UNIVERSITE DE NANTES
FACULTE DE MEDECINE

De la génomique fonctionnelle vers la génomique intégrative de pathologies humaines

THESE DE DOCTORAT

Ecole Doctorale Biologie Santé Nantes Angers
Discipline : Sciences de la Vie et de la Santé
Spécialité : Aspects Moléculaires et Cellulaires de la Biologie

Présentée et soutenue publiquement par

RAHARIJAONA Mahatsangy

Le 15 octobre 2009, devant le jury ci-dessous

Président du jury :
Rapporteurs : Pr Lucie KARAYAN – TAPON, Université de Poitiers
Pr Jacques VAN HELDEN, Université libre de Bruxelles

Examinateurs : Dr Thierry GRANGE, Université Paris 7
Dr Yannick JACQUES, Université de Nantes
Pr Jean MOSSER, Université de Rennes
Dr Gérard RAMSTEIN, Université de Nantes

Directeur de thèse : Dr HOULGATTE Rémi, Université de Nantes

Table des matières

Avant-propos

Ces travaux ont été effectués sous la direction de Rémi Houlgatte, responsable de l'équipe «Génomique et bioinformatique» au sein de l'unité INSERM U915 et de la plate-forme puce à ADN d'OUEST-genopole (biogenouest) à Nantes.
Le projet de recherche de l'équipe, initié en 1999 par Jean Léger et poursuivi depuis 2005 par Rémi Houlgatte, s'inscrit dans la médecine génomique.
L'équipe vise, grâce à des techniques de Génomiques (puces à ADN, bioinformatique…), à détecter de perturbations au niveau du transcriptome, à établir de marqueurs diagnostiques dans de nombreuses pathologies humaines, à comprendre et modéliser les mécanismes moléculaires sous-jacents et trouver des cibles thérapeutiques.
Nous nous intéressons à l'étude de cancers (lymphomes, cancers thyroïdiens), de pathologies musculaires (pathologies cardiaques et dystrophies musculaires) et du métabolisme. Et en tant que plateforme, notre équipe entretient également de nombreuses collaborations aussi bien au niveau régional que national.

Ce manuscrit présente différentes études en transcriptome auxquelles j'ai pris part. Elles incluaient des méta-analyses et l'analyse de données génomiques d'autres natures : un travail sur des motifs d'ADN, l'identification de variations en nombre de copies sur différents modèles. Ce manuscrit décrit enfin un travail préliminaire sur l'intégration de données transcriptome et d'identification de cibles de facteurs de transcription à l'échelle du génome par ChIP-chip, technique que j'ai contribuée à mettre en place au sein de notre laboratoire.

Introduction

1. Explosion de la génomique

Le terme « génome » a été utilisé pour la première fois en 1920 par H. Winkler (Winkler 1920). Composé des termes « GENes » et « ChromosOMEs », il décrivait l'ensemble des chromosomes et leurs gènes. Mais ce n'est qu'en 1986 que le terme « génomique », la discipline scientifique qui étudie le génome, a été proposé par T.H Roderick (McKusick and Ruddle 1987). Cette discipline est apparue dès l'initialisation du projet de séquençage du génome humain en 1990 (Human Genome Project, HGP). Ce projet s'est conclu en 2001, après près de quinze années de travail, par la publication d'un premier « brouillon » du génome humain par le HGP et par une société privée Celera Genomics® (Lander, Linton et al. 2001), (Venter, Adams et al. 2001)). Le travail de finition s'est achevé en avril 2003: une version complète et précise à 99,99% de la séquence du génome humain est aujourd'hui librement accessible ((Schmutz, Wheeler et al. 2004)). Grâce au progrès technique concomitant à ce projet, ce sont aujourd'hui les génomes de plusieurs centaines d'espèces (des bactéries aux mammifères) qui sont entièrement séquencés.

Le séquençage de génomes a naturellement contribué à l'explosion de la génomique. Le but de la génomique n'est pas uniquement de cartographier les différents éléments fonctionnels -gènes codants ou non, pseudogènes, éléments régulateurs… - d'un génome. Elle s'attache également à la façon dont un génome fonctionne, au mode d'action des gènes, à l'étude de l'ensemble des produits qu'il code dans un système biologique donné.

Comme l'information génomique est devenue disponible pour un grand nombre d'organismes, une nouvelle ère post-séquençage - désignée « post-génomique » - émerge dans laquelle une caractérisation moléculaire globale détaillée et objective d'un système biologique peut être réalisée. L'analyse post-génomique ou « post-séquençage » utilise le génome comme fondement pour la compréhension des fonctions des gènes, des processus cellulaires, des spécificités tissulaires en tenant compte de l'évolution, du développement ou de maladies. Des approches expérimentales ou bioinformatiques sont utilisées pour analyser un grand nombre de gènes ou de protéines. En pratique, des méthodes à grande échelle pour mesurer transcrits, protéines, mais aussi pour explorer/analyser l'état de l'ADN sont disponibles.

Toutes les cellules d'un organisme contiennent la même information génétique, mais n'expriment pas toutes les mêmes gènes. Francis Crick posa en 1958 (Crick 1958) la première pierre de ce qui est aujourd'hui la théorie centrale de la biologie moléculaire : selon elle, l'ADN se réplique, est transcrit en acide ribonucléique (ARN) pouvant, ou non, être à son tour traduit en protéines. Au sein de l'ensemble des ARN d'une cellule, on retrouve, en plus des ARNm, des ARN fonctionnels non traduits en protéines. Certains ont un rôle structural (ARNt, ARNr). D'autres (les petits ARN : miRNA, siRNA, shRNA) ont un rôle de régulation post transcriptionnelle ((Carninci, Yasuda et al. 2008)).

Afin de comprendre l'activité d'un génome, déterminer où, quand et comment sont exprimés les gènes est primordial. Pour cela on peut appréhender la génomique par ses différents niveaux d'analyse. On peut examiner l'ADN, support de l'information génétique, présent dans le noyau des cellules, les ARNm, reflet de l'expression des gènes, ou les protéines, les acteurs.

Le Génotypage ou l'étude physique du génome.

Des différences dans la séquence ou la structure primaire de l'ADN peuvent contribuer à la variation du phénotype. Ces variations génétiques au sein d'une même espèce peuvent prendre différentes formes, allant du changement d'un seul nucléotide aux larges anomalies chromosomiques. Elles incluent un certain nombre d'évènements moléculaires touchant la séquence primaire d'ADN : mutations, insertions, délétions, duplications, translocations, inversions… L'analyse de l'ADN permet de cataloguer et de comprendre la variabilité dans un génome et comment cette variabilité contribue aux traits physiologiques. L'ensemble des variations de structures et de séquences d'un génome peut influencer l'expression ou la fonction des gènes, être à l'origine de pathologies ou expliquer certaines susceptibilités à des maladies (Feuk, Carson et al. 2006).

L'épigénome et interactions ADN-protéines.

L'épigénétique s'intéresse aux mécanismes dynamiques, réversibles et transmissibles permettant de réguler l'expression des gènes, sans affecter la structure primaire de l'ADN. Les modifications épigénétiques impliquent la méthylation de l'ADN ou les modifications post-traductionnelles des

protéines qui organisent l'ADN en chromatine, les histones. Ces modifications sont connues pour réguler un large panel de processus physiologiques et pathologiques (Jones and Baylin 2002).

- Les histones sont les protéines autour desquelles s'enroulent l'ADN et qui participent à l'empaquetage de l'ADN dans le noyau, formant ainsi la chromatine. Les modifications post-traductionnelles des histones incluent phosphorylation, acétylation, méthylation, ubiquitination, ADP-ribosylation ou sumoylation de résidus spécifiques (Strahl and Allis 2000). Ces modifications modifient la compaction de la chromatine. Elles peuvent influencer l'expression des gènes en régulant l'accessibilité de l'ADN à des facteurs de transcription ou leur recrutement. Ces mécanismes restent toutefois assez complexes. Par exemple, l'acétylation de lysine est souvent corrélée à une activité transcriptionnelle alors qu'une méthylation de lysine peut avoir des effets différents sur la transcription selon la position de la lysine méthylée (Bannister and Kouzarides 2005) et (Kouzarides 2007) pour revue).

- La méthylation de l'ADN s'effectue par l'ajout d'un groupement méthyl, grâce à des méthyltransférases de l'ADN, au niveau des Cytosines chez la plupart des eucaryotes, et uniquement dans le contexte de dinucléotides CpG chez les mammifères (Bird 2002 pour revue). Ces dinucléotides, assez rares dans le génome (Egger, Liang et al. 2004) suivent une distribution non uniforme : il existe des domaines, appelés îlots CpG, où ce dinucléotide est plus fortement représenté. Les îlots CpG sont fréquemment retrouvés dans le promoteur et le premier exon des gènes (Antequera and Bird 1993). Lorsqu'un îlot CpG d'une région promotrice est méthylé, le gène situé en aval est généralement réprimé (Bird 2002) ; (Jaenisch and Bird 2003). Cette répression peut se faire par deux mécanismes. La méthylation de l'ADN peut empêcher la liaison de facteurs de transcription au niveau de la région promotrice. Cette méthylation peut également entrainer une modification des histones, rendant ainsi l'ADN inaccessible. Cette méthylation est impliquée notamment dans les processus de développement et de différentiation.

Un épigénome est la description de ces modifications au niveau de l'ensemble du génome, à un temps donné, dans des conditions données et dans un système biologique donné. L'étude permet d'avoir une distribution de ces modifications dans le génome. L'épigénome répond à des signaux environnementaux en modulant l'activité des gènes et sa variabilité peut être à l'origine de certaines maladies multifactorielles (Hatchwell and Greally 2007).

Interaction ADN-protéines. Les facteurs de transcription modulent également l'activité des gènes. Ce sont des protéines qui se lient généralement à une séquence bien précise de l'ADN et qui induisent, facilitent ou inactivent la transcription des gènes de diverses façons. Ils peuvent notamment agir au niveau de la condensation de la chromatine, de l'initiation de la transcription et de la phase d'élongation du transcrit. Déterminer la présence (ou non) d'un ou de plusieurs facteurs de transcription sur l'ADN peut renseigner sur l'activité transcriptionnelle des gènes.

Le transcriptome.

Pour comprendre la fonction d'un gène, le fait de savoir quand, où et dans quelle mesure un gène est exprimé est essentiel pour comprendre l'activité et le rôle biologique de sa protéine correspondante. Cependant, accéder à la fonction cellulaire et physiologique d'un gène nécessite de l'étudier dans le contexte de son interaction naturelle avec l'ensemble des autres gènes exprimés dans la cellule.

L'ensemble des transcrits dans un système biologique donné, dans des conditions biologiques précises, à un temps donné constitue le transcriptome. C'est un déterminant majeur du phénotype et des fonctions cellulaires. Il est dynamique et varie d'un type cellulaire à un autre, d'une condition biologique à une autre. L'intérêt d'une étude des transcrits à grande échelle est d'avoir une vue d'ensemble de l'expression du génome et de ses variations dans différentes conditions**.** L'objectif est de comprendre comment les gènes aboutissent ensemble à la fonction biologique impliquée dans l'état physiologique ou pathologique d'un organisme.

L'étude du transcriptome d'un échantillon biologique permet de caractériser des signatures propres à cet échantillon.

Le protéome.

La protéomique s'attache à l'étude de l'ensemble des protéines d'une cellule.. Elle comprend l'analyse de leur niveau d'expression, de leurs localisations cellulaires et de leurs modifications post-traductionnelles ou leurs interactions, essentiels à l'accomplissement de leurs fonctions.

2. Les Biopuces

La communauté scientifique avait rapidement pris conscience que les projets génome allaient révolutionner la biologie moderne. Les techniques de biologie moléculaire ont rapidement présenté des limites pour répondre aux besoins de l'ère « post-génomique ». Un des challenges de cette nouvelle ère était de concevoir des nouvelles stratégies capables d'exploiter au mieux les informations issues du séquençage du génome.

Les techniques d'analyse moléculaire à grande échelle récemment développées permettent de connaître le statut de plusieurs milliers de gènes ou protéines simultanément, dans un échantillon biologique, en une expérience. Une des techniques d'analyse à grande échelle la plus utilisée aujourd'hui réside sur l'utilisation des biopuces. Elles appartiennent à un ensemble de nouvelles techniques développées depuis quelques années à l'interface de nombreuses spécialités comme la biologie moléculaire, la chimie, l'informatique, l'électronique et la robotique. Elles permettent aujourd'hui de mesurer les variations du génome, du transcriptome, du protéome, de l'épigénome d'un échantillon biologique. Il est alors par exemple possible grâce à elles d'identifier des différences, des altérations voire des aberrations moléculaires entre une cellule ou un tissu non pathologique et une situation pathologique.

2.1 Principes généraux

Les biopuces permettent de connaître le statut de plusieurs milliers de gènes ou produits de gènes simultanément dans un échantillon biologique en une expérience. Elles sont constituées d'un réseau de sondes denses ordonnées et immobilisées sur un support solide, le plus souvent en nylon, en verre ou en silicium. Leur utilisation, qui dépend de la nature des sondes et du matériel analysé, couvre aujourd'hui un vaste domaine d'investigation.

Chaque sonde correspond à une molécule précise et identifiée. Ces sondes ou spots peuvent être de différentes natures. Dans leurs emplois les plus courants, elles peuvent être constituées par de l'ADN (puces à ADN (Schena, Shalon et al. 1995)), des protéines (MacBeath and Schreiber) ; (Zhu, Bilgin et al.) ; (Zhu, Klemic et al. 2000), des tissus (Tissue MicroArray, TMA (Kononen, Bubendorf et al. 1998). Nous allons décrire ces trois types de biopuces dans ce chapitre. Cependant, les biopuces peuvent également être constituées de cellules ((Ziauddin and Sabatini 2001) et de petites molécules (Shin, Cho et al. 2004) mais ne seront pas décrites ici.

2.2 Puces à ADN

Les premières biopuces développées ont été les puces (ou plateformes) à ADN, probablement les plus abouties aujourd'hui.

Le champ d'application des puces à ADN est vaste. Initialement développées comme outil de criblage de banque (Gress, Hoheisel et al. 1992), il est possible grâce aux puces à ADN de mesurer l'expression des gènes (Gress, Hoheisel et al. 1992; Schena, Shalon et al. 1995), de décrire physiquement un génome (Pollack, Perou et al. 1999), de définir le statut de méthylation de l'ADN (Huang, Perry et al. 1999; Yan, Chen et al. 2001), des histones (Bernstein, Humphrey et al. 2002), et d'identifier des intéractions ADN-protéines (Ren, Robert et al. 2000) ; (Iyer, Horak et al. 2001)**.** Elles peuvent participer à l'annotation d'un génome (taille ARN, (Hurowitz and Brown 2003) , à la détection des gènes (Penn, Rank et al. 2000). Elles sont également utilisées en reséquençage pour identifier des mutations (Kozal, Shah et al. 1996) ou l'effet des agents pathogènes (Lin, Wang et al. 2006).

La technique repose sur l'hybridation spécifique entre deux brins d'ADN de séquences complémentaires, selon le principe du Northern ou du Southern blot : l'ADN fixé sur la puce (ici la sonde) et l'ADN cible de l'échantillon que l'on cherche à mesurer. Cet ADN à doser est marqué par un isotope radioactif, une enzyme, un fluorochrome, ou de la biotine. Les technologies dites mono-canal peuvent utiliser la radioactivité (P^{33} par exemple) ou les fluorochromes (Cyanine3, Cy3 ou Cyanine5, Cy5 par exemple) pour marquer les cibles. Les technologies à double canaux utilisent deux fluorochromes différents (Cy3 et Cy5 par exemple) pour hybrider simultanément deux cibles sur une même puce de manière compétitive. Le signal obtenu sur chaque sonde de la puce est analysé grâce à un lecteur capable de mesurer la radioactivité ou de discriminer les deux fluorochromes. Il reflète l'abondance de l'ADN complémentaire dans l'ADN cible. Le système à deux couleurs offre une certaine flexibilité, puisque le deuxième échantillon peut correspondre soit à une condition expérimentale, à une condition « normale » ou à un « standard » (ARN commercial).

La particularité des puces à ADN réside dans la miniaturisation du procédé permettant des milliers d'hybridations simultanées en utilisant une moindre quantité de matériel génétique. Selon

la problématique, les sondes utilisées pour fabriquer ces puces à ADN peuvent être des produits de PCR (Schena, Shalon et al. 1995), des clones d'ADNc (Nguyen, Rocha et al. 1995) ou d'ADN génomique (cosmide, BAC … (Pinkel, Segraves et al. 1998)), ou des oligonucléotides de synthèse (Ramakrishnan, Dorris et al. 2002). De plus, chaque sonde doit être spécifique d'une région précise du génome. Elles doivent être conçues de telle manière à éviter les hybridations croisées indésirables.

Ces sondes peuvent être déposées par un robot par contact ou par jet-d'encre sur le support puis fixées de façon covalente sur ce dernier. Elles peuvent être également constituées d'oligonucléotides synthétisés *in situ* par jet d'encre, (Agilent : (Hughes, Mao et al. 2001) photolithographie (Affymetrix : (Lockhart, Dong et al. 1996)). Une technologie plus récente utilise un réseau ordonné de billes sur lesquelles ont été greffés des oligonucléotides (Gunderson, Kruglyak et al. 2004).

La nature de la sonde et la méthode de fabrication de la puce présentent chacune des avantages et des inconvénients. L'utilisation de clones est par exemple utile pour l'étude de génomes pour lesquels on a peu d'information de séquence. Elle exige néanmoins de nombreuses manipulations (gestion de banque, amplification par PCR nécessaire avant dépôt…) offrant autant de sources de contaminations et d'artefacts. Il est maintenant admis que l'emploi d'oligonucléotides constitue le meilleur choix pour des sondes. C'est notamment le choix des principales sociétés commercialisant des puces à ADN, comme Agilent, Affymetrix ou Roche-NimbleGen. Elles ont aussi opté pour la synthèse *in situ*, plus coûteuse, mais plus flexible, s'affranchissant des biais potentiels de spottage et proposant des puces de plus forte densité (et par conséquent susceptibles de recouvrir l'intégralité du génome d'un organisme). Ces sondes offrent une plus grande fiabilité, une meilleure spécificité grâce au contrôle du design des sondes, uniformisant notamment les concentrations et les propriétés d'hybridation (Tm : melting temperature) des spots. Leur emploi se limite cependant aux organismes dont on a une connaissance approfondie de la séquence du génome.

Grâce à près de quinze ans de progrès techniques, les puces à ADN ont atteint aujourd'hui une certaine maturité. Grâce à une densité plus importante, une seule puce peut par exemple contenir tous les gènes d'un génome, même le plus complexe. La miniaturisation et l'amplification des cibles (Ginsberg 2005) ; (Paris 2009) permettent d'analyser des échantillons avec des quantités limitées de matériel, comme des biopsies. L'existence de puces de plus en plus fiables, de moins

en moins onéreuses et de protocoles standardisés ont « vulgarisé » l'utilisation des puces à ADN. Enfin, de nombreux outils bio-informatiques ont été développés pour améliorer la gestion, le traitement, l'analyse et l'intégration de la masse de données générées.

2.2.1 Puces à ADN pour le Transcriptome

L'application la plus commune des puces à ADN est l'étude du profil d'expression des ARNm. La puissance des puces à ADN réside dans leur capacité à mesurer simultanément le niveau d'expression de milliers de gènes fournissant ainsi un instantané du transcriptome dans un système biologique et dans une condition donnée. Les techniques de séquençage d'EST à grande échelle (Adams, Soares et al. 1993), d'analyse en série de l'expression des gènes (SAGE) (Velculescu, Zhang et al. 1995), differential display (Liang and Pardee 1992) donnent également un profil transcriptionnel à grande échelle. Cependant, ces techniques sont relativement plus coûteuses et moins flexibles que la technologie de puce à ADN.

Les puces à ADN pour l'analyse du transcriptome ont été les premières biopuces développées, cela pour des raisons techniques (l'accessibilité relativement simple au transcriptome d'un système, la facilité de manipulation de l'ADN) et biologiques (le niveau d'expression d'un gène reflète sa fonction) (Schena, Shalon et al. 1995).

Les puces conçues pour mesurer l'expression des gènes sont généralement biaisées vers les gènes connus ou prédits codants pour des protéines. Des produits PCR issus de clone d'ADNc, des oligonucléotides -provenant d'EST (Expressed Sequence Tags, marqueur de séquence exprimée) par exemple- sont généralement employés comme sondes.

Pour une analyse « classique » du transcriptome, les sondes sont le plus souvent conçues dans la partie 3' des gènes. Depuis peu, des isoformes par épissage alternatif d'un gène peuvent néanmoins être distingués par des puces à exons, disponibles pour les génomes séquencés comme l'Homme, la Souris ou le Rat (Frey, Mohammad et al. 2005) ;(Clark, Schweitzer et al. 2007). Ces puces à ADN, à très haute densité contiennent plusieurs sondes par exon connu ou prédit.

Les cibles, l'ensemble des ARN de l'échantillon étudié, sont extraites, reverse transcrites en ADNc et marquées, puis hybridées sur la lame. Les signaux de la puce sont ensuite quantifiés par un logiciel informatique.

Enfin, notons que l'étude du niveau d'expression de petits ARN non-codants par les puces à ADN est possible malgré leur petite taille (Liu, Calin et al. 2004) ; (Yin, Zhao et al. 2008)) pour revue).

2.2.2 Puces à ADN pour le génotypage

Les puces à ADN permettent de détecter des altérations chromosomiques à l'échelle génomique avec une résolution très satisfaisante (Pinkel, Segraves et al. 1998).
Un ADN test et un ADN de référence (par exemple tumeur et normal) sont fragmentés, marqués par deux fluorochromes différents et co-hybridés sur la puce. Les intensités des 2 fluorochromes déterminent les CNV (Copy-number variation : variation du nombre de copies) relatifs entre l'ADN test et la référence, comprenant gain ou perte de régions génomiques spécifiques. Les différentes plateformes de CGH (Comparative Genomic Hybridization) diffèrent par la résolution de la détection de CNV. Les facteurs principaux affectant la résolution de la puce CGH sont le nombre, la taille et la distribution chromosomique des sondes. Par exemple, l'utilisation de BACs amplifiés par PCR facilite l'analyse du génome entier (Ishkanian, Malloff et al. 2004) à une résolution que l'on peut néanmoins améliorer. Une alternative peut être l'utilisation d'ADNc (Pollack, Perou et al. 1999) .

D'autre part, des mutations peuvent être identifiées par des puces de reséquençage (Kozal, Shah et al. 1996). La même stratégie est également utilisée pour analyser les points de polymorphisme (SNP, Single Nucleotide Polymorphism). Des SNP connus sont immobilisés sur la puce. Pour chaque SNP, des oligonucléotides correspondant aux différents variants connus sont déposés sur une puce qui sera hybridée avec l'ADN d'un individu à tester. Ces sondes sont petites et sensibles à un mismatch d'un nucléotide. L'intensité du signal est mesurée pour chaque allèle du SNP (Dutt and Beroukhim 2007).
Les essais de génotypage basés sur les puces à ADN ont été utilisés dans des analyses de liaison à grande échelle de marqueurs SNP associés à différentes maladies. Les puces de génotypage peuvent interroger jusqu'à 1 000 000 SNPs en parallèle, faisant d'elles l'outil le plus puissant pour le génotypage.

Cette méthode nécessite une amplification de l'ADN génomique de l'échantillon biologique.

2.2.3 Epigénomique et Interactions protéines ADN

L'utilisation de puces à ADN couplée à des techniques de biologie moléculaire permet l'étude de l'épigénome ainsi que l'identification des interactions ADN-protéines à grande échelle.

- Etudes des interactions protéines ADN
- ChIP-chip : L'immunoprécipitation de la chromatine (ChIP) est une approche pour identifier des interactions ADN-protéines *in vivo* dans des conditions physiologiques, applicable à tout type d'échantillons biologiques (cellules, tissus, organes) (Orlando and Paro 1993). Elle est constituée de 3 étapes : les protéines qui sont fixées sur l'ADN sont liées de manière covalente. Puis l'ADN est découpé en courts fragments par sonication ou digestion enzymatique. Un anticorps contre la protéine souhaitée est alors utilisé pour l'immunoprécipitation et pour récupérer ainsi le complexe protéine-ADN. Enfin l'ADN est purifié après réversion du couplage covalent. L'existence d'un anticorps spécifique est cruciale dans cette technique.
Cet ADN peut être cloné et analysé par PCR, par séquençage, par hybridation sur une puce à ADN (ChIP-Chip ; (Ren, Robert et al. 2000) ; (Iyer, Horak et al. 2001). Pour cette dernière technique, l'ADN est amplifié, marqué puis hybridé sur une puce à ADN contenant des régions régulatrices. Ces puces peuvent être constituées de sondes couvrant une région d'intérêt (promoteurs ((Odom, Zizlsperger et al. 2004)), îlots CpG, riches en régions promotrices (Antequera and Bird 1993)) ou de sondes couvrant partiellement ou totalement le génome.
L'approche la plus commune pour détecter l'ADN enrichi par l'immunoprécipitation se fait en technologie double-canaux par co-hybridation du ChIP et d'un contrôle (ADN génomique, par exemple) sur la puce. En technologie mono-canal, le ChIP et le contrôle sont hybridés sur deux puces différentes puis les résultats sont comparés.
L'analyse de cet ADN à l'aide de puces à ADN fournit l'ensemble des sites de liaison de la protéine d'intérêt sur la chromatine à l'échelle génomique. Bien que le ChIP-chip ait été d'abord utilisé pour localiser les sites de liaison de facteur de transcription sur l'ADN (Ren, Robert et al.

2000) ; (Iyer, Horak et al. 2001), une des applications consiste aussi à caractériser d'autres phénomènes de manière globale, comme les modifications d'histones (Bernstein, Humphrey et al. 2002). Dans ce dernier cas, les anticorps utilisés doivent être sensibles aux modifications post-traductionnelles des histones.

Une des principales limites de cette méthode est la nécessité d'un anticorps spécifique pour les protéines. Des alternatives, comme l'introduction d'un épitope fusionné à la protéine d'intérêt endogène (Lee, Rinaldi et al. 2002), l'identification par une DNA adenine methyltransferase (DamID) peuvent être envisagées. Cette dernière approche détecte les sites de liaison potentiels d'une protéine en analysant la méthylation de l'ADN au niveau des adénines dans les cellules exprimant la protéine d'intérêt fusionnée à une enzyme DNA adenine methyltransferase d'*Escherichia coli*. Les adénines se retrouvent méthylées au voisinage des sites de liaison de la protéine fusion. Ces sites sont alors isolés par action d'une enzyme de restriction et identifiés, comme en ChIP-chip, par hybridation sur puce à ADN (van Steensel and Henikoff 2000). Cependant, contrairement au ChIP-chip, l'introduction d'un épitope et le DamID ne permettent pas de travailler dans des conditions physiologiques.

D'autre part, l'utilisation d'une puce à ADN double brin permet également d'identifier *in vitro* l'ensemble des sites de liaisons à l'ADN d'une protéine purifiée (Bulyk, Gentalen et al. 1999).

- La détection des régions d'ADN méthylées peut être réalisée par l'utilisation d'enzymes de restriction sensibles à la méthylation, isolant ces régions génomiques, suivie de leur identification par puces à ADN (DMH, Differential methylation hybridation, (Ching, Maunakea et al. 2005)). Une autre méthode, similaire au ChIP-chip, implique l'immunoprecipitation de l'ADN méthylé (MeDIP, Methylated DNA ImmunoPrecipitation) avec un anticorps dirigé contre la cytosine méthylée (5'-methyl cytosine) couplée à une hybridation sur puce à ADN. Ces deux methodes emploient des puces contenant des clones d'îlots CpG (Huang, Perry et al. 1999) ou des sondes les couvrant, ou des BAC (Weber, Davies et al. 2005).

2.3 Biopuces pour l'étude du protéome : Puces à protéines

Le développement d'approches à grande échelle pour l'étude des protéines est récent. Le champ d'investigation du protéome est vaste, et son analyse peut être abordée de différentes manières : caractériser par exemple la population protéique d'un échantillon (« dite protéomique analytique ») ou mesurer l'activité biochimique de ces protéines (ou dite « protéomique fonctionnelle »). Les puces à protéines offrent la possibilité d'explorer de tels champs d'investigation. Ils impliquent l'utilisation de deux types de puces à protéines différentes (MacBeath 2002) et (Zhu and Snyder 2003) pour revue).

.

- Les puces à Anticorps sont le plus souvent utilisées pour la protéomique analytique. Elle consiste en un réseau dense d'anticorps immobilisés sur un support. Les protéines sont marquées et un profil d'expression protéique est établi pour l'échantillon analysé.
- Des puces fonctionnelles contiennent des protéines complètes. Le matériel analysé peut être des protéines, de l'ADN, ou des petites molécules. L'identification des interactions protéine-protéine, protéine-ADN, protéine-phospholipide ou protéine-petites molécules est alors possible.

Cependant la conception de telles puces soulève de nombreux challenges techniques par rapport aux puces à ADN, notamment pour la production et la purification des protéines.

2.4 TMA

Le tissue microarray (TMA, (Kononen, Bubendorf et al. 1998) est un autre type de biopuces. Son objectif est différent de celles décrites précédemment. Il consiste à observer le statut d'une molécule particulière à travers un très grand nombre de tissus d'origine multiple (pathologiques, non pathologiques par exemple) de façon simultanée. Les tissus à tester sont déposés de manière ordonnée sur un support solide. Les utilisations les plus courantes sont une immunohistochimie ou une hybridation *in situ* (Wang, Hewitt et al. 2006) ; (Huang, Chin et al. 2004) à grande échelle pour la découverte de biomarqueurs.

3. Analyse des données Génomiques issues de Biopuces

Les expériences sur biopuces génèrent de grandes quantités de données qu'il faut traiter et interpréter. Il est important de distinguer au sein de ces données les variations biologiques des variations techniques. Les variations biologiques sont intrinsèques à tous les organismes. Elles peuvent être influencées par des facteurs génétiques ou environnementaux. Ce sont celles qui sont principalement recherchées, alors que les variations techniques sont liées aux biais expérimentaux, indésirables, parasitant alors les signaux biologiques d'intérêt. Il existe des moyens permettant de minimiser ce dernier problème. Un plan expérimental, décrivant les procédures et conditions expérimentales pour répondre à une question biologique, doit être défini avec soin, en particulier pour les études en transcriptome (Yang and Speed 2002). Des méthodes de normalisation (Kreil and Russell 2005) adaptées à la technologie et aux données doivent rendre comparables les mesures des différentes puces. Enfin, des outils statistiques et bioinformatiques doivent aider à l'interprétation biologique des données traitées, répondant ainsi à la question biologique posée.
Ce chapitre se focalisera essentiellement sur l'analyse pour les puces à ADN.

3.1 Plan expérimental et Sources de variations.
((Churchill 2002) pour revue)

L'objectif du plan expérimental est de rendre l'analyse des données et l'interprétation des résultats aussi simples et aussi puissantes que possibles (Yang and Speed 2002). Il doit optimiser la qualité et la quantité de l'information obtenue.
La difficulté de distinguer les variabilités techniques, systématiques, et biologiques d'intérêt réside dans leurs origines multiples. Les biais expérimentaux induisant des variabilités techniques peuvent provenir de la qualité des lots de lames, des réactifs (efficacité du marquage…), des préparations (quantités de matériel hybridé…) ou de biais systématiques (mesure du signal…). Ils entrainent alors des imprécisions de mesures.
Un plan habituel consiste à répliquer systématiquement les mesures pour contrôler et appréhender ces variabilités.

Réplicats techniques. Les réplicats techniques peuvent correspondre au dépôt multiple d'une même sonde sur une même lame ou à l'hybridation de plusieurs lames avec les mêmes échantillons. Les réplicats techniques permettent de contrôler la qualité des mesures obtenues

Réplicats biologiques Réaliser des réplicats biologiques consiste à analyser le plus grand nombre d'échantillons possibles. Cela permet d'étudier les variations interindividuelles en offrant les critères d'indépendance maximaux entre les mesures. Cette approche est donc préférable puisqu'elle permet une généralisation des résultats expérimentaux lorsque beaucoup d'individus sont utilisés.

Diversification des sources d'échantillons. Une façon supplémentaire d'éviter ou de minimiser les effets des biais inattendus est de diversifier le traitement des échantillons (par exemple : jour d'expérience, lot de lame…), l'origine des échantillons biologiques (par exemple : âge, sexe des individus si on veut étudier l'effet d'un traitement sur une population). Cette diversité permet de répartir les variations indésirables de manière égale ente les groupes biologiques d'intérêt.

3.2 Traitements des données primaires

Une fois l'expérience sur puces réalisée, l'analyse des données implique plusieurs étapes. La première est la quantification des signaux des puces (**voir 3.2**). Puis ces données sont normalisées pour les rendre comparables entre elles et adéquates pour la troisième étape, les analyses statistiques et biologiques.

Des technologies mono-canal et double canaux ont été utilisées dans les travaux de cette thèse. Bien que le système d'acquisition des signaux soit différent, le raisonnement pour le traitement des données est similaire.

Normalisation. La normalisation est le terme utilisé pour décrire le processus d'élimination de variations indésirables dans une expérience d'hybridation mais aussi et entre les expériences mêmes. Idéalement, leurs sources pourraient être déterminées ou minimisées au niveau expérimental. Ce n'est souvent possible qu'en partie et, afin de réaliser la comparaison de différentes mesures, les différences restantes doivent être contrôlées par la normalisation. Les

méthodes employées doivent être adaptées à la plateforme utilisée ainsi qu'aux données analysées (Kreil and Russell 2005).

Une procédure assez couramment utilisée pour traiter les différences de marquage est une normalisation linéaire globale des intensités. Elle consiste à transformer les données de sorte que la médiane des intensités des log-ratios (mesure de l'hybridation relative généralement utilisée pour l'analyse) soit de zéro (Quackenbush 2002). Cette stratégie peut être tout aussi bien utilisée pour les technologies simple-canal, pour des comparaisons de mesures, que double-canaux.

