/journal.pone.0007964.
- Robertson, G. et al. (2007), Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, *Nat. Methods*, 4(8), 651-657, doi:10.1038/nmeth1068.
- Sandelin, A., W. Alkema, P. Engström, W. W. Wasserman, et B. Lenhard (2004), JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res*, 32(Database issue), D91-94, doi:10.1093/nar/gkh012.
- Sanger, F., S. Nicklen, et A. R. Coulson (1977), DNA sequencing with chain-terminating inhibitors, *Biotechnology*, 24, 104-108.
- Scarpulla, R. C. (2008), Transcriptional Paradigms in Mammalian Mitochondrial Biogenesis and Function, *Physiol. Rev.*, 88(2), 611-638, doi:10.1152/physrev.00025.2007.
- Schena, M., D. Shalon, R. W. Davis, et P. O. Brown (1995), Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270(5235), 467-470.
- Schneider, T. D., et R. M. Stephens (1990), Sequence logos: a new way to display consensus sequences., *Nucleic Acids Res*, 18(20), 6097-6100.
- Shinde, K., M. Phatak, F. M. Johannes, J. Chen, Q. Li, J. K. Vineet, Z. Hu, K. Ghosh, J. Meller, et M. Medvedovic (2010), Genomics Portals: integrative web-platform for mining genomics data, *BMC Genomics*, 11, 27-27, doi:10.1186/1471-2164-11-27.
- Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho, et G. M. Church (1999), Systematic determination of genetic network architecture, *Nat. Genet.*, 22(3), 281-285, doi:10.1038/10343.
- Thomas, M. C., et C. Chiang (2006), The general transcription machinery and general cofactors, *Crit. Rev. Biochem. Mol. Biol.*, 41(3), 105-178,

doi:10.1080/10409230600648736.

- Vercauteren, K., N. Gleyzer, et R. C. Scarpulla (2009), Short hairpin RNA-mediated silencing of PRC (PGC-1-related coactivator) results in a severe respiratory chain deficiency associated with the proliferation of aberrant mitochondria, *J. Biol. Chem*, 284(4), 2307-2319, doi:10.1074/jbc.M806434200.
- Vercauteren, K., R. A. Pasko, N. Gleyzer, V. M. Marino, et R. C. Scarpulla (2006), PGC-1-related coactivator: immediate early expression and characterization of a CREB/NRF-1 binding domain associated with cytochrome c promoter occupancy and respiratory growth, *Mol. Cell. Biol*, 26(20), 7409-7419, doi:10.1128/MCB.00585-06.
- Villena, J. A., et A. Kralli (2008), ERR α : a metabolic function for the oldest orphan, *Trends Endocrinol Metab*, 19(8), 269-276, doi:10.1016/j.tem.2008.07.005.
- Wang, Z., M. Gerstein, et M. Snyder (2009), RNA-Seq: a revolutionary tool for transcriptomics, *Nat. Rev. Genet*, 10(1), 57-63, doi:10.1038/nrg2484.
- Wasserman, W. W., et A. Sandelin (2004), Applied bioinformatics for the identification of regulatory elements, *Nat. Rev. Genet*, 5(4), 276-287, doi:10.1038/nrg1315.
- Watson, J. D., et F. H. Crick (1953), Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. J.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953, *Nature*, 248(5451), 765.
- van Waveren, C., et C. T. Moraes (2008), Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system, *BMC Genomics*, 9, 18-18, doi:10.1186/1471-2164-9-18.
- Weinstock, G. M. (2007), ENCODE: more genomic empowerment, *Genome Res*, 17(6), 667-668, doi:10.1101/gr.6534207.
- Wingender, E., P. Dietze, H. Karas, et R. Knüppel (1996), TRANSFAC: a database on transcription factors and their DNA binding sites, *Nucleic Acids Res*, 24(1), 238-241.
- Zaid, A., R. Li, K. Luciakova, P. Barath, S. Nery, et B. D. Nelson (1999), On the role of the general transcription factor Sp1 in the activation and repression of diverse mammalian oxidative phosphorylation genes, *J. Bioenerg. Biomembr*, 31(2), 129-135.
- Zeeberg, B. R. et al. (2003), GoMiner: a resource for biological interpretation of genomic and proteomic data, *Genome Biol*, 4(4), R28-R28, doi:10.1186/gb-2003-4-4-r28.

Annexes

I- Annotation fonctionnelle des gènes positifs en ChIP

Les figures ci-dessous représentent les 30 annotations fonctionnelles les plus significatives obtenues avec goMiner pour chacun des facteurs de transcription étudiés. La synthèse de ces annotations (en gras) est présentée dans la partie résultats, Tableau 6.

CREB

GO.ID	Total	Change	Enrichissement	P.Value	Term
75	81	45	1.74	0	cell cycle checkpoint
87	238	129	1.70	0	M phase of mitotic cell cycle
151	77	45	1.84	0	ubiquitin ligase complex
226	127	65	1.61	0	microtubule cytoskeleton organization
278	422	221	1.64	0	mitotic cell cycle
279	334	169	1.59	0	M phase
280	236	128	1.70	0	nuclear division
502	46	28	1.91	0	proteasome complex
775	102	65	2.00	0	chromosome. centromeric region
776	61	42	2.16	0	kinetochore
777	42	29	2.17	0	condensed chromosome kinetochore
779	46	30	2.05	0	condensed chromosome. centromeric region
785	157	85	1.70	0	chromatin
793	98	55	1.76	0	condensed chromosome
910	35	23	2.06	0	cytokinesis
3697	55	35	2.00	0	single-stranded DNA binding
3713	201	100	1.56	0	transcription coactivator activity
3723	647	312	1.51	0	RNA binding
3735	148	79	1.68	0	structural constituent of ribosome
4298	19	15	2.48	0	threonine-type endopeptidase activity
4386	125	69	1.73	0	helicase activity
4842	126	67	1.67	0	ubiquitin-protein ligase activity
5635	175	100	1.79	0	nuclear envelope
5643	68	45	2.08	0	nuclear pore
5654	753	360	1.50	0	nucleoplasm
5657	34	24	2.22	0	replication fork
5681	129	70	1.70	0	spliceosomal complex
5694	388	214	1.73	0	chromosome
5759	208	100	1.51	0	mitochondrial matrix
6281	263	133	1.59	0	DNA repair

ERRα

GO.ID	Total	Change	Enrichissement	P.Value	Term
785	157	42	2.15	0	chromatin
786	55	19	2.77	0	nucleosome
4725	92	27	2.35	0	protein tyrosine phosphatase activity
5925	93	26	2.24	0	focal adhesion
6333	113	33	2.34	0	chromatin assembly or disassembly
6357	607	116	1.53	0	regulation of transcription from RNA polymerase II promoter
8134	468	89	1.53	0	transcription factor binding
8138	40	18	3.61	0	protein tyrosine/serine/threonine phosphatase activity
9653	1036	198	1.53	0	anatomical structure morphogenesis
9790	434	83	1.53	0	embryonic development
9986	253	56	1.78	0	cell surface
10628	464	88	1.52	0	positive regulation of gene expression
10926	922	177	1.54	0	anatomical structure formation
16564	263	57	1.74	0	transcription repressor activity
17017	13	8	4.94	0	MAP kinase tyrosine/serine/threonine phosphatase activity
30055	100	28	2.25	0	cell-substrate junction
32989	359	71	1.59	0	cellular component morphogenesis
33549	13	8	4.94	0	MAP kinase phosphatase activity
43353	7	6	6.87	0	enucleate erythrocyte differentiation
44427	319	65	1.63	0	chromosomal part
45893	383	77	1.61	0	positive regulation of transcription. DNA-dependent
48646	310	67	1.73	0	anatomical structure formation involved in morphogenesis
51093	478	92	1.54	0	negative regulation of developmental process
51254	385	77	1.60	0	positive regulation of RNA metabolic process
65003	501	94	1.50	0	macromolecular complex assembly
902	328	65	1.59	1.00E-04	cell morphogenesis
5694	388	75	1.55	1.00E-04	chromosome
5924	97	26	2.15	1.00E-04	cell-substrate adherens junction
6605	217	47	1.74	1.00E-04	protein targeting
6869	138	30	1.74	0.0015	lipid transport

GABP

GO.ID	Total	Change	Enrichissement	P.Value	Term
375	152	48	2.04	0	RNA splicing. via transesterification reactions
377	152	48	2.04	0	RNA splicing. via transesterification reactions with bulged adenosine as nucleophile
398	152	48	2.04	0	nuclear mRNA splicing. via spliceosome
3723	647	155	1.55	0	RNA binding
3735	148	51	2.23	0	structural constituent of ribosome
5654	753	188	1.61	0	nucleoplasm
5657	34	16	3.04	0	replication fork
5681	129	48	2.41	0	spliceosomal complex
5730	643	174	1.75	0	nucleolus
5740	355	85	1.55	0	mitochondrial envelope
5840	186	62	2.16	0	ribosome
6396	505	157	2.01	0	RNA processing
6397	301	95	2.04	0	mRNA processing
6399	105	41	2.53	0	tRNA metabolic process
6412	370	108	1.89	0	translation
6886	411	101	1.59	0	intracellular protein transport
8033	66	25	2.45	0	tRNA processing
8380	274	91	2.15	0	RNA splicing
16071	346	102	1.91	0	mRNA metabolic process
22613	165	56	2.19	0	ribonucleoprotein complex biogenesis
22904	52	20	2.49	0	respiratory electron transport chain
30529	447	147	2.13	0	ribonucleoprotein complex
31145	63	23	2.36	0	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
31966	335	80	1.54	0	mitochondrial membrane
31974	1637	400	1.58	0	membrane-enclosed lumen
31981	1286	327	1.64	0	nuclear lumen
33279	118	39	2.14	0	ribosomal subunit
34470	164	54	2.13	0	ncRNA processing
34613	441	107	1.57	0	cellular protein localization
34660	204	71	2.25	0	ncRNA metabolic process