La méthode lowess (Locally Weighted Scatterplot Smoothing) est adéquate si le biais lié au fluorochrome dépend de l'intensité de la mesure. Le principe est basé sur le découpage des données en fenêtres d'intensités de taille connue, suivie d'une somme de régressions locales pondérées. L'ensemble des régressions locales est ensuite lissé pour former la courbe d'ajustement des données. Une version de ce principe prend également en compte les biais de localisation sur la puce liés aux aiguilles lors du dépôt (Yang, Dudoit et al. 2002).

Ces méthodes de normalisation sont le plus utilisées. Mais il en existe bien d'autres comme par exemple l'utilisation de quantiles (Bolstad, Irizarry et al. 2003) afin de réduire l'écart de la variance des mesures entre les lames. Toutefois chaque méthode de normalisation repose sur un certain nombre d'hypothèses qui doivent être par conséquent appropriées à l'étude.

Filtrage. Enfin, les gènes non informatifs (par exemple faiblement exprimés proche du bruit de fond) peuvent être éliminés pour la suite de l'analyse.

3.3 Analyse statistiques des données

3.3.1 Pour le transcriptome

La première étape de la génomique fonctionnelle consiste à identifier les gènes dont les niveaux d'expression varient d'une manière importante lorsque l'on compare deux groupes d'échantillons ou plus. Cette notion d'expression différentielle est centrale à la recherche de gènes cibles pour le développement de nouvelles approches thérapeutiques. Le raisonnement est que les gènes dont l'expression est perturbée (augmentée ou diminuée) sont directement liés aux processus moléculaires qui sous-tendent la perturbation étudiée. Cependant, l'information qu'un gène n'est pas différentiellement exprimé entre deux situations biologiques constitue une information importante.

Les principaux objectifs des expériences de puces pour le transcriptome peuvent être le plus souvent classés dans l'une des catégories suivantes : la « comparaison de classes », la « découverte de classes » ou la « prédiction de classe ». Le but de la comparaison de classe est d'établir si des profils d'expression diffèrent entre des groupes pré-établis. Pour la découverte de classe, il s'agit d'identifier des groupes et des sous-groupes méconnus sur la base de leur profil d'expression. Le but de la prédiction de classe est de prédire un phénotype à partir d'un profil d'expression de gène (Miller, Long et al. 2002).
De nombreux méthodes et outils sont disponibles afin d'extraire ces types d'information à partir des données d'expression. C'est dans ce domaine, mais aussi dans l'interprétation biologique des résultats, que les efforts et progrès dans l'approche « transcriptome » ont été les plus importants ces dernières années. La stratégie d'utilisation de ces différentes approches est bien sûr dépendante de la finalité de la recherche, la règle étant souvent de combiner ces différentes méthodes selon un scénario adapté à chaque étude. Il n'existe aucun scénario universel qui permettrait d'extraire de manière exhaustive l'ensemble des connaissances d'un jeu de données ou qui pourrait être utilisé quelle que soit l'étude réalisée.

3.3.1.1. Découverte de classe Approches non supervisées

Un grand nombre de méthodes ont été développées pour la découverte de classe (Slonim 2002) pour revue). Ce sont des stratégies non-supervisées, objectives, car elles ne requièrent aucune

connaissance *a priori* et se fondent uniquement sur les données. La classification est basée sur la comparaison de profils d'expression des gènes ou des échantillons.

Une stratégie de classification possible est une stratégie de regroupement (« clustering »). L'intérêt des regroupements des gènes par profil d'expression commun a été démontré dès la fin des années 90. A partir d'une étude de cinétique sur *S. Cerevisiae*, M. Eisen a montré que les gènes présentant un profil d'expression génique similaire (« cluster » de gènes coexprimés) sont très souvent impliqués dans une même fonction biologique (Eisen, Spellman et al. 1998).

L'objectif des algorithmes de regroupement est de définir des groupes de gènes ou d'échantillons dont les éléments sont proches les uns des autres et distants des éléments n'appartenant pas au groupe. Par convention, la suite du chapitre décrira l'utilisation de ces méthodes pour le regroupement des gènes. Les mêmes méthodes sont également appliquées aux échantillons biologiques.
La définition de ces groupes est basée sur une métrique entre les profils d'expression. Elle peut mesurer des distances (distance euclidienne par exemple) ou des corrélations (Pearson, Spearman...).Parmi les méthodes de regroupement, les classifications hiérarchiques ascendantes et les méthodes par partitionnement sont les plus utilisées.
La classification hiérarchique, méthode utilisée pour la première fois par Eisen pour l'étude de profils d'expression est toujours populaire. Il s'agit d'une méthode agglomérative où les gènes sont ajoutés à un groupe selon la similarité de leur profil d'expression. Le principe est le suivant : après le choix d'une métrique, par exemple le coefficient de corrélation de Pearson ou la distance euclidienne, les gènes ou groupes de gènes sont agglomérés deux à deux en débutant par les éléments les plus proches. Deux gènes agglomérés deviennent alors un groupe indissociable et ainsi de suite. Des méthodes de calcul des distances inter-groupes doivent donc être définies pour pouvoir agglomérer deux groupes de gènes. Le lien moyen (distance moyenne entre toutes les paires d'objets de ces deux groupes) et le lien entre les centroïdes de chaque groupe sont les méthodes les plus utilisées. En plus de sa simplicité, le succès de la classification hiérarchique vient aussi du fait de la visualisation des résultats. Il s'agit d'un dendrogramme des gènes. La similarité des profils d'expression entre deux objets est représentée par la distance entre le sommet de l'arbre et le premier embranchement commun. Ce dendrogramme est accompagné

d'une carte thermique représentant les données d'expression réorganisées selon le résultat. (Acharya, Sparreboom et al.). L'avantage et l'inconvénient de cette technique est l'absence de définition préalable du nombre de groupe de gènes à identifier. L'intérêt majeur de cette méthode est toutefois de représenter l'ensemble du jeu de données à étudier sous la forme d'une image facile à interpréter.

La plupart des méthodes de partitionnement au contraire imposent la définition préalable du nombre de groupes de gènes attendus. Les méthodes des k-moyennes (k-means) (Tavazoie, Hughes et al. 1999) et SOM (Self-Organizing Maps, (Tamayo, Slonim et al. 1999)) sont les plus utilisées. L'objectif global est de minimiser la distance de chaque gène au centre du groupe auquel il appartient. Ces méthodes sont non déterministes c'est-à-dire qu'appliquées plusieurs fois de suite sur le même jeu de données elles ne produiront pas le même résultat. La détermination du nombre de groupes de gènes est essentielle pour permettre une analyse efficace mais ce nombre est en général inconnu. Une stratégie souvent choisie est de démarrer par une analyse hiérarchique pour évaluer le nombre de groupes de gène attendus puis de lancer ensuite des méthodes par partitionnement qui vont permettre une délimitation automatique des clusters. Il existe par ailleurs des algorithmes de partitionnement itératifs descendants n'imposant pas un nombre de groupes au préalable.

Les méthodes de regroupement ne sont pas les seuls moyens d'explorer les données. Une autre stratégie possible, plus complexe est une réduction de la dimension de l'espace de données (analyses factorielles). Les analyses factorielles reposent sur l'idée qu'une grande partie de la variation des données peut être expliquée par un petit nombre de variables transformées. Parmi elles, l'analyse en composante principale (Raychaudhuri, Stuart et al. 2000) est une technique assez utilisée. Elle consiste à la recherche d'axes qui contiennent la plus grande part de l'information contenue dans les données. L'ACP recherche des axes orthogonaux qui représentent les plus grandes variances dans les données.

Cependant, ces stratégies peuvent entrainer une perte d'information conséquente et l'interprétation biologique de ses résultats peut ne pas être évidente.

3.3.1.2. Analyse d'expression différentielle pour la comparaison de classes, approches supervisées

-Tests statistiques pour la détection de gènes différentiels

Dans cette stratégie, l'utilisation de méthodes statistiques est essentielle pour vérifier que les différences observées de manière quantitative ne sont pas liées au hasard ou au bruit expérimental. Toutefois, la structure même des données issues des études en transcriptome comporte des caractéristiques qui imposent une adaptation des stratégies statistiques classiques. Les matrices de données comportent la plupart du temps les valeurs d'expression de milliers de gènes (et donc autant de biomarqueurs potentiels) pour quelques dizaines d'échantillons. Chaque gène fait l'objet d'un test sous l'hypothèse nulle qu'il n'y a pas d'expression différentielle à un risque donné. La multiplicité des tests augmente la probabilité d'obtenir des faux-positifs. Ce constat a imposé le développement de procédures statistiques afin de mieux traiter en particulier la présence de faux positifs (Benjamini 1995)

Les méthodes proposées, trop nombreuses pour être détaillées ici, doivent dans l'idéal minimiser le nombre de faux-positifs (erreur de type I) et de faux-négatifs (erreur de type II). Elles reposent néanmoins sur le principe général suivant :

i) Calcul pour chaque gène d'une statistique de test permettant de classer les gènes du plus au moins différentiel ii) Comparaison de la statistique obtenue pour chaque gène à sa distribution pour l'hypothèse nulle (il n'y a pas de différence d'expression pour le gène considéré). Cette distribution est soit fixe (cas d'un test de student classique) ou recalculée pour chaque jeu de données par des méthodes de permutations aléatoires (bootstrapping) des échantillons biologiques entre les groupes testés. On en déduit ainsi la probabilité statistique pour le gène d'être différentiellement exprimé.

Les tests basés sur des permutations ne font pas d'hypothèse sur la distribution des données, ni sur la taille des groupes. Le profil d'expression d'un gène sert de modèle pour représenter toutes les valeurs d'expression possibles pour ce gène. On peut générer de nouvelles matrices simulées ayant gardé des groupes de même taille. En calculant les statistiques de tests pour toutes les matrices aléatoires, et en les comparant avec celles de la matrice originale, on peut déduire des p-values. Les méthodes basées sur les permutations sont très robustes pour le jeu de données considéré. Il faut donc s'assurer que le jeu de données soit représentatif des populations étudiées.

Dans le cas d'analyses multifactorielles, des méthodes plus complexes utilisant par exemple une approche bayesienne (Smyth 2004) sont souvent nécessaires.

- Tests multiples.

Afin de contrôler l'augmentation du nombre de faux positifs due à la multiplicité des tests effectués dans cette analyse gène par gène, des procédures de corrections, ajustant les p-values sont nécessaires. Une correction courante est la correction de Bonferroni et consiste à diviser la p-value par le nombre de tests. Ce choix est très conservateur car il diminue considérablement la liste de gènes différentiellement exprimés, induisant alors un taux important de faux-négatifs. Les méthodes de corrections du taux de faux-positifs, introduit par Benjamini et Hochberg sont moins stringentes et plus adaptées aux puces à ADN (Reiner, Yekutieli et al. 2003). Elles mesurent la proportion de faux-positifs dans les gènes détectés comme différentiellement exprimés. Par exemple, si 5000 gènes sont testés, alors le nombre de faux-positifs attendu pour un $p<0.001$ n'est pas plus grand que 5. S'il y a 100 gènes significatifs à ce seuil, alors la proportion attendue de faux-positifs n'est pas plus grand que 5%, ce qui n'est généralement pas un problème.

La méthode SAM (Significance analysis of Microarray) (Tusher, Tibshirani et al. 2001) est un exemple populaire d'analyse pour l'étude de gènes différentiels. Elle permet, sur la base d'un test t modifié et de l'utilisation de permutations aléatoires des échantillons, de sélectionner un groupe de gènes différentiels en contrôlant le taux de faux-positifs.

- Classification supervisée pour la prédiction de classe

Il existe d'autres approches supervisées basée sur les profils d'expression et n'incluant pas de tests multiples. Les méthodes de classification supervisée établissent des règles et un modèle de classification à partir d'un jeu de données connu et annoté (jeu d'apprentissage) afin de prédire la classification de nouveaux cas appartenant à un jeu de données tests. Elles permettent de mettre en évidence des gènes marqueurs d'une classe (classifieur) afin de rendre possible le diagnostic ou le pronostic sur la base de portraits moléculaires.

De nombreux algorithmes de classification existent. Parmi les plus connues et les plus performantes, citons : les k plus proches voisins (KNN, K Nearest Neighbor), dans lequel l'échantillon inconnu est associé à la classe qui possède les k échantillons qui lui sont le plus

similaires. L'analyse discriminante linéaire (LDA, Linear Discriminant Analysis) est une méthode paramétrique dans laquelle une droite ou un hyperplan est calculée afin de séparer au mieux deux ou plusieurs classes connues respectivement. Cette séparation est réalisée de telle sorte que la variation intra-classe soit minimale et que la variation inter-classe soit maximale. L'échantillon inconnu est alors positionné dans l'espace et associé à la classe dont il est le plus proche dans le plan ou dans l'espace.

3.3.1.3. Prédiction, validation des clusters

Des méthodes permettent d'évaluer la stabilité et la validité d'un cluster de gènes d'intérêt (discriminants par exemple un état pathologique d'un état non pathologique) ou d'un prédicteur (classifieur).

Stabilité d'un cluster : Datta et Datta (Datta and Datta 2003) proposent des indices de qualité basés sur la perturbation du jeu de données par retrait, au hasard, d'un échantillon du jeu de données (Leave-one-out). Ils suggèrent ainsi d'évaluer la proportion d'échantillons (ou gènes) mal classés après le retrait d'un échantillon.
Les méthodes par ré-échantillonnage aléatoire (Booststrap, Jackknife) sont également utilisables.

Validité d'un prédicteur. L'une des finalités des études en transcriptome est la création d'outils décisionnels (prédicteur) pour l'aide au diagnostic et au pronostic. Une méthode classique est le partitionnement aléatoire des données du cluster en un jeu de données d'apprentissage (2/3 des données est un bon compromis) et en un jeu de données test (le reste des données). La qualité du prédicteur est estimée par sa capacité de classer chaque échantillon du jeu test dans la bonne catégorie du jeu d'apprentissage. Les approches par ré-échantillonnage aléatoire (bootstrap, Jackknife) et le Leave-one-out (Dudoit and Fridlyand 2002), sont les plus communes.

3.3.2 Analyse en ChIP-chip et CGH

- ChIP-chip

Les données de puce d'expérience de ChIP-chip reflète l'enrichissement (défini par le log ratio du signal ChIP sur le signal contrôle) relatif de séquences spécifiques dans le ChIP comparé à un échantillon contrôle.

Le développement d'outils dédiés à l'analyse de résultats en ChIP-chip est récent compte tenu de la relative nouveauté de la technique.

L'identification des sondes positives en ChIP peut se faire après définition d'un seuil (Kirmizis, Bartley et al. 2004) ou par l'analyse des rangs des log-ratios (Iyer, Horak et al. 2001). D'autres stratégies plus complexes, incluant un modèle d'erreur (single-array error model (Ren, Robert et al. 2000) ou des modèles de mélanges (Martin-Magniette, Mary-Huard et al. 2008) peuvent aussi être employées.

Statistique

Contrairement au transcriptome, les mesures dérivées des expériences en ChIP-chip apparaissent comme le mélange de deux distributions. La première correspond à la population de fragments génomiques enrichis par le ChIP (pour une sonde donnée : log-ratios IP/input positif, l'input étant de l'ADN génomique), et la deuxième au reste de la population non enrichie, et représentant le bruit de fond. L'observation de la distribution des log-ratios est donc asymétrique, avec un biais sur les valeurs positives. La distribution des log-ratios négatifs est approximativement gaussienne, alors que les positifs exhibent une queue non gaussienne (Martin-Magniette, Mary-Huard et al. 2008). Des statistiques sur les valeurs négatives peuvent être réalisées pour déterminer la significativité des valeurs positives.

Tiling arrays. L'utilisation de puces pavantes (tiling microarrays) recouvrant entièrement ou partiellement le génome à très haute résolution (les puces affymetrix sont actuellement les plus informatives). Une séquence positive en ChIP correspond alors à un signal d'enrichissement d'un groupe de plusieurs sondes contigües successives sur le génome (couvrant donc cette séquence). De plus, la distribution de ce signal au sein de ce groupe de sondes ordonnées selon leur position sur le génome n'est pas aléatoire, mais forme un pic, centré sur le site de liaison et correspondant à l'empreinte de la protéine. De nombreux logiciels ont été développés pour identifier de tels pics. (Kim, Barrera et al. 2005) ; (Buck, Nobel et al. 2005).

-CGH array

Le but est d'identifier des régions génomiques amplifiées ou délétées dans un échantillon test par rapport à un contrôle. Les log-ratios sont ordonnés et plottés selon leur position physique sur le génome. Un des challenges réside dans la détection des points de cassures.

3.4 Extraction de l'information biologique de puces à ADN.

Un des problèmes clé des résultats de puces à ADN est d'extraire l'information biologique issue d'une groupe de gènes dont on connait le profil d'expression.

3.4.1 Annotations fonctionnelles des clusters

Eisen a montré que les gènes d'un cluster participaient à un même processus biologique. Une autre approche pour l'interprétation biologique consiste à annoter fonctionnellement les groupes de gènes grâce aux informations sur la fonction de ces derniers.

Ces informations, provenant de sources différentes sont nombreuses et hétérogènes. Leur organisation et leur standardisation par des approches systématiques sont rapidement devenues une nécessité afin de pouvoir les exploiter pleinement.

Le projet Gene Ontology (GO) a été initié afin d'intégrer et d'uniformiser de manière cohérente les descriptions des gènes et produits de gènes de différentes bases de données (Ashburner, Ball et al. 2000). Son objectif est d'établir un vocabulaire structuré, contrôlé (ontologie) décrivant les gènes ou produits de gènes selon le processus biologique, la localisation cellulaire ou les fonctions moléculaires. Ces trois ontologies sont structurées sous forme d'un graphe orienté acyclique ou DAG (Directed Acyclic graph). Ce DAG est un réseau où chaque nœud représente un terme GO, et chaque terme GO peut être l'enfant de un ou plusieurs parents. Le terme enfant est toujours plus spécifique que le ou les termes parents.

Les termes GO associés aux gènes ayant un même profil d'expression peuvent alors être comparés à l'ensemble des catégories fonctionnelles du génome étudié ou aux gènes présents sur la puce. Des outils d'interprétation de puces comme GOMiner (Zeeberg, Feng et al. 2003),

FatiGO (Al-Shahrour, Diaz-Uriarte et al. 2004) ou EASE (Hosack, Dennis et al. 2003) utilisent cette stratégie.

3.4.2 Méta-analyse

Les résultats peuvent également être comparés à d'autres résultats de puces déposés dans des bases de données publiques, comme GEO (Gene experssion Omnibus, (Edgar, Domrachev et al. 2002) ou ArrayExpress (Brazma, Parkinson et al. 2003). Ce travail de méta-analyse (Rhodes, Yu et al. 2004) a notamment été réalisé dans notre équipe sur les dystrophies musculaires (Madmuscle, http://cardioserve.nantes.inserm.fr/mad/madmuscle/, en cours de rédaction) et un travail similaire sur les lymphomes est actuellement en cours.

3.4.3 Microdissection virtuelle pour l'analyse de tissus

L'ARN extrait d'un tissu ou même d'un organe entier reflète l'expression d'une grande variété de types cellulaires. La microdissection virtuelle utilise les profils d'expression géniques des lignées cellulaires correspondant le plus aux types cellulaires composant le tissu à analyser (Shaffer, Rosenwald et al. 2001). Un groupe de gènes dont le niveau d'expression est clairement plus important dans une des lignées par rapport aux autres sera supposé être lié à un type cellulaire du tissu correspondant. Grâce à cette microdissection virtuelle, on peut éliminer les gènes non spécifiques à l'étude. Le point crucial est de disposer de lignées cellulaires de l'espèce, représentatives des types cellulaires étudiés.

3.4.4 Découverte de motifs

La co-expression d'un groupe de gènes peut s'expliquer par des mécanismes de régulation transcriptionnelle semblables impliquant des facteurs de transcription communs. Van Helden et al. et Roth et al.(Roth, Hughes et al. 1998; van Helden, Andre et al. 1998; Tavazoie, Hughes et al. 1999) ont montré que des groupes de gènes co-régulés présentaient des sites de liaison de facteurs de transcription significativement sur-représentés en amont de la séquence codante de ces gènes.

Le site de liaison d'un facteur donné est court (5 à 20 nucléotides) et dégénéré, c'est-à-dire qu'il peut accepter des mutations pour certaines positions. La découverte de sites de liaison de facteurs de transcription sur l'ADN repose sur la découverte d'un motif inconnu statistiquement sur-représenté dans un jeu de séquences régulatrices (promoteurs, enhancer, silencer) apparentées. Deux stratégies différentes de découverte peuvent être distinguées : les méthodes de comptage de mots exacts et des méthodes probabilistes pour des motifs consensus.

La première méthode implique généralement l'énumération exhaustive d'oligonucléotides de taille donnée (Waterman, Arratia et al. 1984; van Helden, Andre et al. 1998). Afin d'identifier des oligonucléotides statistiquement sur-représentés, la probabilité d'observer n occurrences est estimée par des modèles probabilistes (binomiale par exemple).

Les méthodes probabilistes utilisent des matrices poids-position ou PWM (Position Weight Matrix) (Bucher 1990) pour représenter le motif. Ces matrices prennent en compte la dégénérescence des sites de liaison en comptabilisant les fréquences de chaque base à chaque position du motif. C'est une représentation préférée aux séquences consensus car elles intègrent les caractéristiques quantitatives des sites et leur utilisation est moins gourmande en temps de calcul.

La plupart des méthodes probabilistes reposent sur deux techniques statistiques itératives d'optimisation: l'échantillonnage de Gibbs (Lawrence and Reilly 1990) et l'algorithme d'espérance-maximisation (EM, Expectation Maximisation, (Lawrence and Reilly 1990). Elles revoient des matrices issues de l'alignement de sites, chacune correspondant à un motif dégéné.

La stratégie de découverte de motifs est particulièrement adaptée aux résultats de ChIP-chip., puisque cette technique donne accès aux séquences directes du facteur étudié.

4. Apports des puces pour l'étude de pathologies

4.1 Cancer et transcriptome

La plupart des cancers peuvent être catégorisés selon des critères cliniques, histologiques ou moléculaires. Ces catégories permettent de définir leur agressivité (grade), permettant une prédiction de l'évolution de la maladie et une indication thérapeutique. Cependant, ces facteurs anatomopathologiques standards encore très employés aujourd'hui sont insuffisants pour rendre compte de l'hétérogénéité des cancers. Ils peuvent en effet mener à un mauvais diagnostic et conduire des patients à un traitement inadapté, inefficace, inutile, voire toxique traduisant l'existence de sous classes pronostiques non identifiées par ces approches classiques.

Les biopuces, et en particulier les puces à ADN ont contribué à une avancée considérable dans la compréhension de nombreuses pathologies, permettant une caractérisation moléculaire plus globale, détaillée et objective de la maladie. C'est dans le domaine du cancer où l'apport des puces à ADN est le plus évident. Cela est dû au fait que les cancers, du moins les tumeurs solides, s'accompagnent de remaniements géniques majeurs (accumulation d'aberrations génomiques et épigénomiques influençant l'expression des gènes : remaniements chromosomiques, amplifications, insertions, délétions, mutations, acétylation des histones, méthylation des CpG. .) qui ne peuvent être appréhendés dans leur globalité qu'avec des techniques à grande échelle.

L'analyse du transcriptome par les puces à ADN a ainsi pu montrer qu'il est possible de classer des tumeurs d'une façon au moins aussi efficace que l'anatomopathologie classique, au moyen de signatures caractéristiques. Golub et al. (Golub, Slonim et al. 1999) ont fait la distinction entre deux classes de leucémie uniquement sur la base de profils d'expression génique, tandis que Thieblemont et al. ont démontré que 3 sous-types de lymphomes non Hodgkinien ont des profils transcriptionnels spécifiques (Thieblemont, Nasser et al. 2004).

Classification : De plus, l'analyse du transcriptome permet de découvrir de nouvelles classes dans des groupes que l'on pensait homogènes sur des critères histo-cliniques ou dans des groupes d'échantillons inclassables. L'étude d'Alizadeh et al. (Alizadeh, Eisen et al. 2000) a mis en évidence grâce à une classification hiérarchique l'existence de deux sous-classes parmi les lymphomes B diffus à grandes cellules aux pronostics de survie différents, sur la base de leur données d'expression sur près de 18000 clones d'ADNc.. Ballester et al. (Ballester, Ramuz et al.

2006) ont identifié trois nouvelles sous-classes parmi les lymphomes T inclassables. De même, des profils d'expression caractéristiques de cancers du sein ont permis d'améliorer leur classification clinique (Sorlie, Perou et al. 2001); (Bertucci, Finetti et al. 2005). Des études ont également mis en évidence des signatures associées au pronostic (Bertucci, Nasser et al. 2002);(Loi, Haibe-Kains et al. 2008).

L'analyse du transcriptome permet donc une meilleure classification des cancers, traduisant des processus biologiques qui sont propres aux différentes sous classes. Ces études suggèrent le potentiel pronostique des puces en cancérologie. Elles ont défini des combinaisons de gènes dont les niveaux d'expression peuvent distinguer efficacement des sous-types cliniquement distincts, et exigeant des stratégies de traitements différents. Ces gènes peuvent par la suite être utilisés comme biomarqueurs (indicateurs biologiques) pour le pronostic et le diagnostic clinique ou pour l'évaluation de la réponse à un traitement. Les tissue microarrays (TMA) sont par exemple utilisés pour les valider au niveau transcriptionnel ou protéique (Ballester, Ramuz et al. 2006), (Hewitt 2006).L'intérêt des biomarqueurs est fondamental pour permettre la transition de la médecine actuelle globalement appliquée de manière identique à une cohorte de patients vers l'émergence d'une pratique médicale adaptée à chaque individu.

D'abord critiquées, avec notamment des doutes sur la validité ou la reproductibilité les puces à ADN (Tan, Downey et al. 2003) ont connu au fil du temps des progrès dans leur conception, l'expérimentation et dans les méthodes d'analyse bioinformatique des puces. L'étude MAQC (MicroArray Quality ControlMAQC) menée par la FDA (Food and Drug Administration) a démontré récemment la fiabilité des mesures générées par les puces (Shi, Reid et al. 2006). Cette étude a levé des barrières sur l'emploi des puces à des fins cliniques.

Des puces à ADN commerciales sont aujourd'hui disponibles pour prédire le devenir clinique du cancer du sein d'une patiente et ainsi aider le clinicien dans le choix du traitement. Le test MapQuant Dx™ (Ipsogen) mesure le niveau d'expression de 97 gènes pour reclasser des tumeurs de grade II en tumeur de faible (grade I) ou fort risque (grade III) en vue d'une prescription de chimiothérapie. Le test MammaPrint™(Agendia) est le premier essai diagnostic utilisant une puce à ADN approuvé par le FDA . Il permet d'aider dans le pronostic (grade et prolifération) de cancer de stade I ou II sans envahissement ganglionnaire chez les patientes de moins de 61 ans. Les 70 gènes de ce classifieur ont été initialement identifiés dans l'étude de Van't Veer et al. (van 't Veer, Dai et al. 2002) où les auteurs ont montré une corrélation entre l'expression des gènes et

le pronostic à cinq ans, puis revalidés par une étude indépendante (Buyse, Loi et al. 2006). Son pouvoir prédictif se révèle aussi performant que d'autres classifieurs caractérisés dans des études du même type, mais meilleur que des critères classiques (Haibe-Kains, Desmedt et al. 2008). Bien que cette signature ait été validée dans de nombreuses études rétrospectives, sa validation est actuellement en cours dans l'essai prospectif MINDACT (The Microarray in Node Negative and 0 to 3 Positive Lymph Node Disease May Avoid Chemotherapy Trial (Cardoso, Piccart-Gebhart et al. 2007).

4.2 Cancer et CGH

La technique de CGH (hybridation génomique comparative) array a fourni bon nombre d'informations sur les variations de nombres de copies dans le cancer et a également été utilisée dans la classification de cancer. Elle a permis la localisation d'altérations génétiques spécifiques qui sont par exemple associées au pronostic ou à la réponse à des traitements. Les gènes impliqués dans ces aberrations sont susceptibles de contribuer à la cancérogenèse, et l'utilisation de puces CGH haute résolution facilite l'identification de ces gènes associés au cancer. Par exemple, l'identification de petites délétions homozygotes peut suggérer la présence de gènes suppresseurs de tumeurs dans ces régions (Mestre-Escorihuela, Rubio-Moscardo et al. 2007), alors que des amplifications peuvent contenir des oncogènes (Pollack, Perou et al. 1999) ;(Andre, Job et al. 2009).

De telles analyses ont eu un fort impact dans la classification des divers stades des tumeurs et dans l'identification et la stratification de sous-groupes de patients présentant différents pronostics et manifestations cliniques.

Des profils distincts en CGH peuvent ainsi distinguer des sous-groupes différents au sein de cancers que l'on pensait homogènes. Carrasco et al (Carrasco, Tonon et al. 2006) ont pu séparer, sur la base de ces profils, des patients souffrant de myélomes multiples en sous-groupes ayant une mortalité différente. Des sous-classifications similaires ont également été reportées dans d'autres types de tumeurs. Dans le cas du cancer du sein, l'analyse de tels profils génétiques permet la distinction de sous-types spécifiques correspondant à ceux déjà découverts par l'étude du transcriptome (Bergamaschi, Kim et al. 2006). Wessels et al. ont identifié les formes héréditaires des formes sporadiques des cancers du sein (Wessels, van Welsem et al. 2002).

D'autres études ont établi des corrélations entre les profils génétiques et le pronostic : par exemple, les patients souffrant d'un cancer du sein sans envahissement ganglionnaire et avec une délétion 11q pourraient bénéficier d'une chimiothérapie (Climent, Dimitrow et al. 2007). Une amplification de 5q31-q35 est associée à un mauvais pronostic du cancer de l'ovaire (Birrer, Johnson et al. 2007). De plus, un nombre important de telles aberrations est en lien avec un mauvais pronostic, ce nombre augmentant avec la progression de la tumeur (Blaveri, Brewer et al. 2005); (Steinemann, Skawran et al. 2006). Enfin, des aberrations génétiques spécifiques découvertes par CGH array ont été associées à des réponses différentes à des traitements. Les pertes des régions 13q32.1 et 8p21.1 sont les marqueurs les plus fiables pour identifier des tumeurs ovariennes chimiorésistantes (Kim, Kim et al. 2007).

4.3 Cancer et methylome

Une méthylation aberrante de l'ADN des régions promotrices de gènes est une des caractéristiques communes à tous les types de cancer. De telles altérations peuvent provoquer une instabilité génomique (Chen, Pettersson et al. 1998). Une hyperméthylation des gènes suppresseurs de tumeurs ou une hypométhylation des oncogènes contribue également à la formation de tumeurs (Jones and Baylin 2002); (Herman and Baylin 2003) ;(Feinberg and Tycko 2004) ;(Esteller 2005)). Comme en transcriptome, l'étude des profils de méthylation de l'ADN permet de distinguer des sous-types de tumeur (Wei, Chen et al. 2002) ; (Adorjan, Distler et al. 2002). Ces profils peuvent être également utilisés pour déterminer le pronostic de la maladie. Martinez et al. (Martinez, Martin-Subero et al. 2009) ont montré récemment que le méthylome d'une tumeur cérébrale, le glioblastome multiforme, était hétérogène et ont proposé un groupe de gènes dont l'état de méthylation était associé à la survie.

4.3 Cancer et SNP

Récemment, des études d'association à l'échelle du génome ont été conduites dans le cancer du sein, de la prostate et dans le cancer colorectal par le génotypage de 200000 à 500000 SNP par

puces à ADN. Elles ont notamment mis en évidence des allèles comme facteur de risques dans le cancer du sein post-ménopausique (Hunter, Kraft et al. 2007), des loci de prédisposition dans la région 8q24 dans le cancer de la prostate (Yeager, Orr et al. 2007) et dans le cancer colorectal (Zanke, Greenwood et al. 2007), ou plusieurs autres loci gènes de prédisposition au cancer de la prostate (Thomas, Jacobs et al. 2008).
Ces études pourraient avoir des conséquences importantes dans le conseil, le suivi et le traitement des patients portant de tels variants alléliques.

4.4 Autres pathologies

Par l'approche originale qu'elles proposent, l'utilisation des puces s'est étendue à l'étude de nombreuses pathologies. Les profils moléculaires de patients en insuffisance cardiaque semblent pouvoir affiner les classifications cliniques (Liew and Dzau 2004); (Steenman, Lamirault et al. 2005). La mise en évidence de signatures spécifiques de pathologies neuromusculaires permet une classification moléculaire des échantillons (Sanoudou, Frieden et al. 2004) au moins aussi efficacement que leur classification sur de simples critères biochimiques et génétiques. Les puces ont permis une détection plus fine de CNV (Copy Number Variation, Variation du nombre de copies de l'ADN) dans l'autisme (Jacquemont, Sanlaville et al. 2006), (Sebat, Lakshmi et al. 2007) et l'association de SNP au diabète de type 2 (Frayling 2007), a permis d'identifier de nouveaux facteurs de risques pour cette maladie.

5. Génomique intégrative

Ces quinze dernières années, les biopuces, et en particulier les puces à ADN ont fait leurs preuves en tant qu'outil puissant dans la caractérisation moléculaire des mécanismes physiopathologiques des pathologies au niveau cellulaire ou tissulaire (hétérogénéité cellulaire).
Leur usage de plus en plus courant vient du fait de l'existence d'une multitude de plateformes permettant une exploration systématique de diverses propriétés génomiques. Les différentes approches pour extraire ces informations génomiques ont contribué individuellement à la compréhension de pathologies à un certain niveau de complexité. L'intégration de ces données génomiques hétérogènes dans un modèle plus global s'avère être une stratégie afin d'éclaircir les mécanismes moléculaires précis inhérents à ces pathologies et de prédire ainsi de nouvelles stratégies thérapeutiques.

5.1 CGH et transcriptome

Afin de déterminer l'impact des amplifications ou des délétions génomiques sur l'expression des gènes, des études ont combiné une approche de génotypage CGH avec l'analyse du transcriptome. En effet, les biais de l'expression de gènes dans des régions chromosomiques peuvent être dus à des changements de nombre de copies (Zhou, Luoh et al. 2003) ; (Crawley and Furge 2002).
Des études ont employé une telle stratégie. La distinction de trois sous-types de lymphomes B à grandes cellules par des profils d'expression (Staudt and Dave 2005) pour revue) et l'observation d'aberrations génétiques à différentes fréquences au sein de cette pathologie (Tagawa, Suguro et al. 2005) ; (Chen, Houldsworth et al. 2006)) ont poussé Lenz et al. ((Lenz, Wright et al. 2008) à combiner ces deux approches à forte résolution. Leurs résultats apportent des preuves génétiques que ces trois sous-types sont des maladies distinctes et utilisent des voies oncogéniques différentes

5.2 ChIP-chip et transcriptome/epigenomique et transcriptome, Etude de régulations transcriptionnelles

L'expression des gènes est contrôlée par des protéines régulatrices qui se fixent sur des séquences spécifiques de l'ADN et recrutent des cofacteurs. L'analyse du transcriptome seul permet de mettre en évidence des expressions particulières à une situation biologique donnée. Il peut, par exemple, s'agir des gènes sur- ou sous-exprimés dans des tumeurs de mauvais pronostic. Une hypothèse est que ces gènes sont sous le contrôle de facteurs de transcription communs. L'analyse des sites de liaison des facteurs de transcription sur le promoteur de ces gènes permet de comprendre des mécanismes de régulation conduisant au développement tumoral. L'analyse à grande échelle du transcriptome et du rôle des facteurs de transcription dans un système biologique donné ouvre même la voie à la modélisation statique ou dynamique de ce système, la biologie des systèmes. Ainsi, comme évoqué précédemment (paragraphe 4.4.4), la recherche de motifs surreprésentés dans les régions promotrices de ces gènes aide à la compréhension de leur co-régulation (van Helden, Andre et al. 1998; Tavazoie, Hughes et al. 1999), même si une telle stratégie ne fournit pas de preuves biologiques de leur fonctionnalité. Des outils comme RSA-tools permettent d'identifier des co-occurrences de motifs suggérant alors des intéractions entre différents facteurs. Des outils génomiques identifiant *in vivo* des interactions ADN-protéines comme le ChIP-chip peuvent être combinés avec des données de transcriptome. Carroll et al. (Carroll, Meyer et al. 2006) ont, pour la première fois dans des cellules humaines, associé des données de transcriptome et de ChIP-chip sur une lignée de cellules humaines de cancer de sein oestrogeno-dépendant, MCF-7. Leur travail consistait à mesurer l'expression des gènes et à identifier les cibles directes du récepteur aux œstrogènes après une stimulation aux œstrogènes. Ils ont mis en évidence des associations de motifs au niveau des sites de liaison du récepteur aux œstrogènes suggérant des interactions entre facteurs de transcription. Certaines coopérations étaient corrélées à des répressions de gènes. Scharer et al. (Scharer, McCabe et al. 2009) ont étudié le rôle que joue *SOX4* dans la progression du cancer. Ils ont identifié des cibles directes de *SOX4* par ChIP-chip et ils ont mesuré leur expression après la transfection dans leur système de *SOX4*. Ils ont ainsi pu identifier des réseaux de régulation et de signalisation de gènes impliqués dans différents processus biologiques et affectés par *SOX4*. De même, Shaffer et al. (Shaffer,

Emre et al. 2008) ont montré que les myélomes étaient liés à un réseau de régulation anormal contrôlé par IRF4.

5.3 SNP et transcriptome ou la génomique génétique

La génomique génétique permet d'identifier des régions génomiques intervenant dans la variabilité de l'expression des gènes. Elle vise à faciliter la découverte de gènes de susceptibilité dans des pathologies multifactorielles.
L'idée principale est d'utiliser les expressions géniques comme des caractères quantitatifs et de les combiner avec des marqueurs d'ADN afin de déterminer ces loci (expressed quantitative trait loci, eQTL) (Jansen and Nap 2001). Ceci permet de lier les polymorphismes directement à l'expression des gènes. L'ensemble des gènes affectés par un même polymorphisme participe à une même fonction biologique.

6. Projet de thèse

Depuis plusieurs années, notre équipe a fait de la génomique fonctionnelle sa spécialité par, notemment, l'utilisation de puces à ADN dédiées à l'étude du transcriptome. L'analyse du transcriptome seul permet de mettre en évidence des groupes de gènes d'expression corrélée et présentant un niveau d'expression typique d'une pathologie. Ces groupes de gènes sont susceptibles de subir des mécanismes communs de régulation transcriptionnelle, faisant alors intervenir une combinaison de facteurs de transcription communs. Dans la pathologie, cette régulation est perturbée. Le ChIP-chip permet d'identifier les cibles directes d'un facteur de transcription à l'échelle du génome. Comme présentée dans la partie 5.2, l'intégration de données de transcriptome et de ChIP-chip permettrait de caractériser ces mécanismes de régulation.
Mon travail avait initialement pour but de caractériser des mécanismes de régulation transcriptionnelle d'un système biologique en utilisant des données de ChIP-chip et de transcriptome.

La technologie de ChIP-chip n'étant pas disponible au laboratoire, j'ai contribué à sa mise en place au sein de la plateforme puce à ADN de Nantes. Pour cela, j'ai appris la technique d'immunoprécipitation de la chromatine sur culture cellulaire au sein de l'unité INSERM UMR599 à Marseille, dans l'équipe animée par le Dr. Jean Imbert. Comme nous voulions appliquer cette stratégie sur des tissus, j'ai également débuté les mises au point de ChIP sur tissu cardiaque humain. J'ai apporté le ChIP et suivi la mise en œuvre du ChIP-chip sur la plateforme. Toutes les mises au point pour cette technique lourde ont ensuite été réalisées par les ingénieurs de la plateforme.

J'ai dû me familiariser aux méthodes d'analyse de puce à ADN en transcriptome : du traitement primaire des données à l'analyse biologique des donneés. J'ai été impliqué à des degrès différents dans plusieurs études :
- 4 contributions « majeures »: étude sur le métabolisme du foie (étude 1) en collaboration avec l'unité U694 d'Angers ; étude des effets de mutations d'un Facteur de Transcription chez la drosophile (étude 3) en collaboration avec une équipe dijonnaise (UMR CNRS 5548) ; étude des mécanismes de cancérogénèse (étude 4) ; étude de l'effet d'un siRNA contre un cofacteur sur des lignées cellulaires dérivées d'oncocytome thyroïdien (étude 5).
- et des contributions « mineures » (Etude 2, annexes).

Je me suis également intéressé aux problèmes de découverte de motifs (Etude 3)

L'objectif final était d'intégrer les données de ChIP-chip, transcriptome, et découverte de motif pour caractériser des réseaux transcriptionnels (Etude 6). Malheureusement nous avons rencontré des difficultés pour obtenir des résultats fiables en ChIP-chip, et nous commençons juste à avoir les premiers résultats exploitables. Il ne m'a pas été possible de faire ce travail. Cependant les résultats générés au cours de cette thèse ont ou vont faire l'objet de plusieurs articles.

RESULTATS

1 TRANSCRIPTOME

Etude 1 : Restriction calorique dans le foie humain

En 2003 en France, l'obésité atteint 9,6% des adultes et 12% des enfants (enquête ObEpi 2003). A titre indicatif, le surpoids (Indice de Masse Corporelle (IMC) > 25) concerne 30% de la population adulte.

La perte de poids est un des objectifs de la prise en charge des patients obèses. Des stratégies thérapeutiques ou chirurgicales sont appliquées pour le traitement de l'obésité. Pour obtenir une réduction pondérale, la restriction alimentaire volontaire est une des stratégies utilisées.

La perte de poids peut être abordée sous la vision de la théorie énergétique où la variation de poids est expliquée par un changement primitif de la balance énergétique : 1) une réduction de la prise alimentaire génère un déficit énergétique. 2) Ce déficit est compensé par une mobilisation des réserves énergétiques.

En présence d'une perte de poids, la balance énergétique correspond à une situation où l'énergie métabolisable (apport alimentaire) est inférieure à la dépense énergétique totale.

Cependant quelles que soient les stratégies utilisées pour la perte de poids, la perte de poids n'est pas linéaire dans le temps. De plus, elles aboutissent toujours à un plateau pondéral qui correspond à une situation de balance énergétique équilibrée.

On peut expliquer cette balance par une diminution de la dépense énergétique. Dans les situations de restriction calorique intense, des études rapportent la présence d'un « coup de frein métabolique». Chez l'animal, des modifications du métabolisme énergétique sont également observées. La masse musculaire est épargnée, alors que la masse hépatique est réduite en valeur absolue et en proportion de la masse corporelle. Le foie est le siège des réactions métaboliques impliquant les glucides et les lipides et subit des modifications fonctionnelles importantes lors d'une restriction calorique. Les dépenses énergétiques du foie et du muscle représentent 50% de la dépense énergétique au repos et sont donc fortement impliqués dans un éventuel « coup de frein métabolique ». Une variation du rendement de la phosphorylation oxydative peut être sous tendue de ce coup de frein. Cette adaptation métabolique apparait chez les primates et l'homme après environ 10 jours de restriction calorique sévère.

Dans cette étude, nous souhaitons identifier les acteurs moléculaires impliqués dans ce « coup de frein métabolique », expliquant ainsi le plateau pondéral observé. L'étude inclut 20 patients obèses. 9 d'entre eux ont subi une restriction calorique sévère de 10 jours, les 11 autres étant les

contrôles, sous régime normal. Nous avons mené une étude du transcriptome du foie de ces 20 patients afin d'identifier les modifications d'expression des gènes en restriction calorique.

Après avoir détecté les gènes différentiellement exprimés entre les patients témoins et ceux en restriction calorique, nous proposons d'identifier les fonctions affectées par 4 stratégies différentes. Nous avons intégré à nos données 1) les informations sur les ontologies liées à chaque gène, 2) d'autres données de transcriptome (méta-analyse), 3) les informations focalisée sur des fonctions particulières et 4) des réseaux métaboliques.

Dans ce travail, j'ai réalisé l'ensemble de l'analyse de données de puce (acquisition des données ; traitements des données primaires : normalisation, filtrage ; détection de biais éventuels; analyse des données : clustering, tests statistiques, annotation fonctionnelle des clusters), sélectionné les gènes pour la validation par PCR quantitative et analysé ces résultats.
La procédure expérimentale des puces (Extraction d'ARN à partir des tissus de foie, amplification, marquage, hybridation sur puce à ADN, lavages, scan) a été réalisée par les ingénieurs de la plateforme.

Méthodologie

Informations sur les échantillons

Les échantillons utilisés proviennent de biopsies de foie. Chaque échantillon provient d'un individu unique.
2 groupes d'échantillons ont été utilisés : un groupe « contrôle » sous régime alimentaire normal et un groupe de patients obèses en restriction calorique.

Procédures expérimentales

20 000 oligonucléotides caractérisés (Ocimum Biosolutions,Hyderabad, India) ont été spotés sur des lames epoxy-silane à l'aide du Licidea Array Spotter (Amersham®).Cette puce « maison » est quasiment pan-génomique. L'extraction de l'ARN, l'amplification et le marquage de l'ADNc ont été réalisés suivant le protocole en ligne (http://cardioserve.nantes.inserm.fr/ptf-puce/).

Brièvement, 10mg d'ARN total de biopsies de foie ont été extraits avec RNeasy mini kit (Qiagen) selon les recommandations du fabricant. La qualité des ARN à été contrôlée avec le Bioanalyzer 2100 (Agilent) puis les ARN ont été purifiés (oligotex mRNA midi kit, Qiagen) et amplifiés (Amino Allyl MessageAmp™ II RNA amplification kit, Ambion). Pour chacun des 20 échantillons de foie, l'ARNm a été rétro-transcrit et marqué avec Cy3. L'écahntillon référence utilisées dans toutes les hybridations a été un pool de l'ADNc de tous les échantillons et marqué avec Cy3. Chaque ADNc cible a été mélangé avec une quantité identique d'ADNc références. Le mélange a été pré-incubé avec de l'ARNt de levure, et de l'ARN polyA puis hybridées sur la puce.

Les intensités des signaux ont été quantifiées avec le logiciel GenePix Pro 5.1 (Axon) après un scanner sur ScanArray Express II (Packard Bioscience, Billerica, MA, USA).

Analyse des données

Une procédure de normalisation Lowess a été réaliées pour corriger les biais techniques (Yang, Dudoit et al. 2002). Cette procédure a été appliquée canal par canal comme décrit par Workman et al. (Workman, Jensen et al. 2002). Pour chaque puce, les intensités du signal en Cy3 et en Cy5 ont été normalisées séparément par le profil médian de tous les signaux. Les gènes faiblement exprimés (ayant un singla proche du bruit de fond) ont été éliminés pour la suite de l'analyse.

Il est établi que les gènes appartenant à une même fonction biologique ou à un même type cellulaire ont une expression corrélée (Eisen, Spellman et al. 1998; Shaffer, Rosenwald et al. 2001). Afin de déterminer les fonctions biologiques affectées par une restriction calorique dans le foie, j'ai utilisé : 1) une méthode de détection de groupe de gènes corrélés, 2)Une méthode statistique pour détecter une expression différentielle entre le groupe contrôle et le groupe en restriction calorique.

Parmi les méthodes pouvant détecter des expressions corrélées, j'ai choisi la classification hierarchique qui a comme avantage de ne pas fixer *a priori* le nombre de clusters, ni la taille de ces clusters. Pour le clustering, j'ai utilisé le programme Cluster, en utilisant la corrélation de Pearson et l'utilisation de lien centroïde pour les distances inter-groupes. Nous avons utilisé le logiciel TreeView pour la visualisation des données (Eisen, Spellman et al. 1998).

La méthode statistique utilisée est le calcul d'une p-valeur issue d'un test de student. Comme le nombre de gènes attendus n'est pas connu, j'ai calculé le -log p moyenné sur une fenêtre glissante

selon l'ordre du clustering. Je n'ai pas effectué de corrections de multi-testing car l'identification des pics associés aux nœuds du dendrogramme de la classification des gènes a defini les différents clusters. Cette stratégie nous a permis d'identifier des pics de tailles optimales. Ces pics correspondent à des clusters de gènes d'expression corrélée et différentiellement exprimés Ces clusters contiennent des gènes d'expression corrélée et différentiellement exprimés entre les patients contrôles et ceux en restriction calorique. Cette stratégie est semblable à celle utilisée dans l'étude chez la drosophile présentée dans la partie « Transcriptome et Motif ».
Chaque cluster identifié a été repris individuellement et a été testé de manière globale afin d'évaluer son pouvoir discriminatif entre les classes. J'ai utilisé un chi² pour tester si la proportion des classes souhaitées dans chacun des deux clusters d'échantillons est dû au hasard. (H0 : la répartition des classes d'échantillons (restriction-contrôle) dans les clusters est aléatoire)

L'annotation fonctionnelle des différents clusters a été réalisée avec GoMiner et la base de données Gene Ontology. GoMiner determine des enrichissements significatifs de termes GO dans un cluster de gènes comparativement au reste de la puce par un test exact de Fisher.
J'ai également utilisé l'outil gene set enrichment analysis GSEA(Subramanian, Tamayo et al. 2005) pour évaluer si un jeu de gènes prédéfini (provenant de la même localisation chromosoique, participant à la même fonction,étant les cibles d'un facteur de transcription…) présente statistiquement des différences concordantes significatives entre deux états biologiques, à partir des données de puces. Tous les gènes sont classés selon leur expression différentielle entre les deux classes. Le logiciel évalue si la distribution des gènes du jeu testé dans cette liste ordonnée est due au hasard par permutations des classes.

Validation par RT PCR quantitative

Les analyses par PCR ont été réalisées par TaqMan Low Density Arrays (TLDA, Applied Biosystems) afin de valider les données issues des puces. Elles ont été menées sur 58 gènes uniques dans 8 échantillons (4 témoins, 4 contrôles). La majorité de ces gènes sont différentiellement exprimés entre les patients contrôle et ceux en restriction calorique dans notre étude. Nous avons utilisé l'ARN 16S pour la normalisation des données. Afin d'estimer la reproductibilité des données en puces et en RT-PCR, une corrélation de Pearson a été calculée pour chaque gène entre les deux types de mesure. Une classification hierarchique (en utilisant la

corrélation de Pearson) a été réalisée sur les données PCR et transcriptome afin de tester la classification des échantillons.

Résultats

Ce travail étant en cours de rédaction, voici les principaux résultats obtenus.

Nous avons identifié 10 849 gènes significativement exprimés dans la majorité des échantillons. Nous avons cherché des groupes de gènes d'expression corrélée - participant à une même fonction biologique - et différentiellement exprimés entre les patients en restriction calorique et ceux qui ne le sont pas (contrôles). Nous avons utilisé un test-t comme méthode de détection d'expression différentielle. Le logarithme de la p-value déterminée par test-t a été lissé sur le dendrogramme de la classification hiérarchique pour déterminer des groupes de gènes corrélés différentiellement exprimés entre les restreints et les contrôles, comme décrit précédemment (de Fraipont, El Atifi et al. 2005). Les résultats sont présentés dans la figure 1.

Trois clusters comportant des gènes différentiellement exprimés entre les patients témoins et les patients en restriction calorique ont été identifiés. Les clusters 1 et 2 comportent des gènes sous-exprimés chez les personnes en restriction calorique. Le cluster 3 comporte des gènes sur-exprimés chez ces mêmes personnes. Chaque cluster discrimine de manière significative chacune des deux classes testées (χ^2 test ou fisher-test). De plus, le profil médian représentatif du cluster est toujours significatif (t-test).

Afin d'identifier les fonctions biologiques affectées dans le foie lors d'une restriction calorique, nous avons intégré à nos données 1) les informations sur les ontologies liées à chaque gène, 2) d'autres données de transcriptome (méta-analyse), 3) les informations focalisées sur des fonctions particulières et 4) des réseaux métaboliques.

Annotations fonctionnelles basées sur les ontologies

La première stratégie consiste à identifier des termes significativement sur-(ou sous-)représentés au sein de chaque cluster, :. Les gènes impliqués dans la traduction (translation elongation, GO:0006414, cluster 1), dans la glycolyse et dans l'homéostasie du glucose (glucose homeostasis, GO:0042593 ; glycolysis, GO:0006096 cluster 2) sont sous exprimés chez les personnes en

restriction calorique. . Les gènes impliqués dans le métabolisme (Lipid Metabolic Process, GO:0006629 ; Tricarboxylic Acid Cycle, GO:0006099, ; Proteolysis GO:0006508) ou dans la réponse au stress oxydatif sont sur-exprimés (cluster 3). L'ensemble des fonctions sont représentées dans la table 1.

Méta-analyse

La deuxième stratégie consiste à comparer notre jeu de données à d'autres jeux de données de transcriptome disponibles dans les bases de données publiques. Nous avons utilisé un jeu « multi-tissus » afin d'identifier des spécificités tissulaires au sein de chaque cluster.

Nous avons ensuite utilisé un jeu de données de « multi-tissus» donnant le niveau d'expression de l'ensemble des gènes humains dans la plupart des tissus dans des conditions non pathologiques. Ces données mettent notamment en évidence des clusters de gènes spécifiques ou du moins surexprimés dans un tissu particulier. Nous avons étudié le niveau d'expression des gènes de chacun des trois clusters dans ce jeu de données. Ceci permet d'identifier parmi les gènes différentiellement exprimés entre les restreints et les contrôles des gènes fortement exprimés dans le foie ou d'autres tissus et les fonctions associées (figure 2). Le cluster 3 possède la signature de foie la plus nette. Parmi ces gènes de foie, on retrouve des gènes du métabolisme des lipides surexprimés chez les personnes en restriction calorique. Ce cluster présente également une signature de muscle associée au métabolisme énergétique.

GSEA

La troisième stratégie s'intéresse à des fonctions particulières que l'on veut tester. Nous avons constitué des groupes de gènes qui participent à une même fonction biologique : la glycolyse, la gluconeogenese, la beta oxydation des acides gras, le stress oxydatif, du cycle de Krebs.

La méthode GSEA (Gene Set Enrichment Analysis) permet d'identifier des groupes ayant une sur-représentation de gènes parmi les gènes différentiellement exprimés. Les gènes sont classés selon leur expression différentielle entre les deux conditions que l'on teste. L'hypothèse nulle de GSEA est que l'ordre des gènes du groupe testé est dû au hasard, l'hypothèse alternative étant

que cet ordre est associé au critère utilisé pour catégoriser les groupes d'individus. La significativité est déterminée par des tests de permutations des échantillons.

J'ai soumis ces groupes de gènes à GSEA afin de déterminer si un de ces groupes de gènes est enrichi en gènes sur exprimés dans une des 2 classes. Les gènes impliqués dans la beta oxidation sont sur exprimés chez les restreints (figure 3).

Réseaux Métaboliques

La quatrième méthode que j'ai utilisée est l'intégration des données transcriptome avec celles des voies métaboliques. L'idée est de trouver soit des gènes-clé des voies métabolismes affectés soit des voies métaboliques globalement affectées. Pour identifier à quel niveau les différentes voies métaboliques (Gluconeogenèse/Glycolyse, Cycle de Krebs, Cétogénèse et la voie d'entrée PPARalpha), nous avons placé les gènes différentiellement exprimés entre les patients contrôles et ceux en restriction sur les graphes KEGG (Kyoto Encyclopedia of Genes and Genome)(figure 4). La base de données KEGG décrit l'ensemble des réactions chimiques intervenant dans une voie métabolique donnée, et ce de plusieurs espèces, des Bactéries à l'Homme. Cette base détaille le(s) gène(s), l'enzyme(s), le(s) substrats et le(s) produit(s) impliqués dans chaque réaction.
Nous observons ainsi que la néoglucogénèse est activée par l'augmentation de la quantité de substrats glucoformateurs, notamment le glycérol provenant de la lipolyse, les acides aminés glucoformateurs (alanine, glutamine), le lactate. Il y a par ailleurs une augmentation de la synthèse et/ou de l'activité des enzymes clés de la néoglucogenèse (ENO1, ENO3) associée à la diminution de la synthèse et/ou de l'activité des enzymes clés de la glycolyse (ALDOA, PFKL).
Nous sommes dans une phase cétogènique (HMGCS et HMGCL surexprimés), les substrats sont principalement fournis par la lipolyse (ACSL, ACACB). Provenant de la lipolyse, les acides gras sont utilisés par tous les tissus. Les acides gras sont oxydés directement au niveau du foie.
La chaîne respiratoire mitochondriale est activée par la beta oxydation des acides gras et l'entrée des électrons par le complexe II (SDH) de la chaine est favorisée.
La régulation de cette beta oxydation semble être relayée essentiellement par la voie PPARalpha (activation des gènes cibles, APOAS, HMGCS, SCP..). La balance énergétique est équilibrée et

le métabolisme des lipides s'est adapté à un niveau élevé de la beta oxydation, tout en restant strictement contrôlée.

Les résultats observés en transcriptome sont validés par RT-PCR quantitative (figure 5).

Tables and Figures

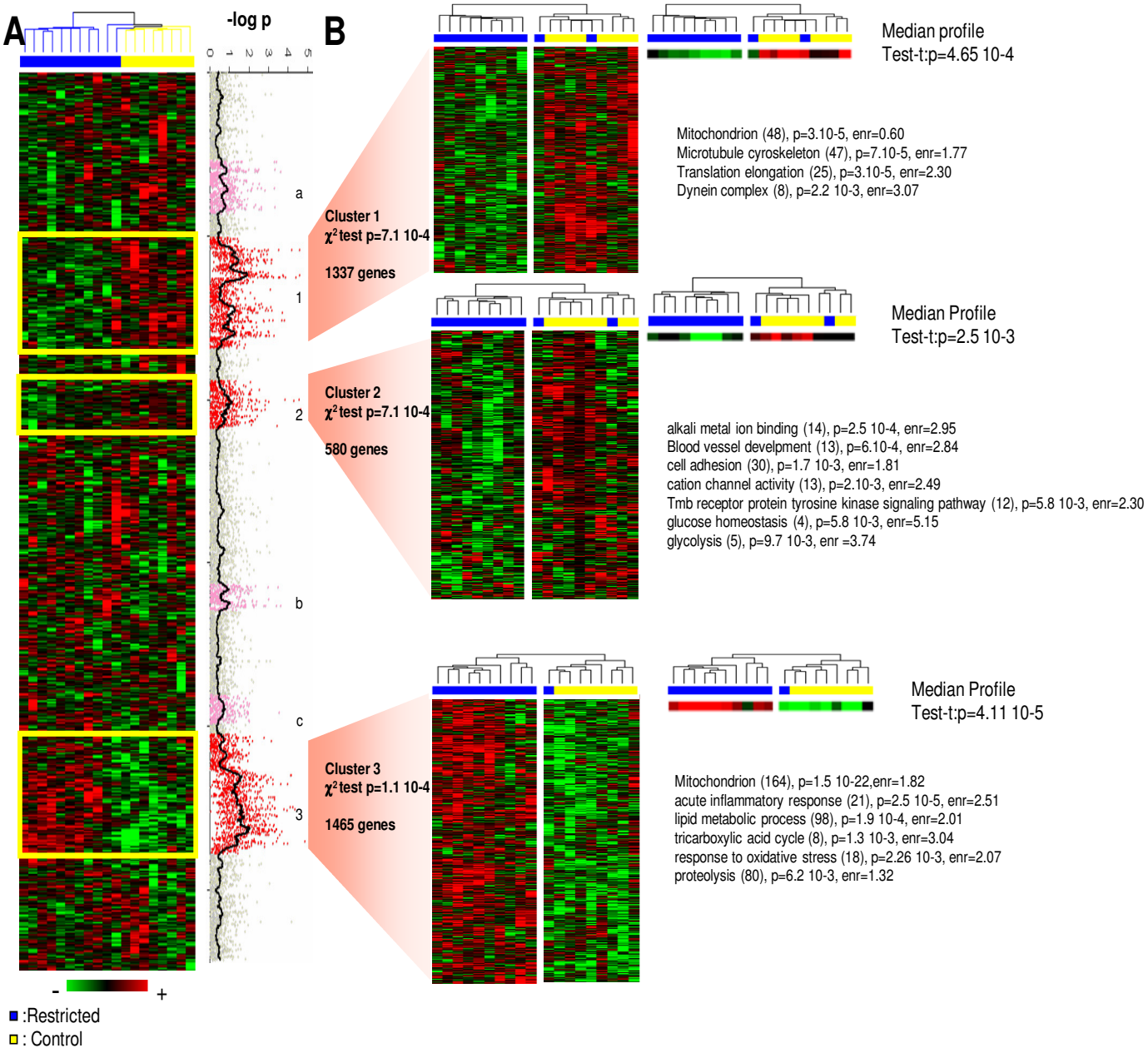


Figure 1: Clustering Hierarchique et identification des clusters

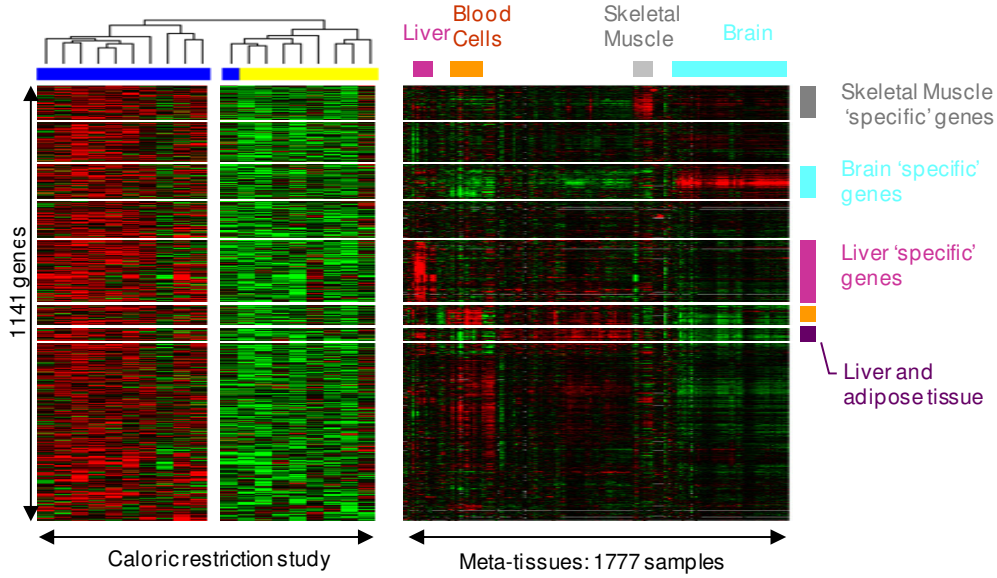
Mesure de l'expression des gènes dans le foie provenant de 11 patients en restriction calorique et de 9 patients contrôle.. **A**. Clustering Hierarchique des 10849 gènes exprimés (en ligne) et des échantillons « diet » et « contrôle » (en colonnes).). Le rouge indique un niveau d'expression supérieur à la valeur médianne dans l'ensemble des échantillons pour un gène donné, alors que le vert indique un niveau d'expression inférieur à cette valeur. **B**. Dot plot. Chaque point gris représente le –log p-value (en ordonnée) du gène dans l'ordre du clustering (en abscisse), le p-value étant calculé à partir d'un test-t entre les valeurs d'expression chez les restrients et les contrôles. La courbe noire correspondent à ces valeurs lissées sur 200 gènes. 5 clusters sont identifiés.. En rouge: terme(s) GO significatif(s) + χ^2 significatif; aucun terme GO significatif ou χ^2 non significatif. A droite, le profil médian de

chaque cluster et quelques termes GO sur- ou sous-représentés dans le cluster sont montrés.. Un enrichissement (enr.) <1 indique un terme GO sous-représenté.

Table 1: Termes GO significativement sur- ou sous- représenté pour chaque cluster

Cluster		GO terms (nb of genes)	P-value (<0.01)	enrichment
1 (underexpressed in Resctricted)	Cluster 1 vs all	Mitochondrion (48)	3.3 10-5	0.60
		microtubule cytoskeleton (47)	7 10-5	1.77
		Transcription (193)	3.7 10-4	1.25
		Translation elongation (25)	2.91 10-5	2.30
		dynein complex (8)	2.26 10-3	3.07
2	Cluster 2 vs all	alkali metal ion binding (14)	2.5 10-4	2.95
		blood vessel development (13)	6 10-4	2.84
		cell adhesion (30)	1.7 10-3	1.81
		cation channel activity (13)	2 10-3	2.49
		transmembrane receptor protein tyrosine kinase signaling pathway (12)	5.8 10-3	2.30
		glucose homeostasis (4)	6.4 10-3	5.15
		glycolysis (5)	9.76 10-3	3.74
3	Cluster 3 vs all	Mitochondrion (164)	1.5 10-22	1.82
		carboxylic acid metabolic process (93)	7.8 10-11	1.92
		oxidation reduction (94)	6.7 10-9	1.77
		transcription (120)	9 10-7	0.69
		peroxisome (25)	2.7 10-6	2.57
		coenzyme metabolic process (29)	7.8 10-6	2.28
		fatty acid metabolic process (37)	1.7 10-5	2.01
		acute inflammatory response (21)	2.5 10-5	2.51
		lipid metabolic process (98)	1.9 10-4	1.42
		amine catabolic process (18)	5.7 10-4	2.28
		complement activation (12)	4.8 10-4	2.77
		MHC protein complex (6)	8.2 10-4	2.93
		fatty acid oxidation (11)	9.5 10-4	2.73
		proteasomal protein catabolic process (20)	1.2 10-3	2.05
		antioxidant activity (11)	1.3 10-3	2.63
		tricarboxylic acid cycle (8)	2.1 10-3	3.04
		acetyl-CoA catabolic process (8)	2.1 10-3	3.04
		nervous system development (38)	2.2 10-3	0.64
		response to oxidative stress (18)	2.26 10-3	2.07
		carboxylic acid catabolic process (10)	3.7 10-3	2.48
		energy derivation by oxidation of organic compounds (19)	4.5 10-3	1.91
		proteolysis (80)	6.2 10-3	1.32

A



B

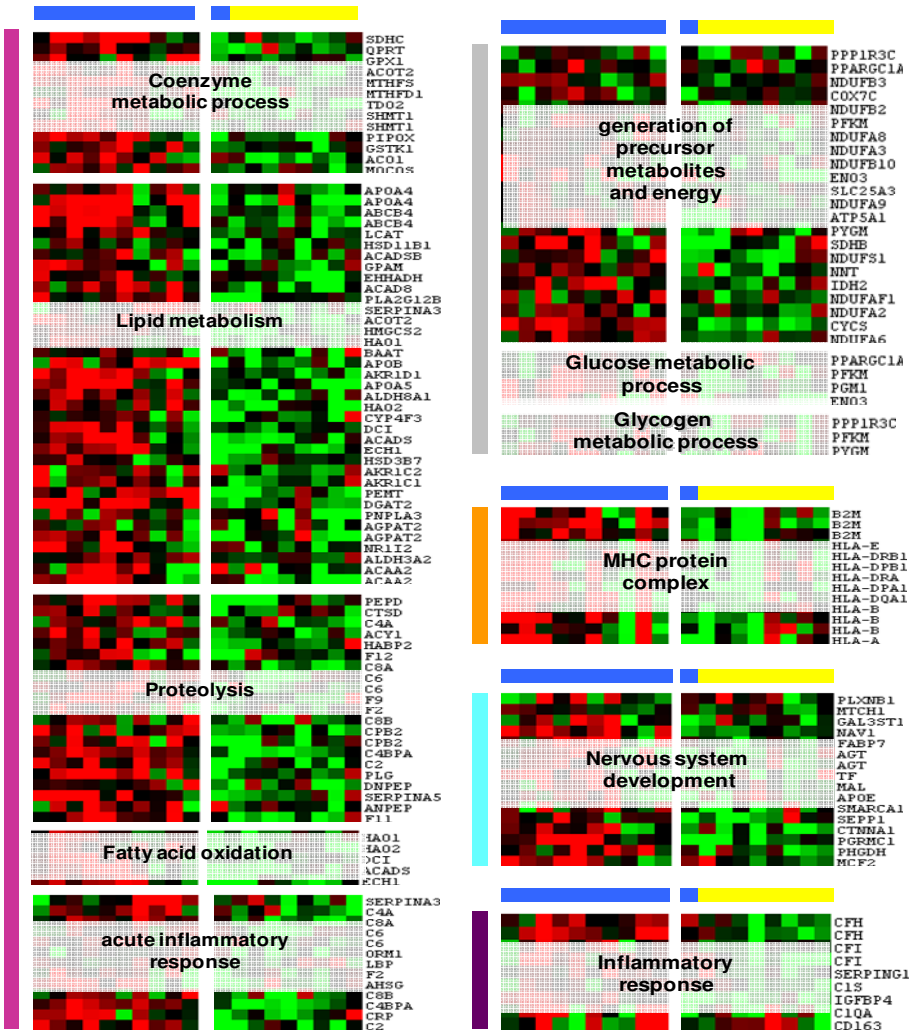


Figure 2: Microdissection virtuelle du cluster 3. A. Une matrice composée de plusieurs jeux de données transcriptome a été générée pour détedter des signatures tissu-spécifiques dans les gènes différentiels. Des signatures de Muscle squelettique (en gris), cerveau (bleu clair), cellules sanguines (orange), foie (violet), et foie + tissus adipeux (violet foncé) sont identifiés. **B**. Termes GO statistiquement sur-représentés pour chaque tissu.