NRF1

GO.ID	Total	Change	Enrichissement	P.Value	Term
75	81	41	1.80	0	cell cycle checkpoint
226	127	61	1.71	0	microtubule cytoskeleton organization
278	422	178	1.50	0	mitotic cell cycle
775	102	48	1.68	0	chromosome. centromeric region
785	157	76	1.73	0	chromatin
4842	126	69	1.95	0	ubiquitin-protein ligase activity
5635	175	80	1.63	0	nuclear envelope
5643	68	35	1.84	0	nuclear pore
5694	388	180	1.65	0	chromosome
5759	208	95	1.63	0	mitochondrial matrix
5819	111	56	1.80	0	spindle
6091	301	129	1.53	0	generation of precursor metabolites and energy
6119	84	41	1.74	0	oxidative phosphorylation
6260	207	88	1.52	0	DNA replication
6325	344	149	1.54	0	chromatin organization
6333	113	58	1.83	0	chromatin assembly or disassembly
6396	505	219	1.55	0	RNA processing
6399	105	53	1.80	0	tRNA metabolic process
6412	370	158	1.52	0	translation
6457	154	71	1.64	0	protein folding
6511	212	93	1.56	0	ubiquitin-dependent protein catabolic process
6886	411	178	1.54	0	intracellular protein transport
6916	193	84	1.55	0	anti-apoptosis
7005	106	51	1.72	0	mitochondrion organization
7093	49	28	2.04	0	mitotic cell cycle checkpoint
8380	274	116	1.51	0	RNA splicing
10608	177	75	1.51	0	posttranscriptional regulation of gene expression
15630	456	193	1.51	0	microtubule cytoskeleton
16538	15	12	2.85	0	cyclin-dependent protein kinase regulator activity
16567	117	64	1.95	0	protein ubiquitination

PRC

GO.ID	Total	Change	Enrichissement	P.Value	Term
785	157	28	2.71	0	chromatin
786	55	14	3.87	0	nucleosome
1501	289	38	2.00	0	skeletal system development
1503	123	21	2.60	0	ossification
1568	240	36	2.28	0	blood vessel development
1666	79	16	3.08	0	response to hypoxia
1944	244	37	2.31	0	vasculature development
5694	388	47	1.84	0	chromosome
6333	113	21	2.83	0	chromatin assembly or disassembly
6334	69	15	3.31	0	nucleosome assembly
6793	1106	111	1.53	0	phosphorus metabolic process
6796	1106	111	1.53	0	phosphate metabolic process
6928	501	57	1.73	0	cell motion
8219	979	99	1.54	0	cell death
8283	945	100	1.61	0	cell proliferation
8285	305	40	1.99	0	negative regulation of cell proliferation
9653	1036	106	1.56	0	anatomical structure morphogenesis
9888	604	69	1.74	0	tissue development
9890	472	57	1.84	0	negative regulation of biosynthetic process
9892	631	76	1.83	0	negative regulation of metabolic process
9967	246	34	2.10	0	positive regulation of signal transduction
10558	454	55	1.84	0	negative regulation of macromolecule biosynthetic process
10605	595	69	1.76	0	negative regulation of macromolecule metabolic process
10627	217	32	2.24	0	regulation of protein kinase cascade
10740	150	24	2.43	0	positive regulation of protein kinase cascade
10926	922	109	1.80	0	anatomical structure formation
10941	657	70	1.62	0	regulation of cell death
16265	983	99	1.53	0	death
16477	317	43	2.06	0	cell migration
22607	668	76	1.73	0	cellular component assembly