Table 2: Termes GO significativement sur- ou sous-représentés pour chaque tissu dans l'analyse du cluster 3.

Tissue (nb of genes)	GO terms (nb of genes in multitissus data)	P-value (<0.01)	enrichment	Significant Go term in cluster functional annotation
Liver (158)	carboxylic acid metabolic process (43)	2.4 10-15	3.25	yes
	lipid metabolic process (33)	3.2 10-8	2.52	yes
	oxidation reduction (32)	1.7 10-7	2.42	yes
	acute inflammatory response (12)	3 10-6	4.08	yes
	gene expression (9)	1 10-5	0.33	yes
	Steroid metabolic process (12)	1.2 10-5	3.71	no
	lipid catabolic process (10)	1 10-4	3.58	no
	coenzyme metabolic process (12)	1.6 10-4	3.02	yes
	fatty acid metabolic process (13)	9.8 10-4	2.45	yes
	proteolysis (19)	5.2 10-3	1.82	yes
Skeletal muscle (80)	generation of precursor metabolites and energy (22)	1.3 10-17	8.21	yes
	oxidative phosphorylation (11)	2.2 10-12	12.32	no
	mitochondrial respiratory chain complex I (10)	3 10-12	13.45	no
	mitochondrion (35)	1.6 10-11	3.07	yes
	oxidation reduction (20)	2.2 10-6	2.99	yes
Brain (92)	nervous system development (15)	4 10-8	4.87	yes
	ion homeostasis (10)	1.4 10-5	6.47	no
Blood cells (51)	immune response (22)	8.14	5.9 10-17	no
	MHC protein complex (9)	21.09	5.9 10-13	yes
Adipose tissue+liver (28)	Acute inflammatory response (6)	5.4 10-4	11.52	yes
	inflammatory response (7)	1.2 10-5	7.91	no

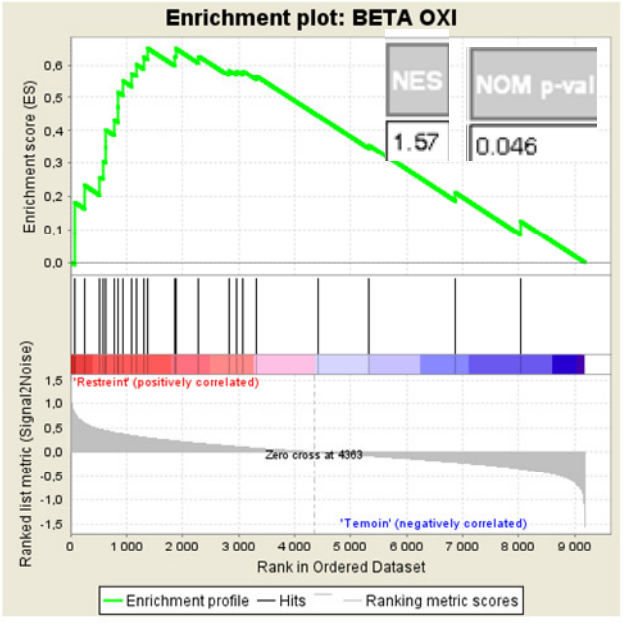
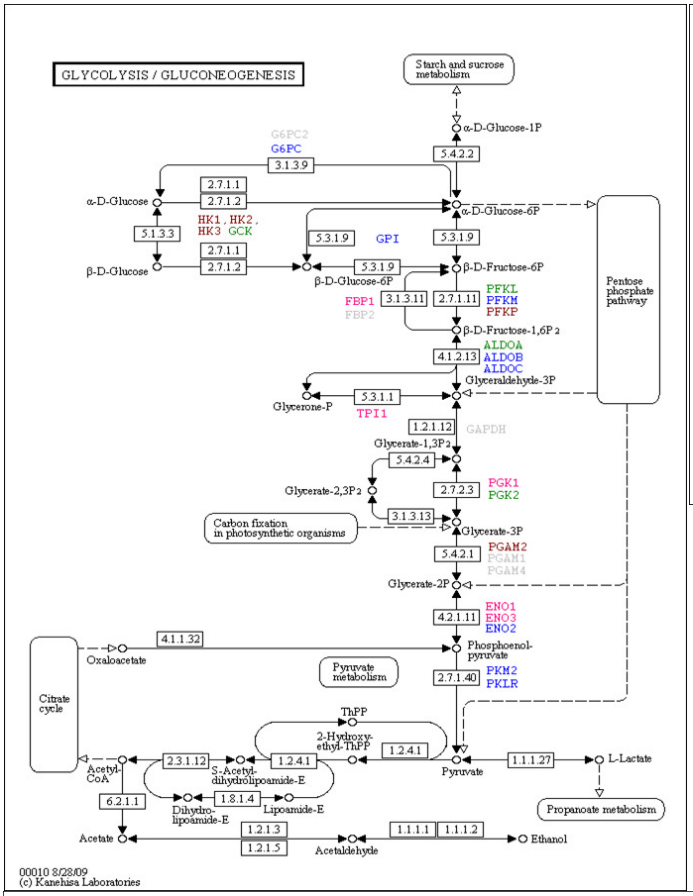
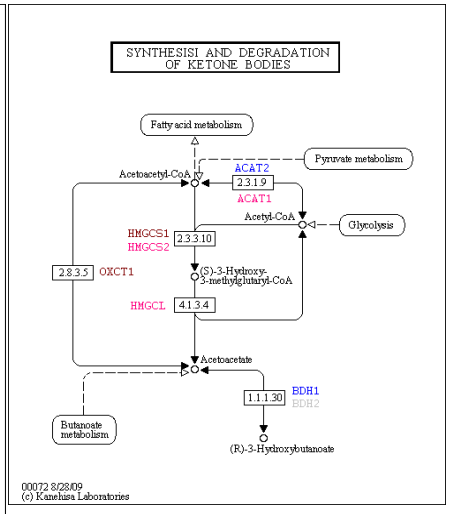


Figure 3: Analyse GSEA. Les gènes de la beta oxidation sont sur exprimés chez les restreints. (enrichissement 1.57, p=0.046)





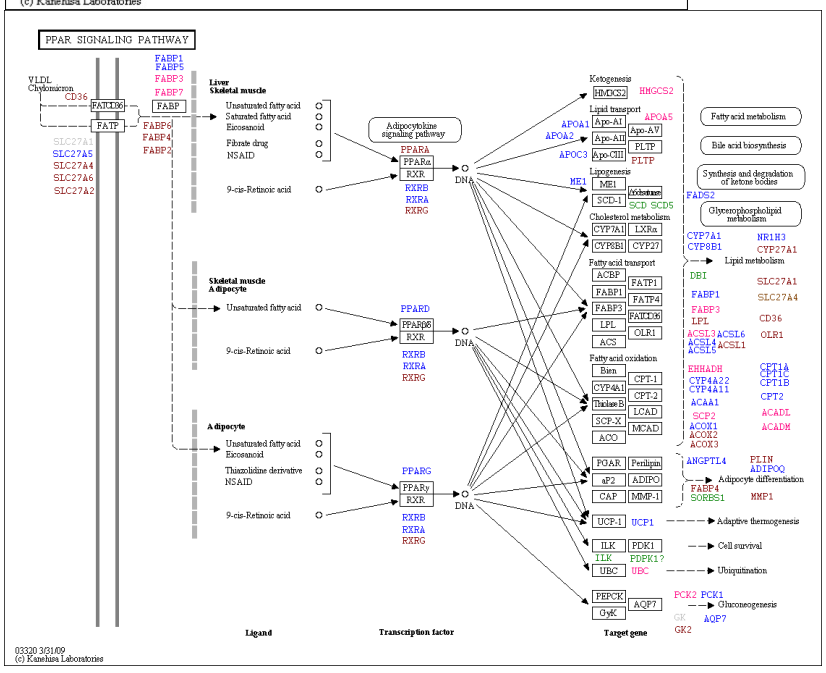
Gènes surexprimés chez les restreints

Gènes sous exprimés chez les restreints

Gènes exprimés non différentiels

Genes non exprimés

Gènes non mesurés



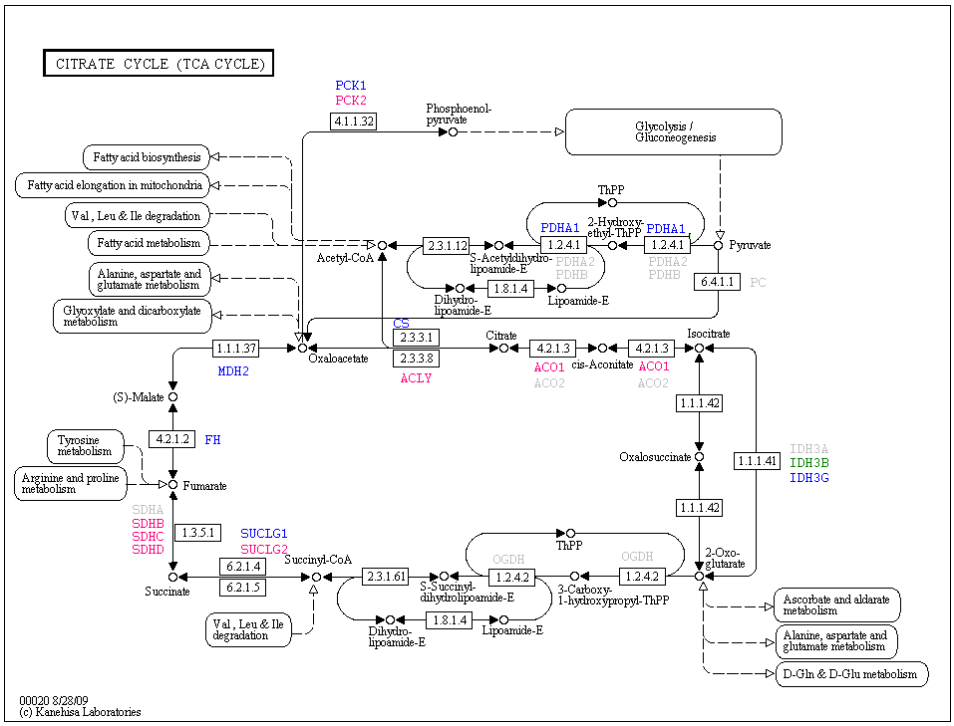


Figure 4: Graphe Kegg. Niveau d'expression des gènes impliqués dans les différentes voies métaboliques.En rouge sont représentés les gènes qui sont présents dans un des clusters et surexprimés chez les patients en restriction calorique. En vert, les gènes sous exprimés. En bleu, les gènes non différentiellement exprimés ; en marron les gènes non exprimés ; en gris les gènes non mesurés.

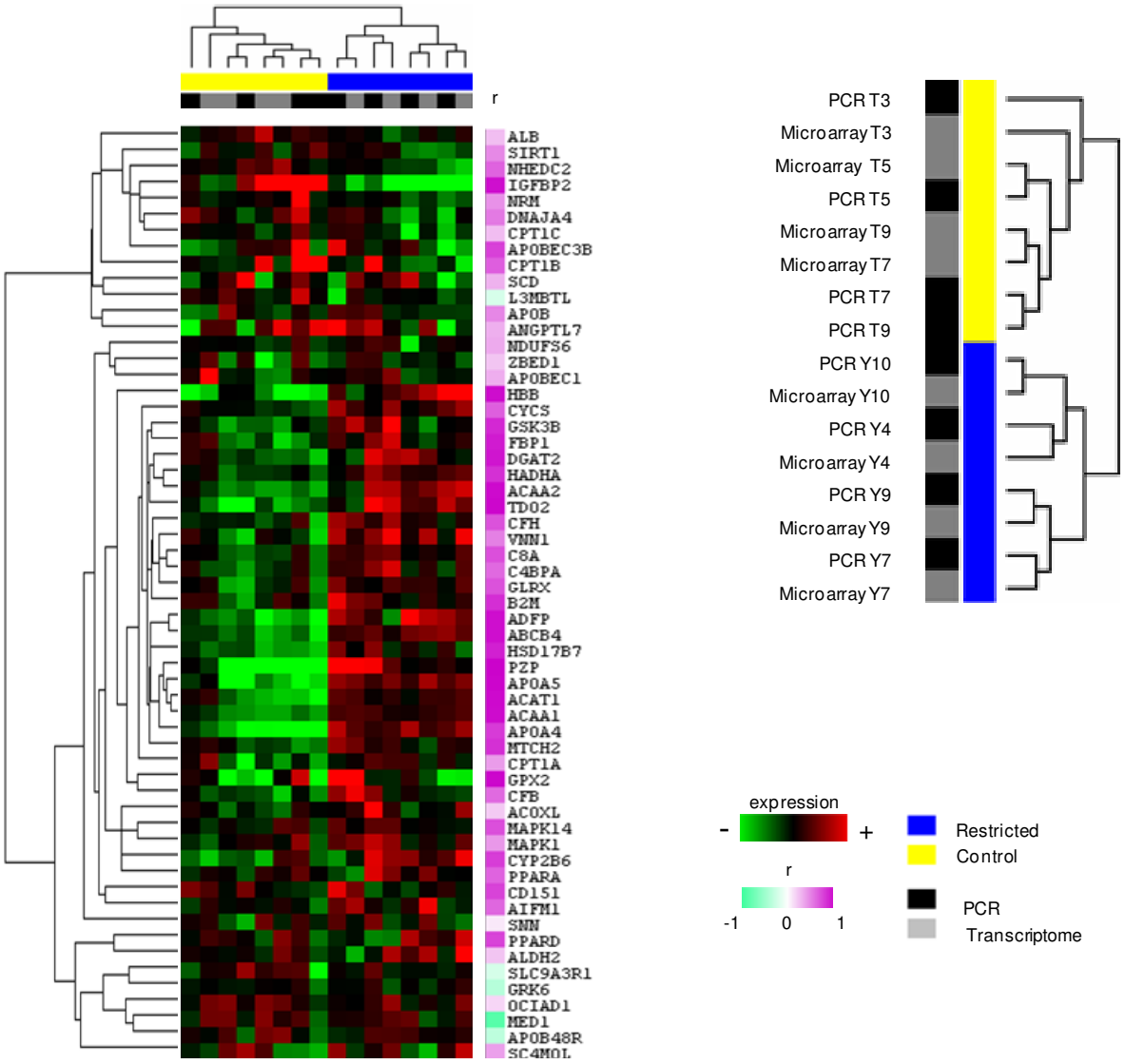


Figure 5: Validation des données de puces. Le clustering hierarchique est réalisé à partir de données de RT-PCR en temps réel et des données de puces de 58 gènes dans 8 echantillons différents (4 contrôles et 4 restreints). Pour chaque gène, la corrélation de Pearson entre les mesures PCR et les mesures des puces a été calculée et représentée à droite de la matrice. Les mesures du niveau d'expression des gènes en RT-PCR et en puces ont une très forte similarité (dendrogramme à gauche)

2. TRANSCRIPTOME + META-ANALYSE

Increasing the number of thyroid lesion classes in microarray analysis improves the relevance of diagnostic markers

Jean-Fred Fontaine, Delphine Mirebeau-Prunie, Mahatsangy Raharijaona, Brigitte Franc7, Stephane Triau, Patrice Rodien, Olivier Goëau-Brissonniére, Lucie Karayan-Tapon, Marielle Mello, Rémi Houlgatte, Yves Malthiery, Frédérique Savagner

Les nodules thyroïdiens sont relativement fréquents chez les adultes mais moins de 20% sont malins. Si le carcinome papillaire (PTC) est la plus fréquente des tumeurs malignes, le diagnostic du carcinome folliculaire (FTC) encapsulé à invasion minime est problématique à cause de ses similarités morphologiques et moléculaires avec les adénomes folliculaires thyroïdiens bénins (FTA). De plus, des caractéristiques atypiques ou oncocytaires rendent difficiles le diagnostic des différentes tumeurs folliculaires sur des critères histologiques, et nécessitent de nouveaux marqueurs moléculaires ou biologiques. Les études de puces à ADN sur les tumeurs thyroïdiennes comparent le plus souvent deux classes de tissus. Cependant, la pertinence des marqueurs identifiés dans ces études est limitée à la caractérisation de ces tumeurs des autres classes. Par exemple le gène CITED1, suggéré comme un des meilleurs marqueurs pour distinguer les PTC des tissus normaux n'est pas présent parmi les 42 meilleurs marqueurs PTC proposés par une méta-analyse incluant des tumeurs bénignes.

Ces méta-analyses pourraient alors augmenter le nombre de classes afin de définir des marqueurs plus pertinents. Cependant, elles pourraient aussi augmenter le déséquilibre du nombre d'échantillons par classes. Ces déséquilibres de classes influent sur l'identification de biomarqueurs et sur les procédures d'apprentissage des clusters et de prédiction de classes.

Griffith et al. ont validé des marqueurs de 21 études discriminant des tissus bénins des tissus malins. Cependant la plupart de ces marqueurs sont pertinents pour le diagnostic des PTC puisqu'ils représentaient plus de 40% des échantillons. Aucun classifieur n'a été défini pour des applications cliniques et cette étude n'avait pas pour objectif d'identifier des marqueurs spécifiques pour chaque pathologie.

Quelques travaux ont comparé simultanément plus de quatre types de pathologies thyroïdiennes. Il a été possible par exemple d'affiner le diagnostic des tumeurs de malignité incertaine par

l'analyse de huit types de tissus thyroïdiens. Ces résultats prometteurs sont encourageants pour la recherche de marqueurs pour la classification de tumeurs thyroïdiennes en tenant compte de plusieurs classes et sous-classes.

Un ensemble de biomarqueurs « idéaux » doit être capable de discriminer au mieux les classes et sous classes d'une pathologie (tumeur) donnée. Il est essentiel que, lors de l'identification de biomarqueurs spécifiques pertinents, toutes ou du moins le maximum de pathologies d'un tissu donné soit représentées.Afin d'améliorer la pertinence du diagnostic, nous avons analysé simultanément les données de transcriptome de 347 échantillons, représentant 11 sous-catégories histologiques de lésions folliculaires et des tissus thyroïdiens normaux, issus de 6 jeux de données publics. L'étude est basée sur notre jeu de données contenant 12 sous-classes représentées par 166 tissus thyroïdiens humains. Compte tenu des déséquilibres dans la réprésentation des différentes classes dans notre étude initiale, nous avons validé nos gènes prédicteurs de classe par l'analyse de 5 jeux de données indépendants traitant ces mêmes classes. Ces études reposent soit sur un nombre de classes moindre que notre étude (études de Giordano et de Weber), soit sur l'identification de gènes différentiels entre un tissu sain et une lésion particulière (études de Jarzab, de He et de Reyes). Ces cinq autres jeux ont été utilisés pour une validation croisée de la classification moléculaire, des classifieurs et de l'analyse fonctionnelle des gènes différentiels de notre jeu de données. Ces méta-analyses permettent de combiner les informations de plusieurs études apparentées mais indépendantes. Elles augmentent la fiabilité ainsi que la puissance statistique des analyses pour l'identification de gènes qui sont différentiellement exprimés. Elles peuvent suggérer une généralisation des résultats, indépendamment de la technologie utilisée ou de la localisation géographique des patients.

Dans ce travail, la méta-analyse a été faite par le premier auteur de l'étude. J'ai participé à l'analyse fonctionnelle des 104 gènes différentiels cross-validés. Il est établi que les gènes appartenant à une même fonction biologique ont une expression corrélée. Afin d'identifier des expressions corrélées, nous avons réalisé une classification hierarchique de ces 104 gènes avec nos données d'expression. 5 clusters ont été identifiés. Chaque cluster a été annoté fonctionnellement par enrichissements en termes Gene ontology.

Résultat

L'analyse non supervisée a montré une séparation des pathologies en trois groupes validés par une autre étude: des lésions bénignes et tissus normaux, des tumeurs malignes et une classe de tissus oncocytaires. Nous avons été capables d'identifier 220 gènes uniques pour la classification moléculaire des 11 sous-classes des pathologies thyroïdiennes folliculaires.

Nous montrons que les outils diagnostics définis à partir de données de puces à ADN sont plus pertinents quand un grand nombre d'échantillons et de classes sont utilisés. Prendre en compte les relations entre les pathologies peut améliorer l'efficacité du diagnostic des échantillons, en concordance avec les principales voies et fonctions biologiques sous-jacentes.

La définition des ces marqueurs spécifiques ont fait l'objet d'un brevet. A partir des ces marqueurs sera conçu un kit diagnostique basé sur le principe des puces à ADN. L'ensemble des ARN sera extrait d'une biopsie, et le niveau d'expression de ces marqueurs sera mesuré. Le profil d'expression de l'échantillon ainsi déterminé permettra d'identifier le type de pathologie dont souffre le patient, par comparaison avec le profil d'expression de chacune des 11 pathologies étudiées.

Title

Increasing the number of thyroid lesion classes in microarray analysis improves the relevance of diagnostic markers

Running Title:

Microarray-based diagnosis of thyroid lesions

Key-words:

Diagnosis accuracy, Microarray, classification, follicular thyroid pathologies

Authors

Jean-Fred Fontaine[1-3], Delphine Mirebeau-Prunier[2-4], Mahatsangy Raharijaona [5-6], Brigitte Franc[7], Stephane Triau[8], Patrice Rodien[2,3,9], Olivier Goëau-Brissonniére[10], Lucie Karayan-Tapon[11], Marielle Mello[12], Rémi Houlgatte[5-6], Yves Malthiery[2-4], Frédérique Savagner[2-5]

Corresponding author:

Frédérique Savagner, Inserm UMR 694, Laboratoire de Biochimie, CHU, 4 rue Larrey, 49033 Angers, France, tel : +33 241 35 33 14, Fax : +33 241 35 40 17, frederique.savagner@univ-angers.fr.

Affiliations

1 : Max Delbrück Center for Molecular Medicine, Berlin, Germany
2 : INSERM, UMR 694, Angers, F-49033, France
3 : Université d'Angers, Angers, F-49033, France
4 : CHU Angers, Laboratoire de Biochimie, Angers, F-49033, France
5 : INSERM, UMR 915, l'institut du thorax, Nantes, F-44035, France
6: Université de Nantes, Nantes, F-44035, France
7 : Hôpital A Paré, Laboratoire d'Anatomie Pathologique, Boulogne, F-92104, France
8 : CHU Angers, Laboratoire de Pathologie Cellulaire et Tissulaire, Angers, F-49033, France
9 : CHU Angers, Département Endocrinologie – Diabétologie - Nutrition, Angers, F-49033, France
10 : Hôpital A Paré, Service de Chirurgie vasculaire, Boulogne, F-92104, France
11 : Université de Poitiers, EA 3805, Poitiers, F-86021, France
12 : INSERM, UMR136, Marseille, F-13288, France

Abstract

Background

Many previous microarray studies compared few classes of thyroid lesions, therefore leading to the identification of marker genes with limited predictive accuracy. To improve diagnosis relevance, we simultaneously analyzed microarray data of 347 tissue samples representing 11 histological subcategories of follicular lesions and normal thyroid tissues from 6 public datasets. Our own dataset represented 48% of the samples and all the categories of lesions.

Methodology/principal findings

Classifier predictions were strongly affected by similarities between the classes and by the number of classes in the training sets. In each dataset, considering class similarities by separating the samples in 3 groups improved sample prediction. We cross validated differential genes showing 4 clusters with functional enrichments. Analyzing 6 of these genes (APOD, APOE, CLGN, CRABP1, SDHA and TIMP1) in 49 new samples showed consistent gene and protein profiles with the observed class similarities. Focusing on 4 follicular tumors subclasses, we explored the diagnosis potential of 12 selected markers (CASP10, CDH16, CLGN, CRABP1, HMGB2, ALPL2, ADAMTS2, CABIN1, ALDH1A3, USP13, NR2F2, KRTHB5) by real time quantitative RT-PCR on 32 other new samples. The gene expression profiles of follicular tumors were examined with reference to the mutational status of the Pax8-PPARγ, TSHR, GNAS and NRAS genes.

Conclusion/significance

We show that diagnosis tools defined from microarray data are more relevant when a large number of samples and tissue classes are used. Taking into account the relationships between the pathologies can improve the diagnosis accuracy of the samples, in concordance to the main involved biological functions and pathways. Our approach is particularly relevant for the classification of microfollicular adenomas.

Introduction

Over the last years, the use of microarray technologies has improved the identification of new markers for the diagnosis and prognosis of human tumors. Analyzing one tumor class and the normal tissue is a current design in cancer studies. In a meta-analysis comparing 40 cancer types from different tissues relative to their normal counterpart, the authors identified a common signature essential to carcinogenesis [1]. However, common markers derived from this analysis may fail to distinguish different tumor classes from the same organ that may yet have different prognoses. More, the signature is likely to be found in a variety of other cellular contexts like inflammatory processes.

Thyroid nodules are extremely common in adult population but less than 20% are malignant [2]. If papillary carcinoma (PTC)

is the most frequent malignant tumors, the diagnosis of minimal invasive follicular carcinoma (FTC) is problematic because of its morphological and molecular similarities to benign follicular thyroid adenoma (FTA) [3]. More, atypical or oncocytic features rendered the differential diagnosis of follicular tumors difficult on histology and ask for new molecular or biological markers [4,5]. Microarray studies on thyroid tumors mostly compared two classes of tissue [6–11]. Depending on the question, authors compared a class of thyroid tumor to normal tissue searching for markers specifically associated to this class or used the most frequent benign and malignant classes (usually the FTA and PTC classes) looking for markers of malignancy. However, the predictive accuracy of the markers identified is therefore limited to classify tumors from other classes. For example, the CITED1 gene that was proposed in a top list to distinguish PTC from normal tissue [9] was less specific when introducing data on FTC samples [8]. In addition, this gene was not present in the best 42 PTC marker genes proposed by a meta-analysis which included benign tumors [12]. Cross analyses may increase number of classes to define more relevant markers but may also increase the imbalance in the representation of some classes. In a previous meta-analysis, cross-validated marker genes from 21 studies differentiated benign from malignant tissues [13]. However, the majority of them are relevant for PTC diagnosis as they represented more than 40% of the samples. Nevertheless, no further classifiers were defined for clinical applications, and the study did not try to identify specific markers for the various individual pathologies.

Few studies have simultaneously compared more than 4 tissue types [5,14–17]. Yet, the predictive accuracy of selected genes has been increased by including several histotypes of FTC (minimal invasive, oncocytic variant) or subcategories of FTA. In the same way, we were able to refine the diagnosis of tumors of uncertain malignancy by the simultaneous analysis of 8 types of thyroid tissues [15]. These promising results encouraged continuing the research for relevant markers of thyroid tumor classification, taking into account more classes and some subclasses of interest. A recent review has emphasized this approach to strengthen the power of proposed markers [18]. In the present study, we have explored gene-expression signatures from the majority of the differentiated follicular pathologies of the thyroid tissue. We postulate that the most powerful meta-analysis to define tissue-specific markers and diagnostic tools, should simultaneously analyze several datasets containing all the pathologies originating from the same organ, including datasets with many histological subclasses.

Results

Molecular classification of thyroid pathologies

To analyze the global classification of the follicular thyroid lesions, we defined the centroid of each class of tissue by a composite signature of mean gene-expression levels. The hierarchical clustering of the class centroids in the Fontaine dataset (Figure 1A) showed a group composed of benign lesions and normal tissue (green boxes for FTA, MNG, WT, AT, and GD), another group of malignant lesions (red boxes for FTC, TUM and PTC), and a third group of oncocytic tumors and microfollicular adenomas (blue boxes for FTAb, OTUM, OTA and OTC). These groups were supported by a high robustness index of 0.938, and a discrepancy index of 0.337 and referred as benign, malignant and oncocytic groups. The 3 groups were cross validated in the Giordano dataset (Figure 1C) presenting almost perfect robustness (=1) and discrepancy indices (=0.001). This

dataset showed a highly correlated oncocytic group (OTA, OTC, correlation >0.6) and a group of malignant tumors (PTC, FTC+). The third group predominantly composed of benign tumors and normal tissues (WT, FTA), also included the FTC- group. The surprising proximity of the FTC and FTA classes has been already described [4]. This should be related to the high homogeneity in the expression profiles of the FTC+ samples, significantly related to the malignant profile of the PTC class. Then, clustering of the FTC- and FTA classes was more related to their heterogeneous expression profiles rather than to their proximity towards a tumoral behavior. This strongly suggests the presence of molecular subclasses for FTA and FTC- samples. We computed the prediction accuracy for the 3 groups of tissue classes defined above (Benign, Malignant and Oncocytic groups). The signature of each group of tissue was trained versus all the other samples, and its accuracy was compared to prediction accuracy of individual lesions (Table 1). In both datasets, the positive accuracy for each group was greater than the best accuracy reported for an individual lesion within this group. Then, considering class similarities by separating the samples in 3 groups was able to improve sample prediction in the 2 datasets.

The matrix of 66 pairwise correlation coefficients between the classes (Figure 1B) allowed a detailed observation of the similarities and dissimilarities. All the classes in the benign/normal group were positively correlated with each other. Some pairs of classes were highly correlated (correlation coefficient > 0.4), e.g. FTA with MNG, and AT with GD. All the classes in the malignant group were positively correlated with each other. The TUM class had a stronger correlation with PTC than with FTC (respectively 0.354 and 0.274). This similarity between TUM and PTC has previously been demonstrated within this dataset and with independent samples [15]. In the oncocytic group, the classes of oncocytic tumors were positively correlated with each other but this was not the case for the FTAb class. Whatever their degree of similarity, we were able to propose 220 non-redundant genes for the molecular classification of each of the 11 subclasses of follicular thyroid pathologies we explored (Table 2).

Class prediction from binary and complete trainings

We compared classifiers defined from many classes, i.e. the entire dataset (complete training), and two classes, i.e. one pathological class versus the normal tissue (binary training). Accuracy to predict the samples of a class, i.e. the positive accuracy, was always higher after a binary training than after a complete training in the two main datasets (Figure 2A). Considering the complete training, only two classes had similar performance for each dataset: PTC and AT for Fontaine, FTC+ and PTC for Giordano. The classifiers defined by a complete training from the Fontaine dataset were evaluated in 5 other thyroid datasets (Table 3). As compared to the positive accuracy computed from intra dataset cross-validations, the PTC, FTC, FTA and OTC classifiers were more accurate than expected. The OTA classifier had lower accuracy than expected (0.59 instead of 0.64) but only one other dataset contained this class of samples.

Functional analysis of cross-validated differential genes

From the 2 largest datasets containing the main kind of thyroid lesions [5,15], we searched for cross-validated differential genes. One hundred and four genes intersected the two lists of differential genes, separating five gene clusters from the hierarchical clustering of classes shared by the 2 datasets (Figure 3 and Table 4). The

identified functions and pathways mainly related to oncocytic and papillary tumors (clusters 1, 3 and 4). Gene ontologies are associated to oxido-reductive mechanisms (cluster 1, 28 genes), nitrogen metabolism and cell communication (cluster 3, 17 genes), glucocorticoid receptor signaling and leukocyte extravasation (cluster 4, 24 genes). A cluster of genes showed a specific profile for tumors with potential for malignancy where transcriptional factors involved in the primary anabolic and catabolic cell pathways are down-regulated (cluster 2, 25 genes). However in the cluster 2, the heterogeneous profiles for some classes (FTA) were suspected of molecular subgroups while the profiles for WT were almost homogeneous. A cluster of 10 genes (cluster 5) was not clearly specific to the tissue classes as their profiles were different in the two datasets. Referring to our data, we founded that 52% of the cross-validated genes belonged to two or more of the 11 classes of lesions (data not shown).

Independent validation of selected genes by real time RT-PCR

To test the relevance of some markers selected by cross validation from the two main datasets, we measured the expression level of 6 differential genes; CASP10, CDH16, CLGN, CRABP1, HMGB2 and ALPL2 on 32 new follicular tumor samples including micro and macrofollicular adenomas, oncocytic adenomas and minimal invasive follicular carcinomas (8 samples of each, Figure 4A). For all the 6 genes, the expression level was differential between follicular adenomas (macro and microfollicular) and follicular carcinomas (T- test, $p<0.05$). However, in all cases except one (CASP10, F-test, $p<0.05$), the gene expression level was not differential comparing macro to micro follicular adenomas but also comparing follicular adenomas to oncocytic adenomas as for CDH16 and CRABP1 genes. To test the relevance of our microarray study to identify subclasses of adenomas, we measured the expression level of 6 differential genes from the microfollicular adenomas subclass (ADAMTS2, CABIN1, ALDH13, USP13, NR2F2 and KRTHB5), on the same 32 follicular thyroid samples as above (Figure 4B). Except for NR2F2 ($p<0.954$), all genes were differentially expressed between the 3 subclasses of adenomas (T- test, $p<0.05$) but also through the 4 groups of follicular tumors (F-tests, $p<0.05$).

Mutational status of follicular tumors

Mutational status of 3 genes (TSHR, GNAS and NRAS) was explored on 59 follicular adenomas (43 from the microarray study and 16 from the RT-PCR study). Rearrangement status for PAX8-PPARG was explored on 11 FTC (3 from microarray and 8 from RT-PCR). For the follicular adenoma samples, we identified mutations in the TSHR gene in 7 cases (12% (7/59) of all the samples), 6 macrofollicular adenomas from the microarray study and 1 macrofollicular adenomas from the RT-PCR study. All mutations were situated in the seventh transmembrane domain of the TSH receptor (from codon 619 to 633). Our results are in accordance with the mutational status of the TSH receptor gene observed in the literature, especially concerning the macrofollicular subtype [19]. One macrofollicular adenoma presented a mutation in the codon 201 of the GNAS gene. None of the microfollicular adenoma showed TSHR and GNAS mutations in the four exons explored. Five microfollicular adenomas (4 from the microarray study and 1 from the RT-PCR study) were positive for NRAS mutation in the codon 61. For all the FTC samples studied, no PAX8-PPARG translocation was found.

Protein expression of selected differential genes in independent samples

We collected 49 new tissue samples to measure the protein expression of 6 selected differential genes in the common classes of the 2 main datasets (FTA, FTC, OTA, OTC, PTC and WT) and also for the AT class as a non-tumoral control (Figure 5). Significant differential expression through the classes (F-tests, $p<0.05$) was found for 5 proteins (SDHA, CLGN, APOD, CRABP1 and TIMP1). The APOE protein was not differentially expressed ($p<0.932$). Among the differentially expressed proteins, 3 showed similar profiles with their gene profiles (SDHA, APOD and TIMP1), but 2 proteins displayed some differences (CLGN and CRABP1), in accordance with previous studies of selected classes [20,21]. Expression of the CLGN protein was increased in FTA and FTC samples but not in oncocytic samples, contrary to that observed for its gene profile. Expression of the CRABP1 protein was differential for FTA samples compared to PTC and FTC samples with a decrease expression of both RNA and protein expression for malignant tumors. In regard to its appurtenance to the gene cluster 2 (Figure 3), CRABP1 protein expression is able to separate benign from malignant groups.

Discussion

The clinical diagnosis of thyroid tumors is subject to significant inter- and intra-observer variations [22,23]. Since 2001, several thyroid microarray studies have proposed molecular markers but none has proved its clinical usefulness. Almost all the previous studies simultaneously compared few tumor classes. Based on our dataset which included the majority of differentiated follicular lesions of the thyroid, we analyzed 6 publicly available microarray datasets that contain a total of 347 thyroid samples defining up to 12 classes of tissues.

Unsupervised analysis of the class signatures showed a separation into 3 groups that belonged to lesions of benign behavior, malignant tumors, and oncocytic tissue classes (Figure 1). Pairwise distinctions of benign/normal and malignant tumors are consistent with published reports, like in the case of PTC and FTA [14]. Considering the group of oncocytic follicular tumors as an independent class or as a variant of follicular tumors was debated through the last two WHO classifications. This question is still relevant in view of the recent results on miRNA profiling, when oncocytic tumors clearly segregated from other conventional follicular tumors [24]. According to their specific profiles for gene and miRNA expression but also to their mutational status for RAS and PAX8-PPARγ genes [7], oncocytic tumors should be reconsidered as a distinct entity from the other follicular tumors.

The role of the “oncocytic group” signature on the appropriateness of the classification of benign and malignant tumors is questionable. Little is known about the etiology of oncocytoma but their signature is mainly composed of numerous mitochondrial and metabolic genes, which role in tumor aggressiveness is still debate [7,20]. Interestingly, we found that microfollicular adenomas (FTAb) were related by their molecular profiles to both oncocytic adenomas and carcinomas. They were mainly connected through genes involved in mitochondrial functions as oxidative phosphorylation. Relationships between thyroid metabolism and microfollicular patterns have been explored [25], but never in the context of mitochondrial functions. However in this latest study, the galectin-3 profile associated with FTAb may be related to its role in the regulation of mitochondrial stability [26]. Galectin-3 has been also involved in thyroid tumorigenesis throughout an antiapoptotic

activity [27]. In this context, the malignant potential of FTAb is still under discussion [28].

Exploring two large datasets allowed the definition of 104 relevant and cross-validated differential genes. Although the gene selection was done on different sets of classes (7 vs 12 classes), discrepancies of some gene profiles (cluster 5) may be related to incomplete microarray platform compatibility or to differential cellular environment. Interestingly, most of genes from this cluster 5 were involved in the process of vascular endothelium cell expansion [29–31], that may represent the stromal reaction. Referring to our 12 sets of classes, we found that more than half genes from the 104 cross validated differential genes are also related to other classes which are not shared by the 2 datasets as AT, GD, TUM, FTAb. This may have consequence on the signature relevance for classifying these follicular pathologies and should explain the low percentage of genes common to both classifiers. Despite the low number of FTC and OTC samples and the microarray platform restricted to 9216 probes, our classifier was able to validate the classification of PTC, FTC, FTA and OTC from 5 other datasets (Table 2). Then, we proved the relevance of our approach to molecularly classify the main kinds of differentiated follicular thyroid pathologies. Looking for the classifiers performance, we showed that poor performances were associated with the FTA subclasses in the two main datasets. In our study, the functional status of the FTA samples may influence the classifier performance to discriminate true FTA from hyperfunctioning nodules. However, only 12% of macrofollicular adenomas presented mutations in TSH receptor or GNAS genes. Then, we believe that the FTA signature we proposed was able to discriminate the true FTA from other follicular thyroid tumors. Relevance for making subclasses of follicular adenomas was tested on 6 cross validated markers and 6 specific markers from our own analysis (Figure 4). For 5 specific markers, we were able to segregate both microfollicular and macrofollicular adenomas from oncocytic adenomas and minimal invasive follicular carcinomas, while only one of the six cross validated markers presented the same performance. Then, distinguishing subclasses of adenoma promotes the identification of relevant molecular markers of follicular adenomas. These markers have never been differentially selected among the previous microarray analyses on thyroid tumors. However, 3 of them (ADAMTS2, ALDH1A3 and CABIN1) were previously proposed as prognosis markers of epithelial tumors recurrence [32–34]. Observation of six protein profiles highlights the difficulty to propose immunomarkers to be used in practice. However, some markers may be relevant for the classification of 6 differentiated follicular pathologies (Macrofollicular FTA, FTC, PTC, OTA and OTC). Then, APOD showing a specific protein profile for FTC and PTC has been recently proposed to modify the proliferative activity of cancer cells [35]. In the context of structure-function relationship, the differential expression of the folding protein Calmegin (CLGN) we observed for FTA and FTC samples may be associated with a difference in protein abundance recently described for thyroid tumors [36]. Finally, one subunit of the complex II of the mitochondrial respiratory chain (SDHA) was able to distinguish OTC from OTA [20]. These markers can be used complementary to the current markers [37], to improve the differential diagnosis for suspicious nodular thyroid lesions.

Enlarging number of classes has revealed new behavior for some makers. TIMP1 and CRABP1 were not selected in our classifier

but were in the intersected differential gene lists. Indeed, these markers belonged to several classes in our dataset: AT, GD and PTC for TIMP1; FTC, PTC, FTA and TUM for CRABP1. These relations were confirmed at protein level on new samples from classes common to both datasets. Our data clearly questioned about the relevance of TIMP1 as a marker of thyroid cancer. As high protein levels were observed in AT and PTC classes, TIMP1 expression should be more related to the frequent described lymphocyte infiltration in PTC [38]. For CRABP1, quantitative RT-PCR analysis on new follicular samples confirmed the upregulation of CRABP1 expression in all the adenomas comparing to follicular carcinomas. Underexpression of CRABP1 mRNA was relevant to distinguish both FTC and PTC classes, in accordance with previous studies [39,40]. However, our results differed at protein level compared to a study on cold thyroid nodules where both expression of CRABP1 mRNA and protein were down regulated compared to normal surrounding tissue [41]. As our FTAs were mostly non functioning nodules, we postulate that this difference in protein level may be due to the histological class studied, macrofollicular adenomas in our case and a mix of follicular adenomas and adenomatous nodules in the other study. Although CRABP1 mRNA level seems to be inversely correlated with the malignant potential of thyroid nodules, its protein level should be more precisely explored in the context of goitres.

In conclusion, diagnosis tools defined from thyroid microarray data are more relevant when many samples and tissue classes are used, especially when including the oncocytic tumors. Considering the model for sequential progression in follicular tumors, we postulate that gene expression profiling remains the most powerful method to identify early markers of transformation. Molecular classification of microfollicular adenomas on fine needle aspirations biopsies may have a direct impact in the management of such tumors.

Materials and Methods

Samples and Microarray datasets

We analyzed gene-expression data from six microarray datasets containing up to 11 classes of the differentiated pathologies from the thyroid organ and the normal tissue. We subcategorized the FTA samples according to follicle architectural criteria to determine whether these subclasses (macro, FTA or microfollicular, FTAb) represent distinct molecular entities. When possible, the FTC class was divided into two subclasses, defined according to the presence or absence of the PAX8/PPARγ translocation in the samples. Atypical (TUM) and/or oncocytic features (OT) were individualized on criteria previously specified [15,20]. We also explored non-tumoral lesions as from Grave's disease (GD) and autoimmune thyroiditis (AT) to identify the immunological environment of the tumoral samples. This study is based on our dataset (Fontaine dataset) containing 12 subclasses represented by 166 anonymized human thyroid tissues. The 5 other datasets, particularly the Giordano dataset that contains 7 tissue subclasses, were used to cross-validate the results of the molecular classification, the classifiers, and the functional analysis of differential genes [5,42–44]. One dataset (He et al.) has not yet associated citation. Informations on the 347 samples from the 6 datasets are summarized in Table 5. The Fontaine and the Giordano datasets are referred as the main datasets. All the datasets are publicly available and were generated by various microarray platforms (Table 5). The Fontaine (GSE6339), the He (GSE3467) and the Reyes datasets (GSE3678) were hosted by the NCBI GEO database. The Weber dataset was

downloaded from EBI ArrayExpress database (E-MEXP-97). Two datasets were downloaded from the author's websites: the Giordano dataset (http://dot.ped.med.umich.edu:2000/pub/PPARG/index.html) and the Jarzab dataset (http://www.genomika.pl/thyroidcancer/PTCCancerRes.html).

Data analysis

For our dataset, hierarchical clustering of the data was computed on log-transformed, median gene-centered and normalized as previously described [15]. Normalized data from the Giordano and Jarzab datasets were downloaded from the authors' websites. For the remaining datasets, raw data were downloaded from GEO or Array Express databases and normalized in BRB Array Tools v3.7.0b1 developed by Dr. Richard Simon. As data originated from three different but one-colour platforms, global normalization to the median was used to center the gene-expression values on each array. Genes showing minimal variation across the set of arrays were excluded from the analysis (log-intensity variation P-value > 0.15). Hierarchical clustering was done on centered genes by using centered correlation distance and an average linkage method. Robustness and Discrepancy indices were computed from 10,000 data perturbations. For the two largest datasets [5,15], we identified genes that were differentially expressed among the classes by using a multivariate 10,000 permutation test, to provide 95% confidence that the number of false discoveries did not exceed 1. Although F-statistics were used for each gene, the multivariate permutation test is non-parametric and does not require the assumption of Gaussian distributions. Functional enrichments in gene clusters were defined at 5% risk for level 3 Gene Ontology terms by Fatigo tools (adjusted P-values, www.fatigo.bioinfo.cipf.es [45]) as compared to the whole genome, and for canonical pathways by Ingenuity Pathway Analysis tools as compared to the Ingenuity knowledge database (www.ingenuity.com) Classifiers were defined for each pathological class and trained either versus the wild type class (binary training) or the entire dataset (complete training). Twenty genes were selected by the Greedy Pairs algorithm, and the 'Diagonal Linear Discriminant Analysis' algorithm was used for training and sample prediction [46]. Intra- and inter-dataset cross-validations were done by the 0.632+ method and the 'Diagonal Linear Discriminant Analysis' algorithm respectively [47]. Since the classical accuracy coefficient is biased in large datasets containing unbalanced class sizes, we preferred to use the positive accuracy defined as the geometric mean of the sensitivity (proportion of true positives for the samples which belong to the targeted class in the entire datasets) and the positive predictive value (proportion of true positives in samples detected as belonging to the targeted class) [48].

Quantitative RT-PCR analysis

An independent set of 32 follicular tumors comprising 8 minimal invasive FTC and 24 follicular adenoma (8 OTA, 8 FTA and 8 FTAb) were subjected to real time quantitative RT PCR to test the relevance of several selected markers. The anonymized samples were obtained through the Department of Pathology, from the Academic Angers University. Total RNA was isolated using Trizol reagent (Invitrogen, Paisley, UK). RNA integrity was determined using a Bio- Analyzer 2100 (Agilent Technologies, Waldbronn, Germany). The reverse transcription was performed on 1µg of RNA with Advantage RT-for-PCR kit (Clontech Laboratory, Palo Alto, CA, USA) following the manufacturer's recommendations. Real-time quantification

was performed in a 96-well plate by using the IQ SYBR Green supermix and Chromo4 (Biorad, CA, USA) with the recommended protocol. Twelve genes were explored for their expression level: CASP10, CDH16, CLGN, CRABP1, HMGB2, ALPL2, ADAMTS2, CABIN1, ALDH1A3, USP13, NR2F2 and KRTHB5. Data were normalized to ß-globin expression level. Significances of differential expression between the classes were assessed by T-tests and those over the classes by F-tests. The sequences of primers used in this study are specified in Table 6.

Detection of mutations and PAX8-PPARG rearrangement

The mutational status for TRSH, GNAS and NRAS genes was determined for all the FTA Coulter, Fullerton, CA, USA) following the manufacturer's instructions.

The presence of the PAX8-PPARG translocation was evaluated as described [51], on the 3 FTC from the microarray study and the 8 FTC from the RT-PCR study. Quantification, degradation and DNA contamination of RNA were assessed using an RNA 6000 Nano Assay kit (Agilent Technologies, Palo Alto, CA, USA). After reverse transcription, we looked for the translocation on 5 µl cDNA, using primers sequences previously specified [5] and the HotGoldstar DNA polymerase according to the manufacturer's recommendations (Eurogentec, Seraing, Belgium).

Dot blot analysis

Two micrograms of protein corresponding to the TRIzol fraction of 49 independent and anonymized thyroid tissues were spotted onto nitrocellulose membranes at room temperature, using a dot-blot apparatus and following the manufacturer's recommendations (Biorad, Hercules, CA). These tissues were provided by the Department of Pathology of the Academic Angers University and were different from and FTAb explored in studies by microarray (26 FTA and 17 FTAb) and RT-PCR (8 FTA and 8 FTAb). DNA was isolated during the guanidium isothiocyanate procedure (Trizol Reagent, Invitrogen Life Technologies). Exon 2 of the NRAS gene was amplified using primer sequences described elsewhere [49]. Exons 9 and 10 of the TSH receptor gene as well as exons 8 and 9 of the Gs alpha gene were amplified using primers sequences previously described [50]. PCR were performed on 5 µl DNA with the HotGoldstar DNA polymerase according to the manufacturer's recommendations (Eurogentec, Seraing, Belgium). Amplified fragments were purified and directly sequenced on a CEQ 8000 apparatus, using a CEQ DTCS Quick Start Kit (Beckman those used for microarray and RT-PCR studies. Six groups of thyroid pathologies were defined on non-ambiguous cytological criteria: AT, FTA, FTC, PTC, OTA and OTC and compared to normal thyroid tissue (WT). Seven samples were tested for each group and the assays were performed in duplicate. Seven primary antibodies (all from Abcam, Cambridge, UK) were used at specific dilutions: 1/750 for TIMP1, 1/1000 for APOD, SDHA, CLGN, CRABP1, APOE and 1/5000 for á-tubulin as a control. Peroxidase anti-rabbit or anti-mouse secondary antibodies were revealed by ECL methodology (ECL Plus reagent kit, Amersham, Chalfont, UK). Spots were quantified on a GelDoc XRS apparatus using the Quantity One software (Biorad) and reported to control value. Significances of differential expression over the classes were assessed by F-tests.

Acknowledgments

We thank Dominique Couturier and Anne Coutoleau for technical help. We thank Marja Steenman for the critical reading of this paper. This work was supported by grants from the French Ministry of Research,

the Institut National de la Recherche Médicale (INSERM), the
University Hospital of Angers and the University of Angers (PHRC 03-10).

References

1. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 101: 9309-9314.
2. Hundahl SA, Fleming ID, Fremgen AM, Menck HR (1998) A National Cancer Data Base report on 53,856 cases of thyroid carcinoma treated in the U.S., 1985-1995. Cancer 83: 2638-2648.
3. Yeh MW, Demircan O, Ituarte P, Clark OH (2004) False-negative fine-needle aspiration cytology results delay treatment and adversely affect outcome in patients with thyroid carcinoma. Thyroid 14: 207-215.
4. Lacroix L, Lazar V, Michiels S, Ripoche H, Dessen P, et al. (2005) Follicular thyroid tumors with the PAX8-PPARgamma1 rearrangement display characteristic genetic alterations. Am J Pathol 167: 223-231.
5. Giordano TJ, Au AY, Kuick R, Thomas DG, Rhodes DR, et al. (2006) Delineation, functional validation, and bioinformatic evaluation of gene expression in thyroid follicular carcinomas with the PAX8-PPARG translocation. Clin Cancer Res 12: 1983-1993.
6. Barden CB, Shister KW, Zhu B, Guiter G, Greenblatt DY, et al. (2003) Classification of follicular thyroid tumors by molecular signature: results of gene profiling. Clin Cancer Res 9: 1792-1800.
7. Nikiforova MN, Lynch RA, Biddinger PW, Alexander EK, Dorn GW, et al. (2003) RAS point mutations and PAX8-PPAR gamma rearrangement in thyroid tumors: evidence for distinct molecular pathways in thyroid follicular carcinoma. J Clin Endocrinol Metab 88: 2318-2326.
8. Aldred MA, Huang Y, Liyanarachchi S, Pellegata NS, Gimm O, et al. (2004) Papillary and follicular thyroid carcinomas show distinctly different microarray expression profiles and can be distinguished by a minimum of five genes. J Clin Oncol 22: 3531-3539.
9. Huang Y, Prasad M, Lemon WJ, Hampel H, Wright FA, et al. (2001) Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. Proc Natl Acad Sci U S A 98: 15044-15049.
10. Jarzab B, Wiench M, Fujarewicz K, Simek K, Jarzab M, et al. (2005) Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. Cancer Res 65: 1587-1597.
11. Weber F, Shen L, Aldred MA, Morrison CD, Frilling A, et al. (2005) Genetic classification of benign and malignant thyroid follicular neoplasia based on a three-gene combination. J Clin Endocrinol Metab 90: 2512-2521.
12. Eszlinger M, Wiench M, Jarzab B, Krohn K, Beck M, et al. (2006) Meta- and reanalysis of gene expression profiles of hot and cold thyroid nodules and papillary thyroid carcinoma for gene groups. J Clin Endocrinol Metab 91: 1934-1942.
13. Griffith OL, Melck A, Jones SJ, Wiseman SM (2006) Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers. J Clin Oncol 24: 5043-5051.
14. Finley DJ, Zhu B, Fahey TJ, III (2004) Molecular analysis of Hurthle cell neoplasms by gene profiling. Surgery 136: 1160-1168.
15. Fontaine JF, Mirebeau-Prunier D, Franc B, Triau S, Rodien P, et al. (2008) Microarray analysis refines classification of

non-medullary thyroid tumours of uncertain malignancy. Oncogene 27: 2228-2236.
16. Stolf BS, Abreu CM, Mahler-Araujo MB, Dellamano M, Martins WK, et al. (2005) Expression profile of malignant and non-malignant diseases of the thyroid gland reveals altered expression of a common set of genes in goiter and papillary carcinomas. Cancer Lett 227: 59-73.
17. Yukinawa N, Oba S, Kato K, Taniguchi K, Iwao-Koizumi K, et al. (2006) A multiclass predictor based on a probabilistic model: application to gene expression profiling-based diagnosis of thyroid tumors. BMC Genomics 7: 190.
18. Eszlinger M, Krohn K, Hauptmann S, Dralle H, Giordano TJ, et al. (2008) Perspectives for improved and more accurate classification of thyroid epithelial tumors. J Clin Endocrinol Metab 93: 3286-3294.
19. Corvilain B, Van SJ, Dumont JE, Vassart G (2001) Somatic and germline mutations of the TSH receptor and thyroid diseases. Clin Endocrinol (Oxf) 55: 143-158.
20. Baris O, Savagner F, Nasser V, Loriod B, Granjeaud S, et al. (2004) Transcriptional profiling reveals coordinated up-regulation of oxidative metabolism genes in thyroid oncocytic tumors. J Clin Endocrinol Metab 89: 994-1005.
21. Wasenius VM, Hemmer S, Kettunen E, Knuutila S, Franssila K, et al. (2003) Hepatocyte growth factor receptor, matrix metalloproteinase-11, tissue inhibitor of metalloproteinase-1, and fibronectin are up-regulated in papillary thyroid carcinoma: a cDNA and tissue microarray study 1. Clin Cancer Res 9: 68-75.
22. Franc B, de la SP, Lange F, Hoang C, Louvel A, et al. (2003) Interobserver and intraobserver reproducibility in the histopathology of follicular thyroid carcinoma. Hum Pathol 34: 1092-1100.
23. Lloyd RV, Erickson LA, Casey MB, Lam KY, Lohse CM, et al. (2004) Observer variation in the diagnosis of follicular variant of papillary thyroid carcinoma. Am J Surg Pathol 28: 1336-1340.
24. Nikiforova MN, Tseng GC, Steward D, Diorio D, Nikiforov YE (2008) MicroRNA expression profiling of thyroid tumors: biological significance and diagnostic utility. J Clin Endocrinol Metab 93: 1600-1608.
25. Nucera C, Mazzon E, Caillou B, Violi MA, Moleti M, et al. (2005) Human galectin-3 immunoexpression in thyroid follicular adenomas with cell atypia. J Endocrinol Invest 28: 106-112.
26. Fukumori T, Oka N, Takenaka Y, Nangia-Makker P, Elsamman E, et al. (2006) Galectin-3 regulates mitochondrial stability and antiapoptotic function in response to anticancer drug in prostate cancer. Cancer Res 66: 3114-3119.
27. Savin S, Cvejic D, Isic T, Paunovic I, Tatic S, et al. (2008) Thyroid peroxidase and galectin-3 immunostaining in differentiated thyroid carcinoma with clinicopathologic correlation. Hum Pathol in press.
28. Schmid KW, Farid NR (2006) How to define follicular thyroid carcinoma?. Virchows Arch 448: 385-393.
29. Ao A, Wang H, Kamarajugadda S, Lu J (2008) Involvement of estrogen-related receptors in transcriptional response to hypoxia and growth of solid tumors. Proc Natl Acad Sci U S A 105: 7821-7826.
30. Sierra JR, Corso S, Caione L, Cepero V, Conrotto P, et al. (2008) Tumor angiogenesis and progression are enhanced by Sema4D produced by tumor-associated macrophages. J Exp Med 205: 1673-1685.
31. Taylor KL, Leaman DW, Grane R, Mechti N, Borden EC, et al. (2008) Identification of interferon-beta-stimulated genes that inhibit angiogenesis in vitro. J Interferon Cytokine Res 28: 733-740.
32. Carinci F, Lo ML, Piattelli A, Rubini C, Chiesa F, et al. (2005) Potential markers of tongue tumor progression selected by cDNA microarray. Int J Immunopathol Pharmacol 18: 513-524.

33. Porter S, Scott SD, Sassoon EM, Williams MR, Jones JL, et al. (2004) Dysregulated expression of adamalysin-thrombospondin genes in human breast carcinoma. Clin Cancer Res 10: 2429-2440.
34. Watanabe T, Kobunai T, Sakamoto E, Yamamoto Y, Konishi T, et al. (2009) Gene expression signature for recurrence in stage III colorectal cancers. Cancer 115: 283-292.
35. Soiland H, Soreide K, Janssen EA, Korner H, Baak JP, et al. (2007) Emerging concepts of apolipoprotein D with possible implications for breast cancer. Cell Oncol 29: 195-209.
36. Netea-Maier RT, Hunsucker SW, Hoevenaars BM, Helmke SM, Slootweg PJ, et al. (2008) Discovery and validation of protein abundance differences between follicular thyroid neoplasms. Cancer Res 68: 1572-1580.
37. Barroeta JE, Baloch ZW, Lal P, Pasha TL, Zhang PJ, et al. (2006) Diagnostic value of differential expression of CK19, Galectin-3, HBME-1, ERK, RET, and p16 in benign and malignant follicular-derived lesions of the thyroid: an immunohistochemical tissue microarray analysis. Endocr Pathol 17: 225-234.
38. Delys L, Detours V, Franc B, Thomas G, Bogdanova T, et al. (2007) Gene expression and the biological phenotype of papillary thyroid carcinomas. Oncogene 26: 7894-7903.
39. Hawthorn L, Stein L, Varma R, Wiseman S, Loree T, et al. (2004) TIMP1 and SERPIN-A overexpression and TFF3 and CRABP1 underexpression as biomarkers for papillary thyroid carcinoma. Head Neck 26: 1069-1083.
40. Huang Y, de la CA, Pellegata NS (2003) Hypermethylation, but not LOH, is associated with the low expression of MT1G and CRABP1 in papillary thyroid carcinoma. Int J Cancer 104: 735-744.
41. Eszlinger M, Krohn K, Berger K, Lauter J, Kropf S, et al. (2005) Gene expression analysis reveals evidence for increased expression of cell cycle-associated genes and Gq-protein-protein kinase C signaling in cold thyroid nodules. J Clin Endocrinol Metab 90: 1163-1170.
42. He H, Jazdzewski K, Li W, Liyanarachchi S, Nagy R, et al. (2005) The role of microRNA genes in papillary thyroid carcinoma. Proc Natl Acad Sci U S A 102: 19075-19080.
43. Jarzab B, Wiench M, Fujarewicz K, Simek K, Jarzab M, et al. (2005) Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. Cancer Res 65: 1587-1597.
44. Weber F, Shen L, Aldred MA, Morrison CD, Frilling A, et al. (2005) Genetic classification of benign and malignant thyroid follicular neoplasia based on a three-gene combination. J Clin Endocrinol Metab 90: 2512-2521.
45. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Nucleic Acids Res 33: W460- W464.
46. Dudoit S, Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol 3: 1-21.
47. Efron B, Tibshirani R (1997) Improvments of cross-validation: the .632+ bootstrap method. 92: 548-560.
48. Kubat M (1998) Decision trees can initialize radial-basis function networks 1. IEEE Trans Neural Netw 9: 813-821.
49. Vasko V, Ferrand M, Di CJ, Carayon P, Henry JF, et al. (2003) Specific pattern of RAS oncogene mutations in follicular thyroid tumors. J Clin Endocrinol Metab 88: 2745-2752.
50. Bourasseau I, Savagner F, Rodien P, Duquenne M, Reynier P, et al. (2000) No evidence of thyrotropin receptor and G(s

alpha) gene mutation in high iodine uptake thyroid carcinoma. Thyroid 10: 761-765.
51. Kroll TG, Sarraf P, Pecciarini L, Chen CJ, Mueller E, et al. (2000) PAX8-PPARgamma1 fusion oncogene in human thyroid carcinoma. Science 289: 1357-1360.

Figures Tables

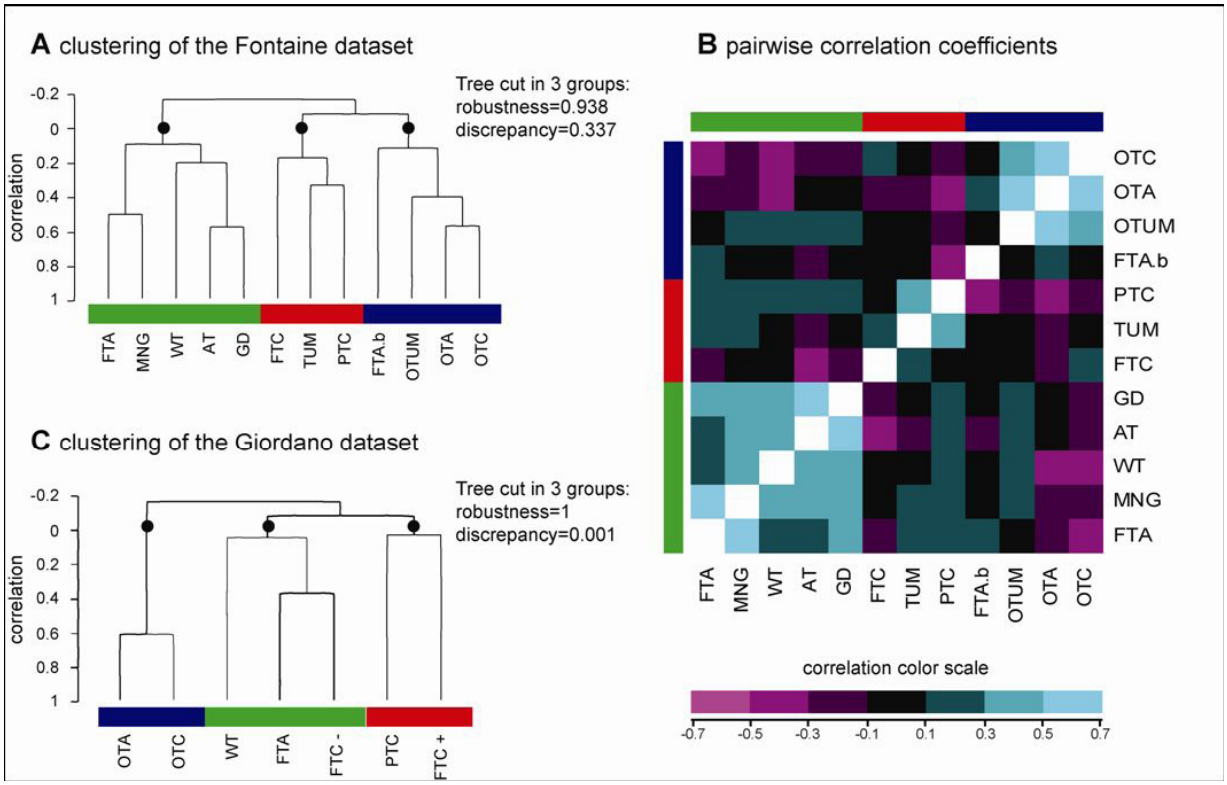


Figure 1: Molecular classification of thyroid tissues.
The centroid signatures of the thyroid tissues were compared to each other in the main dataset [15], and compared to the Giordano dataset [5]. Centroids were defined as the mean geneexpression signature of each tissue class over all the genes. **Panel A** shows the hierarchical clustering of the tissues in the Fontaine dataset. The dendrogram can be cut into 3 robust branches (black circles) which define 3 groups of classes. The groups contain benign lesions and normal tissue (green boxes), malignant tumors (red boxes), or oncocytic tumors and microfollicular adenomas (blue boxes). **Panel B** shows pairwise correlation coefficients between the tissue classes. The heatmap represents a symmetrical matrix of color coded correlation coefficients. Tissue classes are ordered as in the hierarchical clustering and marked with the same 3 colors (green, red and blue). **Panel C** shows the hierarchical clustering of the tissues in the Giordano dataset. The dendrogram can be cut into 3 robust branches (black circles) which define 3 groups of classes. The groups contain heterogeneous lesions and normal tissue (green boxes), malignant tumors (red boxes), or

oncocytic tumors (blue boxes). The two datasets show 3 groups containing similar tissue classes. WT: Wild type tissue; PTC: papillary thyroid carcinoma; FTA: macrofollicular thyroid adenoma (and microfollicular, FTAb); GMN, multinodular goiter; FTC: follicular thyroid carcinoma (FTC+, with Pax8/PPARγ translocation); OTA: oncocytic thyroid adenoma; OTC: oncocytic thyroid carcinoma; TUM, Tumor of uncertain malignancy (OTUM, oncocytic variant); GD, Grave's disease; AT, Autoimmune thyroïditis.