YY1

GO.ID	Total	Change	Enrichissement	P.Value	Term
313	49	14	5.40	0	organellar ribosome
314	19	7	6.96	0	organellar small ribosomal subunit
375	152	27	3.35	0	RNA splicing, via transesterification reactions
377	152	27	3.35	0	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
398	152	27	3.35	0	nuclear mRNA splicing, via spliceosome
3723	647	83	2.42	0	RNA binding
3735	148	29	3.70	0	structural constituent of ribosome
5739	966	94	1.84	0	mitochondrion
5759	208	31	2.81	0	mitochondrial matrix
5761	49	14	5.40	0	mitochondrial ribosome
5763	19	7	6.96	0	mitochondrial small ribosomal subunit
5840	186	34	3.45	0	ribosome
6396	505	72	2.69	0	RNA processing
6397	301	45	2.82	0	mRNA processing
6412	370	61	3.11	0	translation
8380	274	42	2.89	0	RNA splicing
10467	3247	280	1.63	0	gene expression
16070	2157	180	1.58	0	RNA metabolic process
16071	346	48	2.62	0	mRNA metabolic process
16607	93	19	3.86	0	nuclear speck
30529	447	61	2.58	0	ribonucleoprotein complex
31974	1637	136	1.57	0	membrane-enclosed lumen
31980	208	31	2.81	0	mitochondrial lumen
32991	2342	194	1.56	0	macromolecular complex
33279	118	20	3.20	0	ribosomal subunit
34660	204	28	2.59	0	ncRNA metabolic process
43233	1607	135	1.59	0	organelle lumen
44267	2354	193	1.55	0	cellular protein metabolic process
44429	518	58	2.11	0	mitochondrial part
42773	45	10	4.20	1.00E-04	ATP synthesis coupled electron transport

Génomique Intégrative du coactivateur transcriptionnel PRC

L'avènement ces dernières années de technologies à haut débit comme les puces à ADN et plus récemment le « séquençage nouvelle génération » permet maintenant une analyse exhaustive du génome à différents niveaux de régulation parmi lesquels la mesure de l'expression de l'ensemble des gènes d'une cellule, les interactions ADN - facteurs de transcription, ou encore la méthylation de l'ADN. Il devient alors indispensable d'intégrer ces différentes données afin de pouvoir comprendre puis modéliser tous les mécanismes mis en jeu pour le fonctionnement global d'une cellule. Cette compréhension permettra alors de mieux appréhender les sources de dérégulation à l'origine de nombreuses pathologies.

L'objectif de mon travail de thèse est de comprendre l'impact du coactivateur transcriptionnel PRC (PGC-1 Related Coactivator) sur la coordination de la prolifération mitochondriale et cellulaire, à partir de cellules de la lignée tumorale humaine XTC.UC1, riche en mitochondries. Pour cela, PRC et les facteurs de transcription ERRα, NRF1, GABP, CREB et YY1 impliqués dans ces deux processus ont été étudiés par ChIP-chip. Les gènes positifs pour ces différents facteurs sont alors comparés au transcriptome des mêmes cellules sous l'effet d'un siRNA PRC.

L'intégration de ces différentes données de transcriptome et de ChIP-chip, couplées à des études de découverte de motifs, permettra ainsi de construire des réseaux de régulation transcriptionnelle nécessaires à la compréhension des voies d'action de PRC. Ces réseaux rendront possible la détection de cibles thérapeutiques permettant de corriger un état pathologique.

Mots clés : *génomique intégrative, ChIP-chip, transcriptome, découverte de motifs, PRC*

Integrative genomics of transcriptional coactivator PRC

Emergence over the last few years of high throughput technologies like DNA microarray and more recently next generation sequencing now allows an exhaustive analysis of genome with several regulation levels including expression measure of all cell genes, DNA – transcription factors interactions, and also DNA methylation. It thus becomes indispensable to integrate these different data to be able to understand and then to model all these mechanisms involved in global cell function. This understanding will then enable to apprehend the sources of deregulation involved in many pathologies.

The goal of my PhD is to understand the impact of PRC (PGC-1 Related Coactivator) transcriptional coactivator on the coordination of mitochondrial and cellular proliferation, from human tumor cell line XTC.UC1, rich in mitochondria. To do so, PRC and transcription factors ERRα, NRF1, GABP, CREB and YY1 involved in these two processes have been studied with ChIP-chip. Positive genes for these transcription factors are then compared with transcriptome of the same cells under siRNA PRC effects.

Integration of these different data of transcriptome and ChIP-chip, coupled with study of motifs discovery, will allow to construct the transcriptional regulatory networks necessary to the understanding of PRC pathways. These networks will make possible the detection of therapeutic targets allowing to correct pathological condition.

Keywords: *integrative genomics, ChIP-chip, transcriptome, motif discovery, PRC*

Solenne Carat

IRT-UN INSERM U915

8 quai moncoussu 44007 Nantes Cedex 1