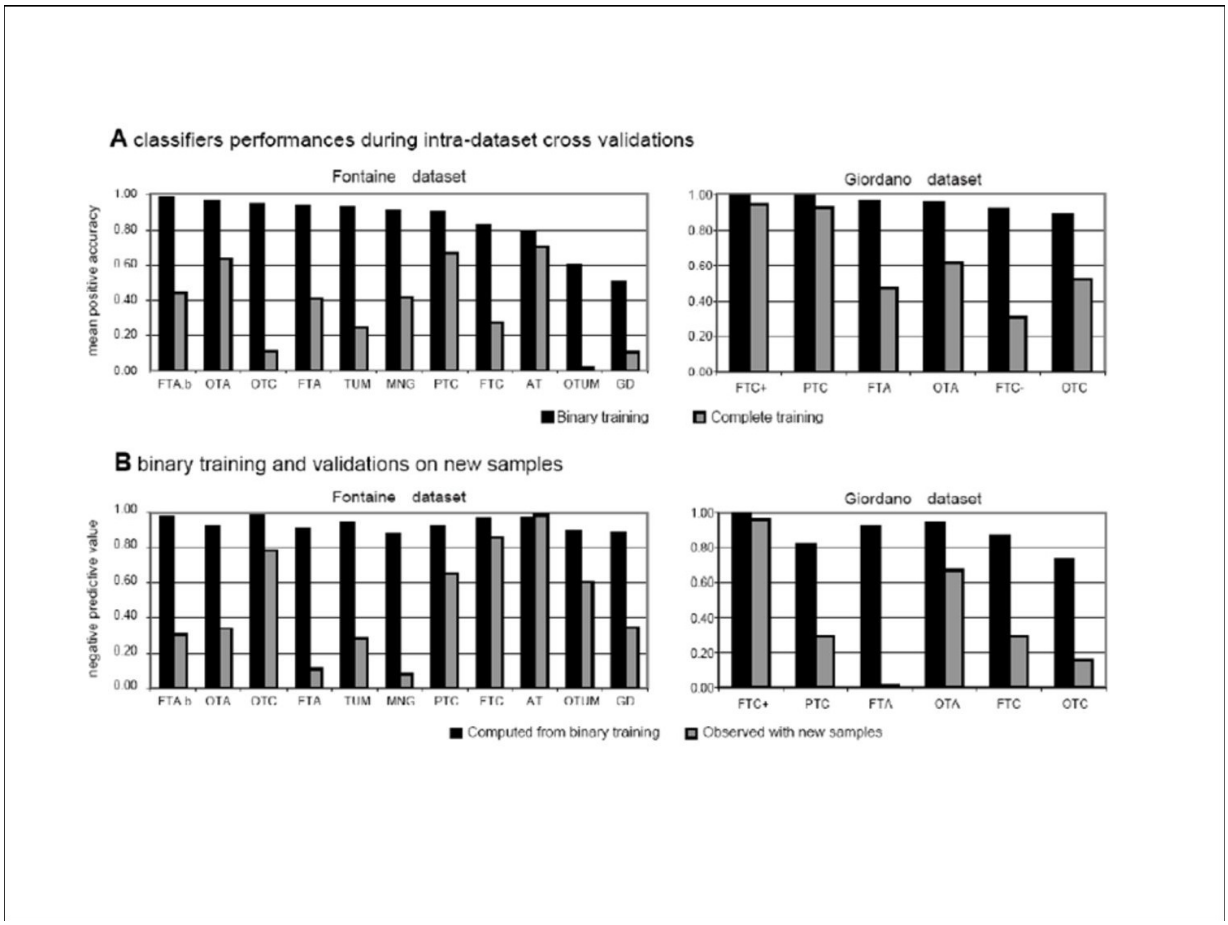


Figure 2: Class prediction of the samples.

Panel A shows for the two main datasets, the mean positive accuracy of classifiers during intra-dataset cross-validations. Each classifier is composed of 20 genes selected by the Greedy pairs algorithm and using the Diagonal Linear Discriminant Analysis for training and predictions. Some classifiers were trained with samples from the class and the normal tissue, i.e. a binary training (black bars). Other classifiers were trained with all the samples of the dataset, i.e. a complete training (gray bars). **Panel B** shows the class prediction of new samples by binary-trained classifiers in the two datasets. All the new samples come from the same dataset and should be predicted as negative. The negative predictive value which was deduced from the binary training (black bars) is compared to observations on the new samples (gray bars).

WT: Wild type tissue; PTC: Papillary thyroid carcinoma; FTA: Macrofollicular thyroid adenoma (and microfollicular, FTAb); GMN, Multinodular goitre; FTC: Follicular thyroid carcinoma (FTC+, with

Pax8/PPARγ translocation); OTA: Oncocytic thyroid adenoma; OTC: Oncocytic thyroid carcinoma; TUM, Tumor of uncertain malignancy (OTUM, oncocytic variant); GD, Grave's disease; AT, Autoimmune thyroïditis.

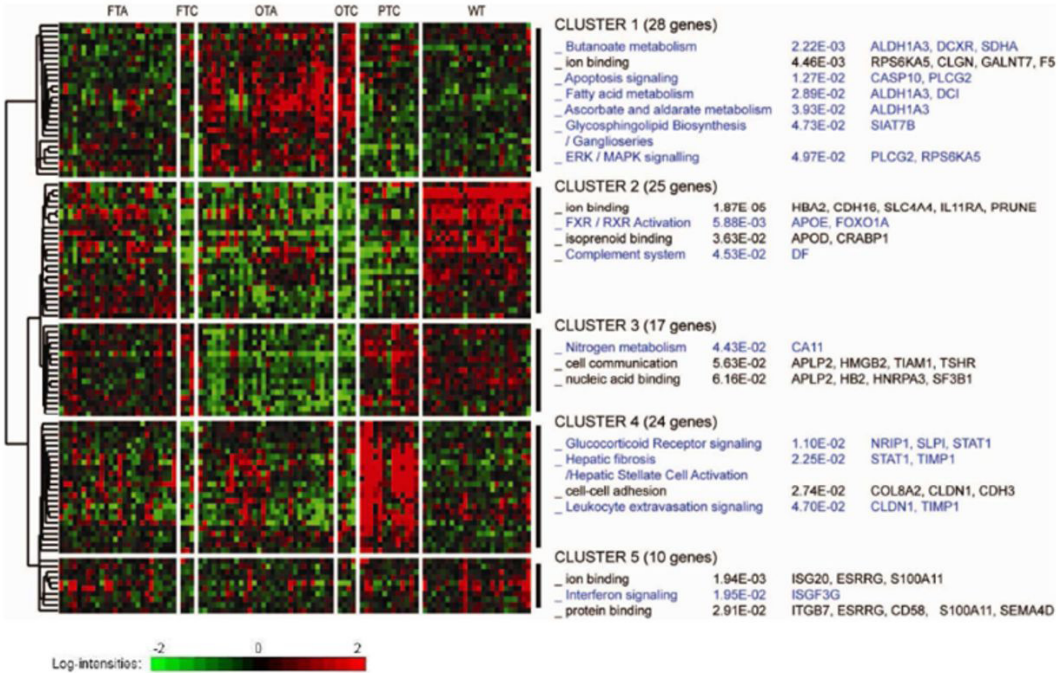


Figure 3: Hierarchical clustering of cross-validated differential genes.
The heatmap shows gene-expression levels in the Fontaine dataset for the tissue classes that are common to the Giordano dataset (104 genes). Functional enrichments are shown for five identified gene clusters on the right, followed by p-values and involved genes. Black text represents level 3 Gene Ontology terms. Blue text represents Ingenuity canonical pathways. Gene profiles of the differential genes reflect the observed class similarities.
WT: Wild type tissue; PTC: Papillary thyroid carcinoma; FTA: Macrofollicular thyroid adenoma; FTC: Follicular thyroid carcinoma; OTA: Oncocytic thyroid adenoma; OTC: Oncocytic thyroid carcinoma.

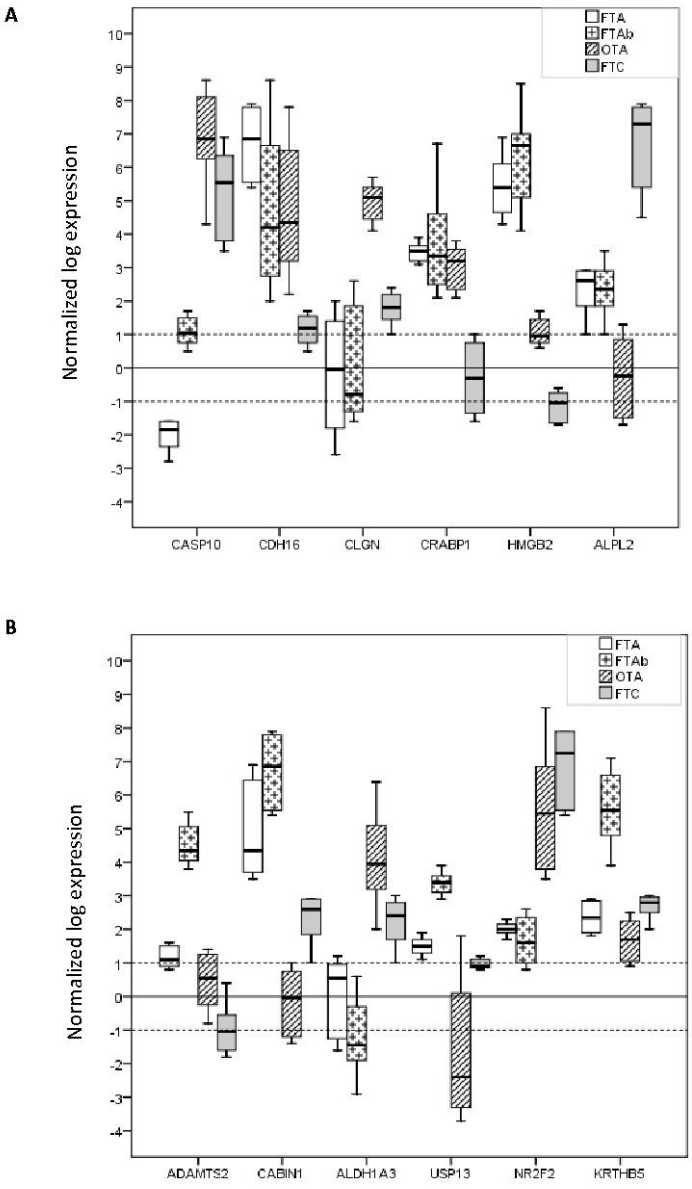


Figure 4: Gene expression of selected markers.
Differential expression of 12 genes measured by real time quantitative RT-PCR on 32 new follicular thyroid tumors (8 FTA, 8 FTAb, 8 OTA and 8 FTC). The upper and the lower limits of each box stand for the upper and the lower quartiles, respectively. Bold lines represent medians. The expression of ADAMTS2, CABIN1, ADLH1A3, USP13, NR2F2, KRTHB5, CASP10, CDH16, CLGN, CRABP1, HMGB2 and ALPL2 genes was referred to ®-globin expression level. Significances of differential expression between the classes were assessed by T-tests and those over the classes by F-tests. FTA: Macrofollicular thyroid adenoma (FTAb, microfollicular); FTC: Follicular thyroid carcinoma; OTA: Oncocytic thyroid adenoma

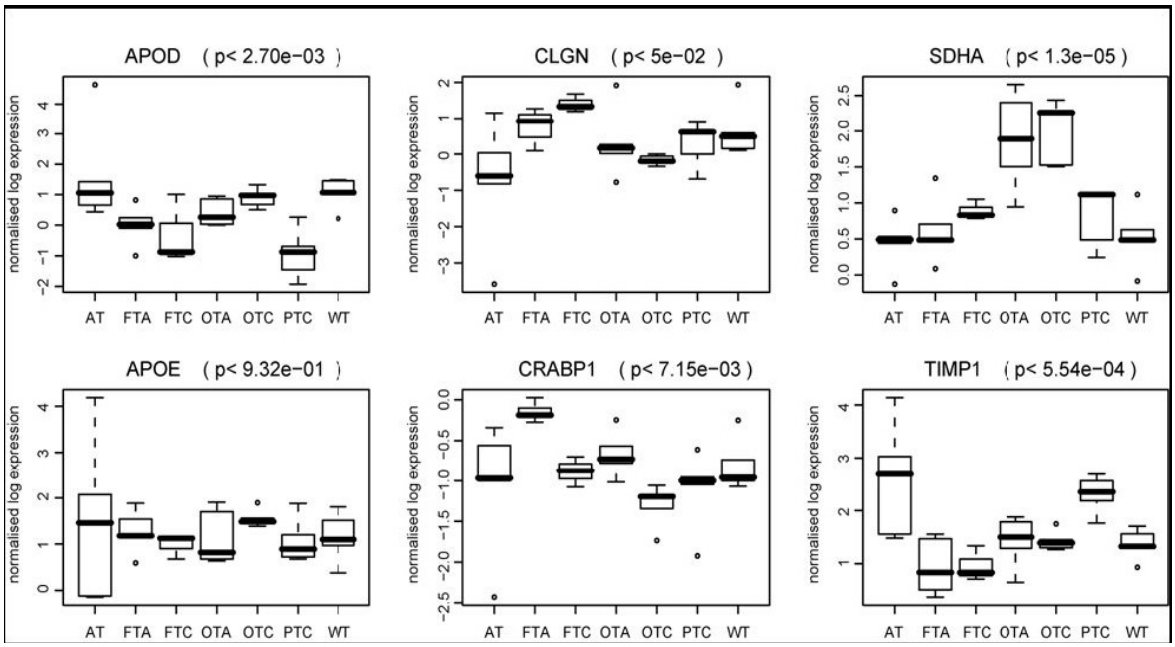


Figure 5: Protein expression of selected genes.
The protein expression of 6 selected differential genes was measured on 49 new tissue samples. These independent samples represent six tissue classes that are common to the two main datasets (WT, PTC, FTA, FTC, OTA, and OTC). The AT class was added to compare the expression levels to a non-tumoral lesion. Seven samples were used for each class. Protein expression profiles of APOD, CLGN, SDHA, APOE, CRABP1 and TIMP1 are shown by boxplots for the tissue classes. Significances of differential expression over the classes were assessed by F-tests. P-values are shown between brackets. WT: Wild type tissue; PTC: Papillary thyroid carcinoma; FTA: Macrofollicular thyroid adenoma; FTC: Follicular thyroid carcinoma; OTA: Oncocytic thyroid adenoma; OTC: Oncocytic thyroid carcinoma.

Table 1: Prediction of samples by considering three groups of tissue classes.
Values can be compared to the range of accuracies for individual lesions that composed each group.

	Fontaine dataset		Giordano dataset	
	Group	Individual lesions	Group	Individual lesions
Benign group	0,79	0.11 - 0.71	0,76	0.31 - 0.47
Malignant group	0,69	0.24 - 0.67	0,93	0.93 - 0.94
Oncocytic group	0,75	0.02 - 0.64	0,86	0.52 - 0.62

Table 2: Classifier genes from the Fontaine dataset (n=220 genes).

Symbol	Clone ID	Class	Cross validation support (%)	Parametric p-value	t-value	Fold-change
ABCB9	36371	AT	100	1,00E-07	6,81	3,28
SEP-02	729542	AT	100	2,67E-01	-1,11	0,75
SEP-05	448163	AT	100	1,00E-07	7,03	4,96
AAAS	740617	AT	99	3,00E-07	5,36	3,34
ABCA3	52740	AT	97	4,99E-01	0,68	1,13
ABCA8	284828	AT	97	1,86E-01	-1,33	0,74
ABCB1	1837488	AT	97	2,26E-01	-1,22	0,72
ABCB8	238705	AT	97	3,51E-01	-0,94	0,84
ABCC3	208097	AT	96	1,54E-01	1,43	1,64
ABCF2	1086914	AT	96	1,00E-07	6,91	3,37
BATF	1929371	AT	83	1,00E-07	10,52	6,03
C18orf51	742649	AT	62	1,00E-07	9,87	7,47
CARD11	665651	AT	46	1,00E-07	10,85	4,72
CBFA2T2	43629	AT	39	1,00E-07	10,18	10,03
CD38	1352408	AT	29	1,00E-07	10,18	7,22
CD79A	1056782	AT	25	1,00E-07	11,14	8,58
CD79B	155717	AT	25	1,00E-07	11,91	10,51
CXCL6	2108870	AT	14	1,00E-07	10,13	4,29
EPPB9	668454	AT	7	1,00E-07	9,67	9,03
DKFZp667B0210	1844689	AT	3	1,00E-07	9,30	4,83
PLXNB2	1420676	FTA	65	6,00E-07	5,21	1,70
TBC1D17	739955	FTA	62	4,00E-07	5,27	1,56
ZNF76	745003	FTA	59	8,00E-07	5,15	1,51
KIAA1196	738938	FTA	52	1,40E-06	5,01	1,52
ASGR1	25883	FTA	48	1,25E-05	4,51	1,75
BCL2L11	300194	FTA	42	6,00E-06	4,68	1,42
DAP	725371	FTA	41	6,20E-06	4,67	1,51
CABIN1	1844968	FTA	40	1,12E-05	4,53	1,59
FLJ14981	24532	FTA	31	1,84E-04	3,83	1,50
RAD51L1	295412	FTA	30	4,23E-05	4,21	1,66
TSPAN1	376003	FTA	26	3,52E-05	4,25	1,57
IMAGE:745465	745465	FTA	25	2,58E-02	2,25	1,24
PLCG1	1174287	FTA	25	8,36E-05	-4,04	0,67
COL18A1	359202	FTA	24	3,20E-05	4,28	1,60
RPS6KA2	22711	FTA	23	3,77E-05	4,24	1,57
CaMKIINalpha	173820	FTA	22	9,52E-04	3,37	1,35
KLF1	208991	FTA	16	3,01E-03	3,01	1,49
BAT3	24392	FTA	13	8,11E-04	3,41	1,42
PDK4	487379	FTA	11	1,00E-02	-2,61	0,70
IMAGE:24065	24065	FTA	6	4,52E-04	-3,58	0,75
KRTHB5	364569	FTAb	86	1,00E-07	6,02	1,94
METRN	487475	FTAb	86	1,00E-07	5,68	2,15

DCI	667892	FTAb	74	6,00E-07	5,19	1,63
IFI30	740931	FTAb	68	4,90E-06	-4,73	0,41
ALCAM	172828	FTAb	64	1,21E-05	-4,51	0,68
ADAMTS2	364844	FTAb	55	1,61E-04	3,86	1,41
ETF1	146976	FTAb	54	1,12E-04	3,96	1,54
TP53I11	667514	FTAb	44	1,22E-04	3,94	1,94
IMAGE:487086	487086	FTAb	42	1,52E-04	-3,88	0,67
IMAGE:666315	666315	FTAb	35	5,63E-04	3,52	1,52
GRB10	564994	FTAb	33	1,80E-04	3,83	1,42
ISG20	740604	FTAb	33	3,51E-05	-4,26	0,60
SPARC	267358	FTAb	33	1,09E-04	3,97	1,47
HAS3	667533	FTAb	28	1,07E-04	3,97	1,52
USP13	666007	FTAb	20	9,96E-05	3,99	1,56
KRT4	1173570	FTAb	13	2,87E-04	3,71	1,65
LHFP	591534	FTAb	9	2,24E-03	-3,11	0,74
PCDHB15	730593	FTAb	9	1,06E-03	-3,33	0,74
TPM2	740620	FTAb	9	9,35E-03	2,63	1,45
DKFZp434M202	743118	FTAb	2	5,77E-03	-2,80	0,76
C22orf9	667250	FTC	89	2,00E-07	-5,43	0,34
SUCLG2	687551	FTC	83	7,00E-07	-5,18	0,27
LAPTM4A	726684	FTC	77	4,00E-07	-5,26	0,24
KIAA1576	743426	FTC	76	1,00E-07	5,70	3,33
NR2F2	72744	FTC	65	2,60E-06	4,87	3,11
CHST3	665327	FTC	62	2,90E-06	-4,85	0,32
PAI-RBP1	236055	FTC	55	5,30E-06	-4,71	0,29
IMAGE:729896	729896	FTC	54	3,00E-07	5,34	3,22
ZBTB16	2467442	FTC	52	4,72E-05	4,18	4,68
ZNF364	471834	FTC	52	5,80E-06	-4,69	0,39
GHITM	25077	FTC	50	5,30E-06	-4,71	0,31
LRP8	415554	FTC	49	1,43E-04	3,90	2,74
CEECAM1	666784	FTC	45	2,67E-04	3,73	2,22
HIST1H4C	682451	FTC	43	1,57E-05	-4,45	0,25
TIE1	743043	FTC	35	5,46E-05	4,14	2,52
SH3RF2	744797	FTC	22	5,37E-03	2,82	2,18
MGC5395	238840	FTC	15	5,07E-04	3,55	2,66
EMX2	365121	FTC	14	6,95E-03	-2,73	0,51
LRRC28	26519	FTC	13	1,51E-03	-3,23	0,40
GFRA3	2716748	FTC	11	4,52E-03	2,88	1,93
TTC19	136366	GD	72	1,57E-04	3,87	1,87
ATP6V1G2	726424	GD	61	1,91E-04	3,82	2,22
IMAGE:23817	23817	GD	61	3,65E-04	-3,64	0,52
IMAGE:178161	178161	GD	52	6,99E-04	3,46	2,53
BDKRB2	665674	GD	50	6,16E-03	2,78	2,15
IMAGE:744385	744385	GD	48	1,37E-02	2,49	1,70
LOC150837	258698	GD	47	5,84E-04	3,51	1,81
SSX1	262894	GD	47	5,97E-04	3,50	1,76
DHX57	24623	GD	46	1,69E-03	-3,19	0,59

RAI14	731711	GD	46	8,19E-03	2,68	1,56
IMAGE:180851	180851	GD	45	5,36E-03	-2,82	0,61
MGC5566	53119	GD	43	1,74E-03	3,18	1,85
D21S2056E	740140	GD	36	2,44E-03	3,08	1,72
IMAGE:729489	729489	GD	31	1,28E-03	3,28	2,95
KIAA1737	122063	GD	29	1,78E-03	3,18	1,67
MLH3	250771	GD	21	8,59E-03	2,66	1,59
IMAGE:53081	53081	GD	18	3,35E-03	2,98	1,77
TERF1	135773	GD	16	2,81E-02	2,22	1,45
DCXR	724596	GD	15	3,39E-03	-2,97	0,61
IMAGE:136686	136686	GD	15	6,29E-02	1,87	1,48
FLJ11127	668625	MNG	88	2,30E-06	4,90	1,54
RODH	471641	MNG	83	1,51E-05	4,46	2,57
MGC17299	713422	MNG	59	6,54E-05	4,10	1,37
NR2F2	72744	MNG	57	3,10E-04	3,69	1,40
ZNF83	486356	MNG	50	3,72E-04	3,64	1,32
IMAGE:744505	744505	MNG	43	2,51E-04	3,74	1,38
IMAGE:745187	745187	MNG	42	1,15E-03	-3,31	0,78
CAV1	309645	MNG	41	3,38E-04	3,66	1,59
CNTN6	257773	MNG	41	2,34E-04	3,76	1,38
C10orf116	740941	MNG	40	2,82E-04	3,71	1,72
CALM1	594510	MNG	39	1,21E-03	-3,30	0,77
IMAGE:726782	726782	MNG	39	3,97E-04	3,62	1,38
IMAGE:250463	250463	MNG	28	6,76E-03	2,74	1,29
IMAGE:731726	731726	MNG	28	3,90E-04	3,62	1,52
TTC17	665668	MNG	28	5,24E-04	3,54	1,31
PRKCA	768246	MNG	25	9,12E-04	-3,38	0,71
NR4A1	1073288	MNG	19	1,88E-03	-3,16	0,64
PDLIM1	135689	MNG	14	9,76E-04	3,36	1,48
IMAGE:731292	731292	MNG	11	2,65E-02	-2,24	0,68
SLC17A2	207920	MNG	9	2,70E-02	-2,23	0,67
MLL	80688	OTA	76	1,00E-07	-8,18	0,55
ALDH1A3	486189	OTA	67	1,00E-07	6,38	1,79
CLECSF12	258865	OTA	63	1,00E-07	7,66	3,51
IMAGE:485104	485104	OTA	60	1,00E-07	7,32	2,25
DCN	666410	OTA	49	1,00E-07	-7,19	0,50
PROZ	430471	OTA	47	1,00E-07	7,05	2,76
CHRDL2	485872	OTA	46	1,00E-07	6,72	1,58
B3GNTL1	53341	OTA	39	9,00E-07	5,10	2,04
SNRP70	729971	OTA	39	1,00E-07	-7,00	0,52
HSPC159	365045	OTA	38	1,00E-07	6,86	1,70
C4A	724366	OTA	32	1,00E-07	-6,30	0,39
SLC6A8	725877	OTA	31	1,00E-07	-7,00	0,23
OBSCN	730926	OTA	29	1,00E-07	6,95	1,55
IMAGE:731616	731616	OTA	27	1,00E-07	6,88	2,42
DC12	724895	OTA	24	1,00E-07	6,67	1,71
UBA52	530069	OTA	16	4,40E-06	-4,75	0,29

KCNQ2	179534	OTA	14	1,00E-07	6,43	1,77
HCNGP	365934	OTA	11	1,92E-05	4,40	1,40
C9orf72	726849	OTA	10	4,40E-06	-4,75	0,69
GNAI1	753215	OTA	6	8,90E-06	4,59	1,48
SDHA	40304	OTC	95	1,00E-07	6,29	3,16
DNALI1	782688	OTC	72	1,00E-07	-5,53	0,24
IMAGE:382423	382423	OTC	50	1,50E-06	5,00	3,52
ATP11C	667991	OTC	47	2,30E-05	4,36	2,17
C1QBP	173371	OTC	46	2,70E-06	4,86	2,47
DCN	666410	OTC	43	6,80E-06	-4,65	0,29
MT1F	78353	OTC	41	3,70E-06	-4,79	0,16
KIAA0543	175080	OTC	38	9,15E-05	4,01	2,03
FBXO46	471664	OTC	37	1,67E-05	4,44	2,85
CD33	1917430	OTC	35	1,06E-05	4,55	3,71
TH	813654	OTC	32	8,32E-04	3,41	1,95
IMAGE:742061	742061	OTC	31	1,99E-05	4,39	2,73
IDH3B	723755	OTC	27	1,90E-05	4,41	2,27
IMAGE:359454	359454	OTC	23	1,31E-05	-4,50	0,27
SST	39593	OTC	23	1,34E-03	3,26	1,81
DHRS6	364412	OTC	20	4,03E-04	-3,61	0,48
IMAGE:136933	136933	OTC	19	6,25E-03	2,77	2,19
TPM2	740620	OTC	18	1,59E-05	-4,45	0,30
TM7SF3	666928	OTC	14	3,58E-04	3,65	2,00
PREB	740347	OTC	5	1,29E-02	-2,51	0,33
LRRIQ2	487152	OTUM	59	1,29E-04	3,92	2,01
LRRC28	26519	OTUM	53	2,04E-04	-3,80	0,43
NDUFC1	796513	OTUM	53	2,65E-03	-3,05	0,61
NUMBL	1855110	OTUM	53	2,94E-04	-3,70	0,56
IMAGE:667527	667527	OTUM	48	2,31E-04	3,77	2,45
MAK	382002	OTUM	43	3,71E-04	3,64	2,14
PCGF4	740457	OTUM	43	3,06E-04	3,69	2,61
IFITM3	713623	OTUM	42	1,74E-02	2,40	1,63
CLGN	1049033	OTUM	40	1,52E-03	3,23	1,86
DEFB1	665086	OTUM	40	9,17E-04	3,38	2,79
IMAGE:669098	669098	OTUM	35	9,47E-04	3,37	1,85
GNAS	382791	OTUM	31	6,58E-04	3,47	2,75
PTTG1	742935	OTUM	29	7,53E-03	-2,71	0,52
USP14	376462	OTUM	28	9,21E-03	-2,64	0,65
SNF7DC2	730401	OTUM	25	1,36E-02	2,49	1,51
C21orf84	730814	OTUM	19	1,24E-03	3,29	2,08
TCTE3	136862	OTUM	19	2,48E-02	-2,27	0,69
TH	813654	OTUM	19	3,81E-02	-2,09	0,69
BAK1	235938	OTUM	16	1,12E-03	3,32	1,80
PAX6	230882	OTUM	12	1,28E-02	-2,52	0,66
CDH3	359051	PTC	99	1,00E-07	9,16	3,95
DPP4	343987	PTC	98	1,00E-07	8,72	4,22
CLDN1	594279	PTC	86	1,00E-07	7,88	8,98

SEP-02	729542	PTC	83	2,60E-01	1,13	1,20
ABCC3	208097	PTC	74	1,00E-07	6,30	3,48
ECM1	301122	PTC	67	1,00E-07	8,24	6,00
TSC	745490	PTC	60	1,00E-07	7,70	4,65
QPCT	711918	PTC	57	1,00E-07	7,27	3,36
ADAMTS9	376153	PTC	48	6,40E-06	4,66	1,95
IMAGE:687667	687667	PTC	48	1,00E-07	7,10	3,74
SLPI	378813	PTC	46	1,00E-07	7,03	2,98
MATN1	1624260	PTC	43	1,00E-07	-6,74	0,42
CITED1	265558	PTC	40	1,00E-07	6,84	4,88
CCND1	324079	PTC	38	1,00E-07	6,39	2,55
IMAGE:136976	136976	PTC	23	1,00E-07	6,15	2,11
PET112L	743125	PTC	20	1,00E-07	6,27	1,96
DAF	627107	PTC	16	1,00E-07	5,68	1,89
GSS	140405	PTC	7	1,00E-07	5,73	2,05
CASP3	823680	PTC	5	6,00E-07	5,20	2,08
CD9	727251	PTC	1	1,45E-05	4,47	1,57
IMAGE:687667	687667	TUM	94	1,00E-07	6,33	3,90
CREB3L2	136399	TUM	63	4,90E-06	-4,73	0,48
EIF1AY	380394	TUM	55	4,02E-05	-4,22	0,57
DJ462O23.2	738970	TUM	51	1,64E-05	4,44	1,63
CITED1	265558	TUM	49	1,21E-05	4,51	3,53
KLF1	208991	TUM	46	3,72E-05	-4,24	0,43
SNCB	50202	TUM	45	4,68E-05	4,18	1,99
FLJ10748	726830	TUM	44	4,24E-05	4,21	1,71
ANK2	2139315	TUM	43	1,31E-03	3,27	1,68
IGFBP3	269873	TUM	39	9,68E-05	4,00	2,06
WDR1	714196	TUM	37	2,06E-04	-3,80	0,65
IFNAR2	123950	TUM	36	2,56E-04	-3,74	0,61
SCNN1A	741305	TUM	36	2,37E-05	-4,35	0,41
TSC	745490	TUM	27	1,95E-04	3,81	2,61
NBL1	503874	TUM	24	2,47E-04	-3,75	0,35
ETF1	146976	TUM	18	6,36E-04	3,48	1,63
ICOSLG	2074228	TUM	18	1,12E-03	3,32	1,55
NCB5OR	743367	TUM	15	2,03E-03	-3,14	0,56
ZNF415	23939	TUM	14	4,94E-03	-2,85	0,63
KLK2	1102600	TUM	9	1,39E-03	-3,25	0,66

AT, Autoimmune thyroïditis ; FTA: Macrofollicular thyroid adenoma (and microfollicular, FTAb); FTC: Follicular thyroid carcinoma ; GD, Grave's disease; GMN, Multinodular goiter; OTA: Oncocytic thyroid adenoma; OTC: Oncocytic thyroid carcinoma; PTC: Papillary thyroid carcinoma; TUM, Tumor of uncertain malignancy (OTUM, oncocytic variant);

Table 3: Inter-dataset cross-validations.
Positive accuracies of the classifiers defined by the Fontaine dataset in five other datasets.

DATASET	*PTC*	*FTC*	*FTA*	*OTC*	*OTA*
Giordano *et al.* 2006	0,93	0,40	0,48	0,64	0,59
Weber *et al.* 2005	-	0,77	0,96	0,72	-
Jarzab *et al.* 2005	1,00	-	-	-	-
Reyes *et al.* 2006	1,00	-	-	-	-
He *et al.* 2005	1,00	-	-	-	-

FTA: Macrofollicular thyroid adenoma; FTC: Follicular thyroid carcinoma ; OTA: Oncocytic thyroid adenoma; OTC: Oncocytic thyroid carcinoma; PTC: Papillary thyroid carcinoma

Table 4: Cross validated differential genes from the two main datasets (n=104)

Cluster	Symbol	UGCluster	Clone ID	P-value *(Fontaine et al. 2008)*	P-value *(Giordano et al. 2006)*
1	CASP10	Hs.5353	241481	1,00E-07	1,00E-07
1	CLGN	Hs.86368	1049033	1,00E-07	1,00E-07
1	DEFB1	Hs.32949	665086	1,00E-07	1,00E-07
1	NTRK2	Hs.494312	2048801	1,00E-07	1,00E-07
1	SDHA	Hs.440475	40304	1,00E-07	1,00E-07
1	HSPC159	Hs.372208	365045	1,00E-07	3,00E-07
1	ALDH1A3	Hs.459538	486189	1,00E-07	8,00E-07
1	PLCG2	Hs.413111	201467	1,00E-07	9,00E-07
1	ARVCF	Hs.370408	40633	1,00E-07	1,50E-06
1	ELK4	Hs.497520	236155	1,00E-07	1,50E-06
1	DCI	Hs.403436	667892	1,00E-07	1,70E-06
1	DIO1	Hs.251415	296702	1,00E-07	3,50E-06
1	CKS2	Hs.83758	725454	2,00E-07	9,00E-07
1	ID1	Hs.504609	1087348	2,00E-07	3,90E-06
1	MRPS11	Hs.111286	471574	4,00E-07	1,00E-07
1	ST6GALNAC2	Hs.592105	823590	6,00E-07	8,00E-07
1	GFPT2	Hs.696497	485085	9,00E-07	1,00E-07
1	PITPNC1	Hs.591185	364436	9,00E-07	2,00E-07
1	HRK	Hs.87247	767779	1,00E-06	1,00E-07
1	MAPRE2	Hs.532824	383868	1,20E-06	1,00E-07
1	DOK5	Hs.656582	25664	1,30E-06	1,00E-07
1	RPS6KA5	Hs.510225	258966	1,40E-06	1,00E-07
1	DCXR	Hs.9857	724596	3,00E-06	1,00E-07
1	ODAM	Hs.143811	364706	3,30E-06	1,00E-07
1	KLK2	Hs.515560	1102600	5,00E-06	1,30E-06
1	F5	Hs.30054	433155	6,60E-06	1,00E-07
1	MFAP3L	Hs.593942	726821	7,60E-06	1,00E-07
1	GALNT7	Hs.548088	381854	1,10E-05	1,00E-07
2	AHNAK	Hs.502756	238840	1,00E-07	1,00E-07
2	APOD	Hs.522555	838611	1,00E-07	1,00E-07
2	CAMK2N1	Hs.197922	173820	1,00E-07	1,00E-07
2	CFD	Hs.155597	666128	1,00E-07	1,00E-07
2	FCGBP	Hs.111732	154172	1,00E-07	1,00E-07
2	FOXO1	Hs.370666	151247	1,00E-07	1,00E-07
2	GLT8D2	Hs.631650	365271	1,00E-07	1,00E-07
2	MATN2	Hs.189445	28584	1,00E-07	1,00E-07
2	PRUNE	Hs.78524	364324	1,00E-07	1,00E-07
2	SLC4A4	Hs.5462	787938	1,00E-07	1,00E-07
2	PAX8	Hs.469728	545475	1,00E-07	3,00E-07
2	HSD17B6	Hs.524513	471641	1,00E-07	1,30E-06
2	HBB	Hs.523443	469549	1,00E-07	2,60E-06
2	APOE	Hs.654439	1870594	1,00E-07	3,70E-06

2	CRABP1	Hs.346950	739193	1,00E-07	4,20E-06
2	ZNF76	Hs.388024	745003	1,00E-07	4,70E-06
2	C20orf19	Hs.187635	366032	1,00E-07	5,00E-06
2	DEPDC6	Hs.112981	669318	2,00E-07	1,00E-07
2	CDH16	Hs.513660	726763	5,00E-07	1,00E-07
2	HBA2	Hs.654744	469647	6,00E-07	1,80E-06
2	SASH1	Hs.193133	31120	1,40E-06	1,00E-07
2	HEXIM1	Hs.15299	342551	2,20E-06	1,00E-07
2	GJA1	Hs.74471	839101	2,30E-06	1,40E-06
2	TIMP3	Hs.644633	501476	4,30E-06	4,80E-06
2	IL11RA	Hs.591088	1101773	5,50E-06	7,00E-07
3	APLP2	Hs.695920	549054	1,00E-07	1,00E-07
3	DCN	Hs.694789	666410	1,00E-07	1,00E-07
3	FRMD4B	Hs.371681	669564	1,00E-07	1,00E-07
3	HMGB2	Hs.434953	884365	1,00E-07	1,00E-07
3	IGF2BP2	Hs.35354	743774	1,00E-07	1,00E-07
3	LOC400451	Hs.27373	667174	1,00E-07	1,00E-07
3	NPAL3	Hs.523442	738970	1,00E-07	1,00E-07
3	PBXIP1	Hs.505806	366042	1,00E-07	1,00E-07
3	PDGFRL	Hs.458573	139242	1,00E-07	1,00E-07
3	PDLIM1	Hs.368525	135689	1,00E-07	1,00E-07
3	PSAT1	Hs.494261	366388	1,00E-07	1,00E-07
3	SF3B1	Hs.632554	739247	1,00E-07	1,00E-07
3	TIAM1	Hs.517228	23612	1,00E-07	1,00E-07
3	TSHR	Hs.160411	565317	1,00E-07	1,00E-07
3	FAM111A	Hs.150651	137454	1,00E-07	4,00E-07
3	CA11	Hs.428446	282587	3,00E-07	1,00E-07
3	HNRPA3	Hs.516539	365349	5,00E-07	2,60E-06
4	ABCC3	Hs.463421	208097	1,00E-07	1,00E-07
4	CDH3	Hs.461074	359051	1,00E-07	1,00E-07
4	CITED1	Hs.40403	265558	1,00E-07	1,00E-07
4	CLDN1	Hs.439060	594279	1,00E-07	1,00E-07
4	DPP4	Hs.368912	343987	1,00E-07	1,00E-07
4	DUSP5	Hs.2128	33285	1,00E-07	1,00E-07
4	NOTCH2	Hs.487360	1641901	1,00E-07	1,00E-07
4	QPCT	Hs.79033	711918	1,00E-07	1,00E-07
4	S100A4	Hs.654444	868577	1,00E-07	1,00E-07
4	SCEL	Hs.534699	668239	1,00E-07	1,00E-07
4	SLPI	Hs.517070	378813	1,00E-07	1,00E-07
4	TESC	Hs.525709	745490	1,00E-07	1,00E-07
4	TIMP1	Hs.522632	162246	1,00E-07	1,00E-07
4	MGAT3	Hs.276808	731060	1,00E-07	8,00E-07
4	MDK	Hs.82045	309009	2,00E-07	1,00E-07
4	BID	Hs.591054	128065	6,00E-07	1,00E-07
4	NRIP1	Hs.155017	38775	6,00E-07	1,00E-07
4	COL8A2	Hs.353001	486204	1,00E-06	1,00E-07
4	MRC2	Hs.7835	235882	1,60E-06	9,00E-07

4	STAT1	Hs.699271	110101	2,10E-06	1,00E-07
4	ARMCX3	Hs.592225	251452	2,40E-06	1,00E-07
4	LAMB3	Hs.497636	1103402	9,90E-06	1,00E-07
4	CD55	Hs.527653	627107	1,30E-05	1,00E-07
5	DUS4L	Hs.97627	726904	1,00E-07	1,00E-07
5	ESRRG	Hs.444225	44064	1,00E-07	1,00E-07
5	ISG20	Hs.459265	740604	1,00E-07	1,00E-07
5	ROR2	Hs.644776	1089025	1,00E-07	1,00E-07
5	S100A11	Hs.417004	143957	1,00E-07	1,00E-07
5	SEMA4D	Hs.655281	210587	1,00E-07	1,00E-07
5	CD58	Hs.34341	2103105	1,00E-07	2,00E-07
5	IRF9	Hs.1706	724588	1,10E-06	3,20E-06
5	MCOLN1	Hs.631858	726156	2,00E-06	2,00E-07
5	ITGB7	Hs.654470	1337232	1,09E-05	2,90E-06

Table 5: Datasets, classes and samples.

Thyroid tissue class	*Symbol*	*Fontaine et al. 2008*	*Giordano et al. 2006*	*Weber et al. 2005*	*Jarzab et al. 2005*	*He et al. 2005*	*Reyes et al. 2006*
Microarray platform		Single channel cDNA microarrays	Affymetrix GeneChip HG U133A	Affymetrix GeneChip HG U133A	Affymetrix GeneChip HG U133A	Affymetrix GeneChip HG U133 Plus 2.0	Affymetrix GeneChip HG U133 Plus 2.0
(probes)		*(9216)*	*(22283)*	*(22283)*	*(22283)*	*(54681)*	*(54681)*
Tissue samples		166	93	24	32	18	14
Wild Type tissue	WT	24	4	-	16	9	7
Autoimmune Thyroiditis	AT	5	-	-	-	-	-
Grave's Disease	GD	5	-	-	-	-	-
Follicular Thyroid Adenoma	FTA	26	10	12	-	-	-
Microfollicular Thyroid Adenoma	FTAb	17	-	-	-	-	-
Multi Nodular Goitre	MNG	24	-	-	-	-	-
Follicular Thyroid Carcinoma	FTC	3	-	6	-	-	-
- FTC with PAX8/PPARγ fusion	FTC+	-	7	-	-	-	-
- FTC without PAX8/PPARγ fusion	FTC-	-	6	-	-	-	-
Oncocytic Thyroid Adenoma	OTA	30	7	-	-	-	-
Oncocytic Thyroid Carcinoma	OTC	4	8	6	-	-	-
Oncocytic Tumor of Uncertain Malignancy	OTUM	5	-	-	-	-	-
Papillary Thyroid Carcinoma	PTC	13	51	-	16	9	7
Tumor of Uncertain Malignancy	TUM	10	-	-	-	-	-

Table 6: Primers sequence used for the real time quantitative RT-PCR study

Gene	Forward Sequence	Reverse Sequence
ADAMTS2	5'-GGG-AAG-GAG-CAC-GTA-CAG-AA-3'	5'-CGT-GAT-CGT-GGT-ATT-CAT-CG-3'
ALDH1A3	5'-TCT-CGA-CAA-AGC-CCT-GAA-GT-3'	5'-GTC-CGA-TGT-TTG-AGG-AAG-GA-3'
APLP2	5'-GCT-GTC-GTT-CCG-GTT-ATG-TT-3'	5'-GAG-GGT-CTC-TCA-CGT-GCT-TC-3'
CABIN1	5'-AGG-CCC-TGG-AGG-TGT-ACT-TT-3'	5'-CCA-GCT-GAA-GAG-TGG-GAG-TC-3'
CASP10	5'-TCT-CAG-GAT-CAC-TGG-GCT-CT-3'	5'-ATG-AAG-GCG-TTA-ACC-ACA-GG-3'
CDH16	5'-CAA-GTC-ATG-AGG-TGG-TGG-TG-3'	5'-GAT-GGT-CAG-CAG-GAA-AGA-GC-3'
CLGN	5'-CAA-TGG-ACC-TGG-AAG-AGG-AA-3'	5'-TGA-CTT-TAT-CGG-CCC-ATC-TC-3'
CRABP1	5'-CAG-GAC-GGG-GAT-CAG-TTC-TA-3'	5'-CGC-CAA-ACG-TCA-GGA-TAA-GT-3'
HMGB2	5'-AGA-GGC-TGA-GGA-TTG-CGT-TA-3'	5'-GGG-GTC-TCC-TTT-ACC-CAT-GT-3'
KRTHB5	5'-AGC-TCT-CAG-GGA-CAA-GAC-CA-3'	5'-AAC-AGG-TTA-GCC-CAG-AAG-CA-3'
NR2F2	5'-TGC-CTG-TGG-TCT-CTC-TGA-TG-3'	5'-ATA-TCC-CGG-ATG-AGG-GTT-TC-3'
USP13	5'-GAC-CTG-CGA-GAA-AAC-CTC-TG-3'	5'-CAG-GAG-TGA-TGG-TTC-CCA-GT-3'
β-GLOBIN	5'-CAA-CTT-CAT-CCA-CGT-TCA-CC-3'	5'-ACA-CAA-CTG-TGT-TCA-CTA-GC-3'

3. TRANSCRIPTOME + MOTIF

Article Expression profiling of prospero in the Drosophila larval chemosensory organ: Between growth and outgrowth (en révision BMC Genomics)

Laure Guenin, Mahatsangy Raharijaona, Rémi Houlgatte, Fawzia Baba-Aïssa

Le complexe antenno-maxilaire (Antenno-maxillary complex, AMC) forme le complexe chimio-sensoriel externe de la région antérieur de la larve de drosophile. Des études ont montré qu'un allèle mutant du facteur de transcription *prospero* (*Voila1, V1*) était associé avec plusieurs altérations dans l'AMC et dans le Système Nerveux Central (Central Nervous System CNS). La plupart des homozygotes *V1* meurent avant le stade de pupe, les survivants sont beaucoup plus petits que les individus sauvages et sont affectés dans leur réponse au sel et au sucrose. Plusieurs mutants caractérisés expriment la protéine Prospero (Mestre-Escorihuela, Rubio-Moscardo et al.) à différents niveaux. Par leur utilisation, des études suggèrent que le niveau d'expression de *Pros* dans la région précurseur embryonnaire de l'AMC est lié avec le degré d'altération de l'organe gustatif de la larve.

Le gène *Pros* code pour un facteur de transcription qui possède un domaine protéique contenant plusieurs résidus caractéristiques des homéodomaines connus mais plus basique que ces derniers, et un domaine Prospero conservé. Ces deux domaines protéiques sont nécessaires pour l'adressage au noyau et à la liaison sur l'ADN. *Pros* est connu pour être exprimé dans les cellules précurseurs neurales et participent à la décision du devenir cellulaire des neuroblastes et du lignage des organes sensoriels. *Pros* contrôle la croissance axonale et dendritique, le développement glial. C'est aussi un régulateur de l'activité mitotique dans l'embryon. Il est possible qu'un rôle-clé de *Pros* dans l'AMC larvaire soit le contrôle du développement neuronal. Cependant, il semble qu'il régule des gènes qui soient impliqués dans d'autres fonctions puisqu'il est également exprimé dans des cellules non neurales (cellules accessoires : cellules à l'origine des soies…) dans l'AMC, mais pas dans les cellules gliales.

Afin d'identifier les cibles potentielles de *Pros* et les fonctions que ce facteur pouvait contrôler, nous avons réalisé une analyse du transcriptome du tissu AMC par puce à ADN de différents mutants *Pros* au stade larvaire. 7500 clones de cDNA ont été déposés sur une puce sur support

nylon. 17 échantillons d'ARN ont été analysés. Nous avons utilisé des individus sauvages, des individus portant la mutation *V1, V13 (prosVoila13)* ou *V24 (prosVoila24,* voir *Table1,* et présentant chacun des niveaux d'expression différents de *Pros*. Nous avons inclus des échantillons CNS pour ces quatre allèles pour établir la spécificité à l'AMC.

Dans ce travail j'ai participé à l'analyse du transcriptome et à la découverte d'un motif partagé dans les séquences promotrices de gènes dont l'expression est dépendante de Pros.
La partie expérimentale des puces a été réalisée par Laure Guenin, premier auteur de cet article, au sein de la plateforme du laboratoire INSERM TAGC ERM 206 : conception des puces, extraction de l'ARN, reverse-transcription, marquage, hybridation des échantillons sur les puces, acquisition des données et traitement primaire des données (correction vecteur, normalisation, filtrage). A partir de ces données, mon travail a consisté à détecter les fonctions biologiques perturbées chez les mutants *Pros*.

Méthodologie

Analyse des données de puces

Il est établi que les gènes appartenant à une même fonction biologique ou à un même type cellulaire ont une expression corrélée (Eisen, Spellman et al. 1998; Shaffer, Rosenwald et al. 2001). J'ai donc cherché des groupes de gènes ayant une expression corrélée et différentiellement exprimés chez les mutants pour répondre à notre question. J'ai utilisé une méthode de détection de groupes de gènes corrélés et une méthode statistique pour détecter une expression différentielle parmi ces groupes.
Différentes méthodes peuvent détecter des expressions corrélées. J'ai choisi la classification hierarchique qui a comme avantage de ne pas fixer *a priori* le nombre de clusters ni la taille de ces clusters. Pour le clustering, j'ai utilisé le programme Cluster, en utilisant la corrélation de Pearson et l'utilisation de lien centroïde pour les distances inter-groupes. Nous avons utilisé le logiciel TreeView pour la visualisation des données (Eisen, Spellman et al. 1998).
La méthode statistique utilisée pour identifier les gènes différentiellement exprimés entre les individus sauvages et les mutants *Pros* a été le calcul d'un score discriminant (DS, Discriminating Score). Comme le nombre gènes attendus n'est pas connu, j'ai calculé un DS

moyenné sur une fenêtre glissante selon l'ordre du clustering, pour chaque tissu. Cette stratégie nous a permis d'identifier des pics de taille optimale. Ces pics correspondent à des clusters de gènes d'expression corrélée et différentiellement exprimés entre les mutants.

J'ai ensuite annoté fonctionnellement ces différents clusters, en identifiant des enrichissements significatifs de termes de Gene Ontology, et en utilisant GoMiner. La fréquence de chaque terme d'ontologie est comparée avec celle observée dans un cluster par un test exact de fisher. Ces annotations ont été par la suite améliorées manuellement.

Découverte de motifs

Un des clusters que nous avons identifiés est commun aux deux tissus CNS et AMC. Au sein de ce cluster, un groupe de 28 gènes présente une corrélation forte de leur niveau d'expression ($r>0.9$). Cette observation pouvait suggèrer que l'expression des gènes sensibles à *Pros* est contrôlée par des facteurs de transcription communs. J'ai entrepris une découverte de motifs, correspondant aux sites de liaison de facteurs de transcription sur l'ADN, significativement surreprésentés dans le promoteur de ces gènes par rapport au reste de la puce. Cette découverte a été réalisée dans la région -1700 +300 par rapport au site d'initiation de la transcription du génome masqué (sans les éléments répétés du génome) de la drosophile.

Les stratégies existantes de découverte de motifs peuvent se diviser en deux catégories principales : les méthodes basées sur le comptage de mots (oligonucléotides) et les méthodes probabilistes, découvrant des motifs dégénérés.

Pour des raisons biologiques (un facteur de transcription reconnaît plusieurs sites sur l'ADN présentant de légères variabilités), j'ai choisi une méthode probabiliste, l'échantillonnage de Gibbs, pour l'identification de motifs dégénérés, présentés sous forme de matrices. Nous avons décidé de trouver des motifs de taille de 10 nucléotides. Nous n'aurions pu tester toutes les tailles de motifs, pour être le plus exhaustif possible, mais ce choix nous avait paru raisonnable car proche de la taille « moyenne » des sites de liaisons de facteurs de transcription. Les matrices, une fois sous forme de Matrice Poids-Position (PWM), permettent de calculer un score reflétant la présence du motif à une position donnée. La position dans une séquence correspondant au score maximal reflète la plus grande probabilité qu'à cette position se trouve le motif.

Pour valider statistiquement le motif sur nos données, j'ai comparé la distribution des scores maximaux, calculé à partir de la PWM, du groupe testé à celle d'un contôle négatif. Ce contrôle

négatif est constitué de l'ensemble des promoteurs présents sur la puce, en excluant les promoteurs testés, correspondant au bruit de fond. Un Chi² d'homogénéité a été utilisé pour cette comparaison (H0 : pas de différence de distribution).

Résultat

Les gènes et échantillons ont d'abord été classés selon une classification hiérarchique. Cette approche a permis d'identifier un cluster de gènes différentiellement exprimés entre l'AMC et le CNS, indépendamment de *Pros*. L'association de cette classification et du calcul d'un score discriminant a permis d'identifier d'autres clusters de gènes d'expression corrélée et différentielle dans chaque organe entre les individus mutants et les individus sauvages. Trois clusters, dont un commun à l'AMC et au CNS ont été identifiés. Un motif commun, pouvant correspondre au site de liaison putatif de *Pros* est significativement surreprésenté ($p=0.04$) dans le promoteur des gènes sensibles à la mutation de *Pros*, communs à l'AMC et au CNS et fortement corrélés ($r>0.9$). Des fonctions ont pu être attribuées à chaque cluster : un premier impliqué dans la croissance des neurites, un deuxième dans le système olfactif de la larve, et un troisième impliquant l'autophagie et la croissance. Les voies de signalisation *EGFR* et *Notch* sont représentées dans ces trois clusters.

Nous proposons que Pros réalise ces différentes fonctions en modulant ces deux voies antagonistes et synergiques.

Cette étude contribue à la clarification de fonction de Prospero dans l'AMC de la larve de drosophile. Le mécanisme par lequel Pros pourrait accomplir sa fonction devra être explorée en détail. Bien que ce motif semble intéressant, rien n'indique que ces gènes sont des cibles directes de *Pros*, et aucune preuve biologique dans notre étude valide sa liaison sur ce motif. Une approche par ChIP-chip étudiant les cibles de *Pros* et des facteurs impliqués dans la voie *Notch/EGFR* validerait nos résultats.

Expression profiling of prospero in the Drosophila larval chemosensory organ: Between growth and outgrowth

Laure Guenin[1,2], Mahatsangy Raharijaona[3,4], Rémi Houlgatte [3,4], Fawzia Baba-Aïssa [2§]

[1] Institut Pasteur, Pathogénomique Mycobactérienne Intégrée, 25, Rue du Dr. Roux 75724 Paris Cedex 15, France
[2] Université de Bourgogne, Facultés des Sciences, Unité Mixte de Recherche 5548 Associée au Centre National de la Recherche Scientifique, 6, Bd Gabriel, 21 000 Dijon, France.
[3] INSERM, U915, Nantes, F-44000 France
[4] Université de Nantes, l'Institut du Thorax, Nantes, F-44000 France
[§]To whom correspondence should be addressed (tel: 33(0)380396297; fax: 33(0)380396289 E-mail: fawzia.baba-aissa @u-bourgogne.fr)

Abstract

Background

The antenno-maxilary complex (AMC) forms the chemosensory system of the *Drosophila* larva, and is involved in gustatory and olfactory perception. We have previously shown that a mutant allele of the homeodomain transcription factor Prospero (*prosVoila1*, *V1*), presents several developmental defects including abnormal growth and altered taste responses. In addition, many neural tracts connecting the AMC to the central nervous system (CNS) were affected. Our earlier reports on larval AMC did not argue in favour of a role of *pros* in cell fate decision, but strongly suggested that *pros* could be involved in the control of other aspect of neuronal development. In order to identify these functions, we used microarray analysis of larval AMC and CNS tissue isolated from the wild type, and three other previously characterised *prospero* alleles, including the *V1* mutant, considered as a null allele for the AMC.

Results

A total of 17 samples were first analysed with hierarchical clustering. To determine those genes affected by loss of *pros* function, we calculated a discriminating score reflecting the differential expression between V1 mutant and other *pros* alleles. We identified a total of 64 genes in the AMC. The construction of a database that compiled all of the information available on the role attributed to each of these genes in the *Drosophila* larvae nervous system, enabled us to identify a first functional category of potential Prospero target genes known to be involved in neurite outgrowth, synaptic transmission and more specifically in neuronal connectivity remodelling. The second category of genes found to be differentially expressed between the null mutant AMC and the other alleles concerned the development of the sensory organs and more particularly the larval olfactory system. Surprisingly, a third category emerged from our analyses and suggests an association of *pros* with the genes that regulate autophagy, growth and insulin pathways. Interestingly, EGFR and Notch pathways were represented in all of these three functional categories. We now propose that Pros could perform all of these different functions through the modulation of these two antagonistic and synergic pathways.

Conclusions

The current data contribute to the clarification of the Prospero function in the larval AMC and show that *pros* regulates different function in larvae as compared to those controlled by this gene in embryos. In the future, the possible mechanism by which Pros could achieve its function in the AMC and the possible involvement of EGFR and Notch pathway will be explored in detail.

Background

In *Drosophila,* some external sensory organs found in the anterior region of larvae are composed of many neurons and support cells that seem to represent an aggregation of several sensory units. This is the case for the antenno-maxillary complex (AMC) that forms the chemosensory system of the *Drosophila* larva. The chemosensory apparatus of the larval head is formed during embryogenesis [1] and consists essentially of three major sensilla complexes on the cephalic lobe, the dorsal (DO), terminal (TO) and ventral organs (VO), and a series of pharyngeal sensilla [2, 3]. While the DO appears to be a mixed smell and taste organ, the TO, VO and pharyngeal sensilla may be exclusively gustatory [4-8].

In previous studies, we described a mutant allele of the transcription factor *prospero* (*Voila1, V1*) that is associated with several alterations in both AMC and the CNS [9]. *V1* homozygotes died before forming pupae. Surviving larvae remain much smaller than wild-type individuals and are impaired for their response to salt and sucrose [10]. Using a set of previously characterised *Voila* alleles (*prosV*) that express different levels of Prospero (Mestre-Escorihuela, Rubio-Moscardo et al.) protein, we found that the level of Pros expression detected in the embryonic precursor region of the AMC, was related to the degree of alteration of larval taste [11]. In embryonic and larval AMC, Pros is expressed in the same cell cluster (~50 cells), including neuronal cells (~10 cells) and many accessory cells but no glial cells [9] .

The pros gene encodes a transcription factor protein that contains a highly divergent putative homeodomain and a conserved Prospero domain that are both necessary for sequence-specific DNA binding and Prospero nuclear localisation [12-14]. Pros is known to be expressed in neuronal precursor cells [15, 16] and participates in cell fate decision in both neuroblasts and sensory organ lineages [17-20]. Pros has been shown to control axonal and dendritic outgrowth [16], glial development [21, 22] and to be a key regulator of mitotic activity in embryos [23]. Pros affects several cell cycle genes and can either promote or inhibit them depending on the cellular or the developmental context [24, 25]. More recently Choksi et al. [26] showed that in the embryonic nerve cord Pros repressed target genes such as cell cycle genes required for self-renewal, and was also required to activate genes involved in terminal differentiation.

In a previous study on the larval antenno-maxilary complex, we showed that loss of *pros* function did not alter the mitotic activity or the final number of neurons. By contrast, many neural tracts connecting the AMC to the CNS are affected [9]. Therefore, it is possible that one key role of Pros in the larval AMC is to control the expression of genes involved in neuron-specific development such as axon routing and or neurite outgrowth. However, as Pros is expressed in non-neuronal cells in the AMC (accessory cells), it is likely that, *it* regulates genes that are also involved in other functions.

In order to identify the Pros target genes associated with this organ, we performed microarray analysis on larval AMC tissue isolated from the wild type, the *V1* mutant and two previously characterised *prospero* alleles [9], *V13 (prosVoila13)* and *V24* (*prosVoila24,* see also Table1). To establish the AMC specificity of these genes, we included analysis of samples from isolated larval CNS for these four alleles. Our findings indicate that, in this sensory organ, *pros* is mainly associated with the regulation of genes that are essential for correct routing of neural processes and synaptic transmission. Many of these genes are involved in the development and remodelling of the nervous system during metamorphosis. Interestingly, we also found that loss of *pros* function induced the misregulation of a subset of genes important for growth, and autophagy. Finally, our data suggest that the *EGFR* (the epidermal growth factor receptor) and the *N* (Notch) pathway could play a central role in regulating all of these functions.

Results

The developing AMC and Pros expression

We have previously shown [9], that in the developing AMC, Pros is always expressed in the same cluster of cells. In addition, neither mitotic activity nor apoptosis was observed during the third instar larval stage or at late embryonic stages suggesting that the final number of Pros expressing cells is fixed before the end of embryogenesis [9]. This hypothesis was further confirmed by analysing mitotic activity in the developing wild type AMC using an H3p marker (*prosV14, V14)*. Our results showed that the H3p labelling disappeared completely after the embryonic stages 12-13 (see supplemental data Fig. S1) indicating that additional cells are not provided until the last larval stage. However, some Pros expressing cells grew in size at the LIII stage. Scoring the different Pros+ cell type morphology (Fig. 1, Table 2), we found that the wild type larval AMC (Terminal organ and dorsal organ) was composed of 8 (±1) large Pros+ cells (most probably accessory cells) and 40 (±4.1) small Pros+ cells. Among the latter, 10.7 (± 2.8) were neuronal cells [9]. Pros is never expressed in glial cells. Interestingly, loss of function affected the axonal pathway in the AMC in the larval AMC, but produced the correct number of neuronal cells but [9] and curiously induced an excess of glial cells, which, we suspect, originated from incorrect peripheral glial cells migration. Therefore, if Pros was expressed in the same number of cells in the embryo and larvae and since no additional cell division was seen after stage 13, it is likely that *Pros* is not involved in cell fate choice in the larval AMC.

To better clarify the role of *pros* in the AMC, we carried out microarray analysis on wild type (*V14*) and three previously characterised *prospero* mutants *(V1, V13, V24),* which present different expression levels of Pros [9, 11]. These alleles were selected as they present distinct phenotypes either in the larval AMC and/or in the CNS (see also Table 1). *V1* is considered as a null *pros* allele for the AMC. It presents an abnormal taste response to sucrose and NaCl (indifferent to both substances), and shows an alteration of the neural connections between the AMC and CNS as well as arborisation defects in larval neuromuscular [11]. In the *V1* larval CNS, Pros is still expressed but at a lower level than that in the wild type *V14*. *V1* larval CNS also shows several defects, which include early initiation of cell death and abnormal subcellular localization of the Pros protein.

Both *V13* and *V24* alleles show a normal Pros expression pattern in the larval AMC, but in the CNS both alleles present atypical cell proliferation. While *V13* presents a standard taste response as it is attracted by sucrose and repulsed by NaCl, *V24* shows an intermediate taste response (repulsed by NaCl but remained indifferent to sucrose). Therefore, in larvae, *V14, V13* and *V24* present a normal Pros AMC pattern, distinct taste responses, and different Pros expression in the CNS. All these alleles were used to identify the genes that are differentially expressed between *V1* AMC and the other alleles. To establish the AMC specificity-profile, CNS tissues isolated from the four different alleles were also included in the analysis.

Transcription profile of *pros*^{V1} larvae

A total of 17 samples were analysed, including both CNS and AMC samples for the four *prosV* alleles and 2 to 3 independent RNA extractions for each allele. Gene expression profiles were used to group genes and biological samples according to their overall similarities in gene expression. With regard to this, the 17 expression profiles were first analysed with

hierarchical clustering. This method is known to highlight clusters of correlated expression genes that participate in the same biological function [27]. Results are presented using the TreeView program [27]. As it can be seen, hierarchical clustering performed on the overall data (Fig. 2A) revealed genes that are differentially expressed between AMC and CNS tissues and in a Pros independent manner. The genes of this cluster (AMC tissue specific signature) are clearly overexpressed in the AMC (Fig. 2A) while the same genes are underexpressed for all alleles in the CNS (Fig 2A).

In the next step, we determined those genes affected by loss of *pros* function for each organ. For the AMC, we calculated a discriminating score (DS, see experimental procedures) reflecting the differential expression between wild type and *pros* mutants. As *V13, V14* and *V24* have a normal Pros expression pattern in the AMC, the DS was calculated for each gene between *V1* and all other alleles.

A similar method was used in the CNS to determine the genes affected in the *V1* mutant. However, since *V13* and *V24* exhibited a complex pattern, the DS for each gene was calculated between *V1* and the *V14* alleles only.

To visualize only groups of correlated genes that were differentially expressed between *V1* and other alleles, the DS score obtained for the AMC or CNS was plotted alongside the hierarchical clustering and smoothed using a sliding window of 100 genes. As shown in Fig.2B different peaks can be seen. To avoid the analysis of non-significant variations, we decided to assess the biological functions of these groups of genes. We therefore searched for significant enrichments of Gene Ontology terms (GO) in each cluster using GoMiner [28]. In the AMC, only 3 peaks (peaks 1-3) could be associated with significant GO functions (Fig. 2B peaks 1, 2 and 3). Complete gene lists for peaks 1-3 are presented in Table S1 and S2. In peak 3, the 26 genes overexpressed in *V1* AMC were significantly associated with the overrepresented GO term « signal transducer activity » (GO:0004871, p=0.0008, see also Table 3). Significant enrichments of the GO term "proteasome complex" (GO:0000502, p<10-5, see also Table 3) were found for 9 genes in peak 2. All of these genes were underexpressed in *V1* AMC. Peak 1, with the highest DS, was common to AMC and CNS and associated with the over-represented GO term "cell fate commitment" (GO:0045165, p= 0.0003). Inside this peak, a cluster of 29 genes highly overexpressed in *V1* AMC and highly correlated (r>0.9, Fig. 2C, Table 3) was isolated. Similarly, we found a cluster of 86 genes highly overexpressed in *V1* CNS and correlated (r>0.9, Fig. 2D, see also Table S3) in the same peak. Interestingly, among the 86 CNS genes, 28 also belonged to the AMC gene list (Table 3 and Table S3). Finally, to assess our microarray analysis, we quantified the expression of seven selected genes by Q-PCR. As shown in Table 4, our results were consistent with the microarray data except for the *hb (hunchback)* gene found to be overexpressed in the CNS but not in the AMC. Because the two methods have different sensitivities, the magnitude of the change determined by microarray and real time PCR is not the same. The orientation of changes, however, is identical.

A subset of genes overexpressed in both larval AMC and CNS share a common putative binding site.

To determine whether the 28 candidate genes overexpressed in both AMC and CNS could be regulated by Pros, we searched for a common sequence motif in their promoters. We used a Gibbs sampling method [29] on the -1700 to +300bp promoter region of these 28 genes. Gibbs Sampling allowed us to determine degenerated motifs described by a position weight matrix (PWM) in a set of sequences by iterative sampling. The best PWM found for all promoters is shown as a Logo [30] in Fig. 3A. The consensus is GCAGCTGC, which includes a **CAGCTG** core. Interestingly, this is partially similar to the motif described by Hassan et al [31] **(CA/TC/TNNCTC** see also alignment in Fig. 3B) and identified with a site selection assay (SELEX). A third *pros* binding motif was reported by [32]. These authors showed that in the adult *Drosophila* eye, Prospero, which is specifically expressed in R7 photoreceptors, preferentially recognizes the **AAGACG** sequence (Fig. 3B) present in the promoter of some *rhodopsine* genes (*rh6* and *rh5*) and mediates their repression in these cells. Cook et al [32] also tested the Hassan et al [31] binding site and found that Pros could bind weakly to individual *SELEX sites*, but only when these sites were multimerized. More recently, using both the DamID method and the Cook motif, Choksi et al. [26] identified 1000 Pros target genes in *Drosophila* embryos that contain this sequence. However, further microarray analysis done by these authors on ventral nerve cord cells only, allowed the identification of additional Pros target genes that do not contain this consensus sequence. Therefore, it is likely, as previously suggested by these authors, that Pros exhibits some flexibility in its binding specificity.

Building a database specific to *Drosophila* larvae to be used for manual annotation

Our microarray analyses showed that peak 1 contained 29 overexpressed genes associated with the GO annotation, "cell fate commitment" (Table 3). These data are not consistent with our previous studies showing that *pros* is not involved in cell fate determination in the larval AMC [9]. We were intrigued by this discrepancy and therefore we looked more deeply for the function of these genes in the larvae. Interestingly, though most of the genes present in this peak were associated with cell fate determination in embryos, such evidence was mostly missing for the larvae peripheral nervous system (PNS). Thus, it is likely that the GO annotation was mostly deduced from the reported function of these genes in *Drosophila* embryos. Therefore, to avoid any misinterpretation of the data we decided to use manual annotation.

The first step consisted in constructing a database that compiled all of the information available on the role attributed to each of the 64 genes identified from peaks 1-3, but most specifically in *Drosophila* larvae. The information was collected using Flybase, mutant analysis, associated phenotypes, research articles and microarray data. As much as possible, we selected only data that reported the function of these genes in the larval nervous system and more specifically the sensory system. Out of the 64 genes, we found that 27 had unknown biological functions or had not been documented in larvae (Table 3). Since the genes encoding the subunit of the proteasome complex are under-expressed and the remaining genes are overexpressed in the *V1* AMC, a description of the phenotype generated by the upregulation and downregulation of the corresponding genes in larvae was included in the database (Table S4-S6). Though many phenotypes were available for a gene, we selected only those reported for the nervous system and preferentially for the PNS.

Analysis of the resulting *Drosophila* larval database (Table S4-S6) revealed that the 37 genes fell into at least one of the following functional categories: (1) neurite outgrowth and/or synaptic transmission; (2) growth, autophagy; (3) sensory organ (mainly olfactory) development. A list of these genes and their associated annotation terms is summarized in Table 3, and a schematic representation is given in Fig. 4. As shown in Fig. 4, some genes can be associated with two functional classes, and four genes (*EGFR, Notch, Ash2* and *pros*β) are associated with the three functional categories: neurite outgrowth, autophagy, and olfactory system development.

Genes involved in neural processes and synaptic transmission are misregulated in the *V1* AMC mutant

One of the functional categories deduced from manual annotation associates some putative Pros target genes with neurite outgrowth and/or synaptic transmission (Fig. 4). Although synaptic transmission and neurite outgrowth belong to different functional categories, we decided to keep these genes in the same class since many of them are involved in both synaptic transmission and neurite outgrowth.

Mostly, the genes are overexpressed in the larval *V1* AMC, (Table 3). It has been reported that the upregulation of most of these candidates inhibits neurite outgrowth in larval neurons (reported phenotypes are shown in Table S4). This is clearly the case for *N (Notch), EGFR (Epidermal Growth Factor Receptor), bnl (branchless)* and *Rac1* [33, 34] whose overexpression was previously shown to inhibit axon extension in larval neurons. This is also the case for *gwl (greatwall*), *limK1 (lim-kinase 1)*, *Nej (Nejire)* or *CG6388* whose upregulation induces axon pathfinding defects or impaired neurotransmitter release in the larval neuromuscular junction (NJM) [35-38]. Interestingly, similar axon pathfinding and NMJ defects were observed in the *prosV1* mutant [9]. Mutations in *EGFR*, *pvr* (PDGF/VEGF-receptor related) or *Ash2* (*absent, small, or homeotic discs 2*) also induce abnormal axon guidance in the larval nervous system [34, 39, 40]. However the consequence of overexpression of these genes remains to be clarified.

We have noticed that most of the genes included in this class more specifically drive neural connectivity remodelling in larvae, a process particularly important during metamorphosis. *bnl, Rac1* [34], *EGFR, Notch* [33], and the genes associated with the ubiquitin-proteasome system were all reported to be involved in axon extension/retraction, pruning and morphogenesis of larval peripheral sensory neurons (See also Table S4). Indeed, though most larval sensory neurons will degenerate during metamorphosis, some persist as neurons and undergo stereotyped pruning of their dendrites and axon terminal branches during early metamorphosis [41].

We were not surprised to find that the genes encoding the different proteasome subunits (*Prosα7, Prosβ2, Prosα6, Prosα26)* are downregulated in *V1* AMC (Fig. 4, Table 3). Indeed, the acute regulation of their protein level is a primary determinant of protein turnover and neurotransmission strength [42, 43].

Recently, an elegant study Choksi et al. [26] showed that *pros* is required for activation of neuronal

differentiation genes in embryos and identified *N, bnl, LimKI, EGFR and PVR, prosα6* as putative *pros* targets in embryos. This reinforces our finding suggesting that in the larval AMC, *pros* plays a crucial role in the modulation of neuronal activity through the control of genes involved in neurotransmission and synaptic plasticity.

Loss of *pros* function alters the expression of genes involved in autophagy and growth

The second functional group that emerges from our analysis includes candidates that play a critical role in the control of autophagy, a process used to provide energy and nutrients during metamorphosis and early adulthood.

The association between *pros* and the regulation of autophagy is mainly attested by the upregulation of genes such as *CG10702, EGFR, Keren* (EGFR ligand), *Ftz-F1 (transcription factor 1), FK506-bp1 (FK506-binding protein 1), Iap2 (Inhibitor of apoptosis 2), nej, Notch* and genes associated with proteasome complex (Table 3, Fig. 4). Juhasz et al. [44] showed that *FK506-pb1* is an inhibitor of autophagy (reported phenotypes are shown in Table S5). Mutations in the steroid regulated gene *Ftz-F1* was shown to prevent autophagic programmed cell death [45]. More recently, proteomic analyses have shown that the genes identified in this study and involved in the ubiquitin-proteasome system are highly expressed during autophagic cell death [46]. Therefore, it seems that loss of *pros* function correlates with the inhibition of autophagy.

Some of the genes cited below were also found to mediate cell growth (See also Table 3 and Table S5). However, it is not yet clear if the overexpression of these genes systematically inhibits cell growth. For example, the upregulation of *LK6 (protein serine/threonine kinase)* or *FK506*-bp1 leads to either the activation or inhibition of cell growth in a context-dependent way [47, 48]. The downregulation of *ash2* promotes cell growth, yet the effect of its overexpression is not known. Finally, *EGFR* and *ftz-F1,* and *Notch* were all reported to foster cell growth [49-51]. We have already mentioned that Notch and EGFR pathways were involved in neurite outgrowth (see below). Interestingly, we found that these two pathways were also associated with both the regulation of autophagy [44, 52] and cell growth control (Fig. 4), suggesting that Pros could mediate all of these functions through the modulation of these two pathways.

Our finding that *pros* is associated with the expression of genes involved in growth and or autophagy is consistent with the phenotypic defects observed in *V1* homozygote mutants: i) individuals died before reaching puparium formation ii) surviving larvae and pupae remained much smaller than wild-type individuals; iii) numerous labelled cells were observed in the fat body using PGal4 enhancer trap line *V1* [10].

It is interesting to note that many of the genes found in this functional group are directly or indirectly associated with insulin-signalling pathways and more specifically the insulin/TOR (target of rapamycin) pathway, an important mediator of growth, autophagy and nutrient sensing (Table S5).

Pros and the olfactory system

Pros was detected both in the terminal (TO: mainly gustatory) and in the dorsal (DO: mainly olfactory) organs of the larval AMC [9]. Accordingly, we found that *pros* loss of function in the AMC induced the upregulation of all candidate genes (except *pros*β2) that were known to be involved in the development of sensory organs (Fig. 4, Table 3). Most of our knowledge on the function of these genes came from studies done on adult *Drosophila* sensory organs (see Table S6). For example, it has been reported that mutations in the genes *ash2 or ckII alpha (*Casein kinase II*)*, can elicit supernumerary or ectopic adult sensory organs [53, 54]. Similarly, overexpression of *Iap2 or limK1* induces respectively additional macrochaetes [55] or ectopic glomeruli in adult antennae lobes [56]. The transmembrane receptor *Notch* and the epidermal growth factor receptor *EGFR* also seem to play an important role in the organisation, remodelling and function of the olfactory system (Garcia-Alonso et al. 2000; Endo et al. 2007). This confirms previous observations which showed that they were respectively required for selecting the sensory organ precursor lineages [57, 58] and for the development of some of the neurons and cuticular structures of the antenno-maxillary sensory complex (for review see [59].

Discussion

Pros may regulate genes essential for neurite outgrowth and remodelling

In the AMC, the transcription factor Prospero is expressed in a cluster of cells (composed of neuronal and support cells, but not glial cells) that emerge during embryonic life and are maintained till the end of the larval stages. In embryos, Pros was reported to be involved in cell fate decision and in cell-cycle control. By contrast, our earlier data from the larval

AMC rather suggested that *pros* could assume more restricted functions, such as the control of neuron-specific functions [9]. The present study confirms this hypothesis and shows that in the chemosensory organs dedicated to larval olfactory and gustatory sensing, *prospero* could regulate genes involved in neurite outgrowth and synaptic transmission.

Since *pros* was clearly shown to control axonal and dendritic outgrowth [16, 60], we cannot exclude the possibility that the connection of *pros* with several genes that drive synaptic activity could be the indirect consequence of its involvement in neurite outgrowth control. In this respect, it is interesting to mention that a recent paper [61] showed that axon targeting of the R7 *Drosophila* photoreceptor cells to their synaptic partner requires R7-specific transcription factor Prospero. These authors proposed that Pros could promote cell-typespecific expression of sensory receptors and cell-surface proteins regulating synaptic target specificity.

As previously mentioned, some of the genes identified in this functional class are more specifically involved in neural connectivity remodelling. These genes were reported to control axon extension/ retraction and pruning during metamorphosis. Indeed, with few exceptions, larval sensory neurons will degenerate during metamorphosis to be replaced by adult neurons, which emerge from imaginal discs [62]. In the olfactory system, [41], showed that some neurons participate in both the larval and adult olfactory system and undergo stereotyped pruning of their dendrites and axon terminal branches locally during early metamorphosis. Does Pros play any role in AMC neuronal remodelling? Some elements deserve some thought. First, Pros is associated with the regulation of *Rac1, bnl, EGFR, limK1, N*, all found to be involved in this process (see also Table S4) and previously identified as putative targets of Prospero by Choski and al. [26]. Second, *limK1* was shown to control morphogenesis through the regulation of the expression of several ecdysone-responsive genes, including the ecdysone receptor itself [63]. Similarly, *EGFR* and *Notch* pathways were both found to be activated by the ecdysone regulatory network [64]. In addition, [65] found that Notch was crucial for developmental processes that establish specific neuronal connections and olfactory neuron identities. Finally, the insulin and epidermal growth factor signalling pathways, as well as ubiquitin-specific proteases are all required for the regulation of *Drosophila* neuronal remodelling [66]. Interestingly, all of these components emerge clearly from our analysis . Therefore, it is likely that a relationship between *pros* and neuronal remodelling does exist but should be clarified by further analyses.

In summary, our data collected from larval AMC and the previous genome wide expression profiling done on embryos [26] confirm that *pros* is required for the regulation of neuronal specific genes. In this respect, it is essential to note that except for a few genes (126), most of the Pros target genes identified (~1000) in Choksi et al. [26] were not represented on our microarrays. For this reason, and because our experiments were performed on isolated individual larval tissues, it is not possible to determine whether the genes identified by these authors are specifically expressed in embryos and /or in tissues other than AMC.

Prospero and the insulin pathway

In *Drosophila,* the insulin/TOR signalling pathway [67] is divided into two branches. The insulin and its downstream effectors P13 and FOXO (forkhead box) represent one branch [68] of this pathway, while the other branch acts through the TOR family of Serine-Threonin kinases [69, 70]. It has been shown that the insulin/TOR signalling pathway inhibits autophagy [69] and controls growth by regulating ribosome biogenesis and protein biosynthetic capacity [71, 72]. Columbani et al. [72], also demonstrated that the TOR is a nutritional checkpoint that participates in the systemic control of larval growth emanating from the Fat body.

Our microarray analysis has identified a group of highly correlated *pros* candidate genes (correlation index: 0.9) that may be involved in both autophagy and growth in larvae. Most likely, these genes assume both functions because they are either being controlled by the insulin/TOR signalling pathway or are directly involved in the signalling cascade. This is the case for *Ash2* [73] which was found to be regulated by TOR signalling. Similarly, FK506-bp1 affects autophagy through the modulation of FOXO (Juhàsz et al. 2007) and *Lk6* was reported to be a direct FOXO Target [71]. Moreover, LK6 also designated by Choksi et al. [26] as a potential Pros target gene, was shown to interact with Notch and EGFR [74], two pathways that seem intimately linked to *pros*-mediated neurite outgrowth and remodelling in the larval chemosensory organ. In this respect, a recent report has shown that in larvae, neurogenesis and EGFR are both regulated by the insulin/TOR pathway [75].

Therefore it seems that in the larval AMC, Pros could be associated with growth, autophagy and nutrient sensing through the regulation of genes that are directly or indirectly linked to the insulin/TOR pathway. Interestingly, *TOR* was found to be differentially expressed in the *V1 pros* mutant in the CNS (see supplemental data Table S3). In the AMC,

it seems that Pros is associated with TOR through the regulation of other effectors of the pathway, where LK6, EGFR and Notch could assume a pivotal role.

Conclusion

As previously described, loss of *pros* function in the AMC induced several alterations including axon pathfinding defects and abnormal growth and taste responses. This is consistent with our microarray results showing that in the larval AMC, Pros expression is associated with the regulation of genes involved in the control of neurite outgrowth, mediation of growth and autophagy and in the organisation and function of the olfactory system. The mechanism by which all of these functions are achieved by *pros* in the AMC is presently not known but we can suggest that EGFR and/or Notch pathways present in the 3 main functional groups (see Fig. 4) could play an important role. Several lines of evidence support this hypothesis.

1-The possible regulation of EGFR by Pros is further reinforced by our observations showing that although Prospero is not expressed in the glia of the AMC, the *V1* mutant showed an excess of glial cells in this structure as compared to the wild type. We have previously suggested that this could be due to the abnormal migration of glial cells to the AMC [9]. Indeed, the glial cells present in the AMC derive either from SOP lineages or originate by migration from the CNS (this process starts during embryonic stages). Interestingly, [76] showed that the migration of glial cells to the antennal segment is regulated by the normal function of the Epidermal Growth Factor receptor and small GTPases.

2- Four ligands are known to bind EGFR receptor: Keren, Gurken, Spitz, and Vein. Two of these were identified as potential targets of Prospero: keren in both larval AMC and CNS and gurken (grk) in the larval CNS only (see Table 3 and S3). Moreover, Notch and EGFR were identified as Pros target genes in both embryos [26] and larval AMC, indicating that they could play a central role.

3- It has been reported that EGFR signalling is required for the development of some of the neurons and cuticular structures present in the AMC [77, 78]. In this respect, it is interesting to point out that EGFR involvement has been reported during the development of mouse gustatory epithelia in the palate and tongue [79].

4- The expression of Notch, EGFR and Pros have been shown to be tightly linked. It has been demonstrated that normal levels of Pros expression in photoreceptor R7 cells in the *Drosophila* eye require EGFR signalling [80, 81] as well as Notch activation [32]. In addition, a recent analysis has shown that in R7 cells Notch and EGFR cooperate in a complex way to promote *pros* transcription [82]. Finally, [83] found that Pros is required for Notch expression in embryonic Lateral glial cells (LG) and that a positive-feedback loop maintains Notch and Pros expression in selected LG cells.

All these data strongly suggest that Notch and EGFR could play a central role in the mechanism by which Prospero carries out its function in the larval AMC.

In the future, it will be of great interest to explore in detail the mechanism by which all of these functions are accomplished by the homeodomain transcription factor Prospero.

Methods

Drosophila strains:

The *prosV* strains used in this study have already been described [9]. Briefly, they were generated by insertion (*prosV1*) and remobilization (*prosV13, prosV24*, *ProsV14*) of a *PGal4* transposon, located 216 pb upstream of the *pros* transcription initiation site. The resulting behavioural and developmental anomalies observed in these mutants are summarised in Table 1.

Isolation of AMC and CNS tissue

Around 150 larvae were used to obtain the AMC and CNS samples. The anterior region of the Larva was dissected to isolate CNS and AMC. The AMC region is not well-defined tissue but rather constituted by a small group of cells located in front of the hooks. Therefore, to maintain AMC integrity we kept the cuticle around it as well as the hooks.

Immunohistochemistry experiments

Staining experiments were performed as previously described by Guenin et al. [9] Briefly, isolated larval AMC from embryos at stage 10-17 were incubated with various primary antibodies: MR1A mouse anti-Prospero at 1:4 dilution, rat anti-Elav at 1/1000 (a neuronal marker; provided by A. Giangrande), and rabbit anti-phosphohistone H3 at 1/1000 (p-Histone H3) and a marker for mitotic activity; SIGMA). The following secondary antibodies were used to visualize these primary antibodies: anti-mouse Cy3 at 1/100 (Sigma); anti-mouse Alexa 594 anti-rat Alexa 488 at 1/400 (Molecular probes, USA); anti-rabbit Alexa 488 at 1/400 (Molecular probes, USA). was performed with standard methods (Sullivan et al. 2000). AMC and CNS were mounted on Vectashield (Vector Laboratories, CA) before inspection under a fluorescence microscope (Leica DMRB) or a confocal microscope (Leica 4SD).

Microarray experiments

Total RNA from third instar larvae (3 µg) was extracted from isolated AMC and CNS, according to Guenin et al. [9]. Four independent extractions were performed for each sample condition. RNA integrity was checked by denaturing formaldehyde agarose gel electrophoresis, and the quantity was assessed by optical density (OD) 260/280 ratios. cDNA were synthesised from 3.0µg total RNA in the presence of 33[P] dATP (Amersham Pharmacia Biotech, Bucks, United Kingdom). Nylon membrane microarrays provided by the TAGC platform (Marseille-Nice Genopole) were used. They contained 7500 amplification PCR products of unique full length cDNA clones from the *Drosophila* Gene collection release version 1.0 (Berkeley *Drosophila* Genome project). To verify the quality of spotting on the microarrays and the amount of DNA accessible for each spot, a first hybridization with a 33[P] labelled oligonucleotide common to all spotted PCR products (called "vector") was performed. Then, after stripping, 33[P] labelled probes made from larvae, RNA were hybridized. Radioactive labelling, hybridization, and posthybridization washes were performed according to the procedures available online at http://tagc.univ-mrs.fr/pub. The microarrays were scanned (with a Fuji BAS 5000, Raytest, Paris France) to quantify signal intensities (ArrayGauge software V1.3; Fuji, Paris, France). The values were corrected for background level, normalized for the amount of spotted cDNA and the variability of experimental conditions, then log-transformed.

Signal intensities depend on the accessible amount of PCR product spotted on the microarray. Hence, for each spot, the corresponding vector signal was used to check the reliability of the measurement, and one sample was discarded (CNS *V14)* because of bad vector signals. Then, the variability due to experimental conditions was eliminated by using a local weighted scattered plot smoother analysis (LOWESS, [84]). The data were then filtered and only values found to be twice the mean background value were kept. Data were then log transformed. Hierarchical clustering is known to highlight clusters of correlated genes participating in the same biological function [27]. For this clustering, we used the Cluster program with Pearson correlation distance and average linkage as the aggregation strategy. The results were displayed using TreeView [27]. Genes differentially expressed between the wild type and *pros* mutants were determined with a Discriminating Score DS [85]. This score measures the difference of gene expression between 2 groups of samples. If M1 represents mean expression of a given gene in wild type samples, and M2 the mean expression of the same gene in *pros* mutant samples, and SD the standard deviation of this gene in all considered samples, DS = (M1-M2)/(SD). As *V13, V14* and *V24* have a normal Pros expression pattern in the AMC, they were considered wild type. The DS between *V1* and all of the other alleles for the AMC was calculated for each gene. As *V13* and *V24* exhibited a complex pattern in the CNS, for each gene, we calculated a DS between the *V1* and *V14* allele for the CNS.

In order to find clusters of differentially expressed genes, DS were ordered according to the hierarchical clustering. The score for each gene was smoothed by calculating the mean score in a sliding window of 100 genes. Smoothed scores were plotted alongside the hierarchical clustering.

The complete dataset is available through the National Center for Biotechnology Information (NCBI), in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under the GSE12178 accession number.

Functional Annotation

Functional annotations of gene clusters were performed using GoMiner software [28] and the Gene Ontology database [86]. GoMiner determines significant enrichments of GO terms in a cluster of genes. This is performed by comparing the frequencies of each GO term in the cluster and in the microarray using Fisher's exact test.

Research of a putative common motif in the promoter of coexpressed genes

Promoting regions of coexpressed genes were collected from the Ensembl ftp site (ftp://ftp.ensembl.org/pub/current. Drosophila_melanogaster/data/fasta/cdna/). The sequence located from -1700 to +300 bp according to the +1 transcription start site of each gene was extracted. These sequences were searched for a common motif using the Gibbs sampling method [29] available at the RSA Tools website: http://rsat.ulb.ac.be/rsat/). Gibbs Sampling makes it possible to determine degenerated motifs described by a position weight matrix (PWM, a probabilistic model of residue frequencies at each position), in a set of sequences by iterative sampling. The initialisation step of the algorithm selects a random subsequence in each sequence to be searched. The predictive step builds a PWM from all the subsequences except one. The sampling step selects a new subsequence from the excluded sequence using a weighting strategy based on the PWM scores. The predictive and the sampling steps are iterated a given number of times or until convergence. Usually the

information content of a PWM is shown by a Logo representation that highlights conserved positions.

Q-PCR validation

In order to validate the Microarray results, seven genes were selected and their expression levels were quantified by Q-PCR for both *prosV1* and *prosV14*. For each RNA extraction used for microarrays, a sample was collected in order to perform Q-PCR experiments. Reverse transcriptions were done from 2µg of total RNA as described in Guenin et al. [9]. The selected genes and the corresponding primers, designed using Primer Express™ (Applied Biosystems) parameters, are indicated in Table 4. Q-PCR reactions were performed with 1 :10 diluted cDNA in 2x SybrGreen PCR master mix (Applied Biosystems) with each specific primer or control primers (*actin 5C* F 5'GCCCATCTACGAGGGTTATGC3' and *actin 5C* R 5'CAAATCGCGACCAGCCAG3'). Signals were measured with ABI Prism 7000™ Sequence Detection System software (Applied Biosystems). All signal thresholds to be compared were standardized with the *actin 5C* mRNA [9]

Authors' contributions

LG and FB generated the microarray data and drafted the manuscript. LG carry out immunohistochemistry and Q-PCR experiments. LG, MR and RH, performed the biostatistic analysis. MR performed the motif discovery. FB generated the Drosopihila larvae data base for manual annotation. FB and RH provided direction and oversight of the experiments. FB holds the grant. All authors read and approved the final manuscript.

Acknowledgements

We thank Nicolas Boulanger and Béatrice Loriod from the TAGC platform for the excellent technical assistance. This work was funded by the Burgundy Regional Council and the Centre National de la Recherche Scientifique.

References

1. Frederick RD, Denell RE: **Embryological origin of the antenno-maxillary complex of the larva of Drosophila melanogaster Meigen (Diptera: Drosophilidae).** *Int J Insect Morphol Embryol* 1982, **11**:227-233.

2. Singh RN, Singh K: **Fine structure of the sensory organs of Drosophila melanogaster Meigen larva (Diptera: Drosophilidae).** *Int J Insect Morphol Embryol* 1984, **13**:255-273.

3. Gerber B, Stocker RF: **The Drosophila larva as a model for studying chemosensation and chemosensory learning: a review**. *Chem Senses* 2007, **32**:65- 89.

4. Chu-Wang IW, Axtell RC: **Fine structure of the terminal organ of the house fly larva, Musca domestica** L. *Z Zellforsch Mikrosk Anat* 1972a, **127:**287-305.

5. Chu-Wang IW, Axtell RC: **Fine structure of the ventral organ of the house fly larva, Musca domestica** L. *Z Zellforsch Mikrosk Anat* 1972b, **130**:489-495.

6. Singh RN, Singh K: **Fine structure of the sensory organs of Drosophila melanogaster Meigen larva (Diptera: Drosophilidae).** *Int J Insect Morphol Embryol* 1984, **13**:255-273.

7. Singh RN: **Neurobiology of the gustatory systems of Drosophila and some terrestrial insects.** *Microsc Res Tech* 1997, **39**:547-563.

8. Heimbeck G, Bugnon V, Gendre N, Haberlin C, Stocker RF: **Smell and taste perception in Drosophila melanogaster larva: toxin expression studies in chemosensory neurons**. *J Neurosci* 1999, **19**:6599-6609.

9. Guenin L, Grosjean Y, Fraichard S, Acebes A, Baba-Aissa F, Ferveur JF: **Spatiotemporal expression of Prospero is finely tuned to allow the correct development and function of the nervous system in Drosophila melanogaster**. *Dev Biol* 2007, **304:** 62-74.

10. Balakireva M, Gendre N, Stocker RF, Ferveur JF: **The genetic variant Voila causes gustatory defects during Drosophila development**. *J Neurosci* 2000, **20:**3425-3433.

11. Grosjean Y, Lacaille F, Acebes A, Clemencet J, Ferveur JF: **Taste, movement, and death: varying effects of new prospero mutants during Drosophila development**. *J Neurobiol* 2003, 55:1-13.

12. Chu-Lagraff Q, Wright DM, McNeil LK, Doe CQ: **The prospero gene encodes a divergent homeodomain protein that controls neuronal identity in Drosophila.** *Development* 1991, Suppl **2**:79-85.

13. Matsuzaki F, Koizumi K, Hama C, Yoshioka T, Nabeshima Y: **Cloning of the Drosophila prospero gene and its expression in ganglion mother cells**. *Biochem Biophys Res Commun* 1992, **182:**1326-1332.

14. Hassan B, Li L, Bremer KA, Chang W, Pinsonneault J, Vaessin H: **Prospero is a panneural transcription factor that modulates**

homeodomain protein activity. *Proc Natl Acad Sci U S A* 1997, **94:**10991-10996.
15. Doe CQ, Chu-LaGraff Q, Wright DM, Scott MP: **The prospero gene specifies cell fates in the Drosophila central nervous system.** *Cell* 1991, **65:**451-464.
16. Vaessin H, Grell E, Wolff E, Bier E, Jan LY, Jan YN: **prospero is expressed in neuronal precursors and encodes a nuclear protein that is involved in the control of axonal outgrowth in Drosophila**. *cell* 1991, **67**:941-953.
17. Hirata J, Nakagoshi H, Nabeshima Y, Matsuzaki F: **Asymmetric segregation of the homeodomain protein Prospero during Drosophila development**. *Nature* 1995, **377**:627-630.
18. Knoblich JA, Jan LY, Jan YN: **Asymmetric segregation of Numb and Prospero during cell division.** *Nature* 1995, **377**:324-327.
19. Spana EP, Doe CQ: **The prospero transcription factor is asymmetrically localized to the cell cortex during neuroblast mitosis in Drosophila**. *Development* 1995**, 121:** 3187-3195.
20. Manning L, Doe CQ: **Prospero distinguishes sibling cell fate without asymmetric localization in the Drosophila adult external sense organ lineage**. *Development* 1999, **126**:2063-2071.
21. Akiyama-Oda Y, Hosoya T, Hotta Y: **Asymmetric cell division of thoracic neuroblast 6-4 to bifurcate glial and neuronal lineage in Drosophila**. *Development* 1999, **126**:1967-1974.
22. Freeman MR, Doe CQ: **Asymmetric Prospero localization is required to generate mixed neuronal/glial lineages in the Drosophila CNS**. *Development* 2001, **128**:4103-4112.
23. Li L, Vaessin H: **Pan-neural Prospero terminates cell proliferation during Drosophila neurogenesis**. *Genes Dev* 2000, **14**:147-151.
24. Griffiths RL, Hidalgo **A: Prospero maintains the mitotic potential of glial precursors enabling them to respond to neurons**. *EMBO J* 2004, **23**:2440-2450.
25. Liu TH, Li L, Vaessin H: **Transcription of the Drosophila CKI gene dacapo is regulated by a modular array of cis-regulatory sequences**. *Mech Dev* 2002, **112**:25-36.
26. Choksi S, P., Southall T, D., Bossing T, Edoff K, de Wit E, Fischer B, E., van Steensel B, Micklem G, Brand A, H.: **Prospero acts as a binary switch between self-renewal and differentiation in Drosophila neural stem cells**. *Dev Cell* 2006**, 11**:775-789.
27. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**:14863-14868.
28. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S *et al*: GoMiner: **A resource for biological interpretation of genomic and proteomic data**. *Genome Biol* 2003, **4**:R28.
29. Lawrence CE, S.F. A, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment**. *Science* 1993, **262**:208-214.
30. Crooks GE, Hon G, Chandonia JM, Brenner SE: WebLogo: **a sequence logo generator.** *Genome Res* 2004, **14:** 1188-1190.
31. Hassan B, Li L, Bremer KA, Chang W, Pinsonneault J, Vaessin H: **Prospero is a panneural transcription factor that modulates homeodomain protein activity**. *Proc Natl Acad Sci U S A* 1997**, 94**:10991-10996.
32. Cook T, Pichaud F, Sonneville R, Papatsenko D, Desplan C: **Distinction between color photoreceptor cell fates is controlled by Prospero in Drosophila**. *Dev Cell* 2003, **4:**853-864.
33. Hassan BA, Bermingham NA, He Y, Sun Y, Jan YN, Zoghbi HY, Bellen HJ: **atonal regulates neurite arborization but does not act as a proneural gene in the Drosophila brain.** *Neuron* 2000, **25:** 549-561.
34. Srahna M, Leyssen M, Choi CM, Fradkin LG, Noordermeer JN, Hassan BA: **A signaling network for patterning of neuronal connectivity in the Drosophila brain**. *PLoS Biol* 2006, **4:**e348.
35. Kraut R, Menon K, Zinn K: **A gain-of-function screen for genes controlling motor axon guidance and synaptogenesis in Drosophila**. *Curr Biol* 2001, **11:**417-430.
36. Ng J, Luo L: **Rho GTPases regulate axon growth through convergent and divergent signaling pathways.** *Neuron* 2004, **44:**779-793.
37. Marek KW, Ng N, Fetter R, Smolik S, Goodman CS, Davis GW**: A genetic analysis of synaptic development: pre- and postsynaptic dCBP control transmitter release at the Drosophila NMJ**. *Neuron* 2000, **25**:537-547.
38. Laviolette MJ, Nunes P, Peyre JB, Aigaki T, Stewart BA: **A genetic screen for suppressors of Drosophila NSF2 neuromuscular junction overgrowth**. *Genetics* 2005, **170**:779-792.
39. Garcia-Alonso LA: **Postembryonic sensory axon guidance in Drosophila.** *Cell Mol Life Sci* 1999, **55** :1386-1398.
40. Learte AR, Forero MG, Hidalgo A: **Gliatrophic and gliatropic roles of PVF/PVR signaling during axon guidance**. *Glia* 2008, **56** :164-176.
41. Marin EC, Watts RJ, Tanaka NK, Ito K, Luo L: **Developmentally programmed remodeling of**

the Drosophila olfactory circuit. *Dev* 2005, **132**:725-737.
42. Speese SD, Trotta N, Rodesch CK, Aravamudan B, Broadie K: **The ubiquitin proteasome system acutely regulates presynaptic protein turnover and synaptic efficacy.** *Curr Biol* 2003, **13**:899-910.
43. Haas KF, Miller SL, Friedman DB, Broadie K: **The ubiquitin-proteasome system postsynaptically regulates glutamatergic synaptic function**. *Mol Cell Neurosci* 2007, **35:** 64-75.
44. Juhasz G, Puskas LG, Komonyi O, Erdi B, Maroy P, Neufeld TP, Sass M: **Gene expression profiling identifies FKBP39 as an inhibitor of autophagy in larval Drosophila fat body**. *Cell Death Differ* 2007, **14:**1181-1190.
45. Martin DN, Baehrecke EH: **Caspases function in autophagic programmed cell death in Drosophila.** *Development* 2003, **131:**275-284.
46. Martin DN, Balgley B, Dutta S, Chen J, Rudnick P, Cranford J, Kantartzis S, DeVoe DL, Lee C, Baehrecke EH: **Proteomic analysis of steroid-triggered autophagic programmed cell death during Drosophila development**. *Cell Death Differ* 2007, **14:** 916-923.
47. Reiling JH, Doepfner KT, Hafen E, Stocker H: **Diet-dependent effects of the Drosophila Mnk1/Mnk2 homolog Lk6 on growth via eIF4E**. *Curr Biol* 2005, **15:** 24-30.
48. Arquier N, Bourouis M, Colombani J, Léopold P: **Drosophila Lk6 kinase controls phosphorylation of eukaryotic translation initiation factor 4E and promotes normal growth and development**. *Curr Biol* 2005, **15:**19-23.
49. Yamada M, Murata T, Hirose S, Lavorgna G, Suzuki E, Ueda H: Temporally restricted expression of transcription factor betaFTZ-F1: **significance for embryogenesis, molting and metamorphosis in Drosophila melanogaster**. *Development* 2000, **127**:5083-5092.
50. Parker J: **Control of compartment size by an EGF ligand from neighboring cells**. *Curr Biol* 2006, **16**:2058-2065.
51. Rafel N, Milán M: **Notch signalling coordinates tissue growth and wing fate specification in Drosophila**. *Development* 2008, **135**:3995-4001.
52. Yue X, Song W, Zhang W, Chen L, Xi Z, Xin Z, Jiang X: **Mitochondrially localized EGFR is subjected to autophagic regulation and implicated in cell survival**.. *Autophagy* 2008**,** **4**:701-703.
53. Adamson AL, Shearn A: **Molecular genetic analysis of Drosophila ash2, a member of the trithorax group required for imaginal disc pattern formation**. *Genetics* 1996, 1**44**:621-633.
54. Bose A, Kahali B, Zhang S, Lin JM, Allada R, Karandikar U, Bidwai AP: **Drosophila CK2 regulates lateral-inhibition during eye and bristle development.** *Mech Dev* 2006, **123**:649-664.
55. Kanuka H, Kuranaga E, Takemoto K, Hiratou T, Okano H, Miura M: **Drosophila caspase transduces Shaggy/GSK-3beta kinase activity in neural precursor development.** *EMBO J* 2005, **24**:3793-3806.
56. Ang LH, Kim J, Stepensky V, Hing H: **Dock and Pak regulate olfactory axon pathfinding in Drosophila**. *Development* 2003, **130:**1307-1316.
57. Mollereau B, Domingos PM: **Photoreceptor differentiation in Drosophila: from immature neurons to functional photoreceptors.** *Dev Dyn* 2005, **232**:585-592.
58. Fichelson P, Audibert A, Simon F, Gho M: **Cell cycle and cell-fate determination in Drosophila neural cell lineages**. *Trends Genet* 2005, **21:**413-420.
59. Lai EC, Orgogozo V: **A hidden program in Drosophila peripheral neurogenesis revealed: fundamental principles underlying sensory organ diversity**. *Dev Biol* 2004, **269**:1-17.
60. Gao FB, Brenman JE, Jan LY, Jan YN: **Genes regulating dendritic outgrowth, branching, and routing in Drosophila**. *Genes Dev* 1999, **13:**2549-2561.
61. Morey M, Yee SK, Herman T, Nern A, Blanco E, Zipursky S: **Coordinate control of synaptic-layer specificity and rhodopsins in photoreceptor neurons**. *Nature* 2008, **456**: 795-799.
62. Tissot M, Stocker RF: **Metamorphosis in drosophila and other insects: the fate of neurons throughout the stages**. *Prog Neurobiol* 2000, **62:**89-111.
63. Chen GC, Gajowniczek P, Settleman J: **Rho-LIM kinase signaling regulates ecdysone-induced gene expression and morphogenesis during Drosophila metamorphosis**. *Curr Biol* 2004, **14:**309-313.
64. Li TR, White KP: **Tissue-specific gene expression and ecdysone-regulated genomic networks in Drosophila**. *Dev Cell* 2003, **5:**59-72.
65. Endo K, Aoki T, Yoda Y, Kimura K, Hama C: **Notch signal organizes the Drosophila olfactory circuitry by diversifying the sensory neuronal lineages**. *Nat Neurosci* 2007, **10**:153-160.
66. Zhao T, Gu T, Rice HC, McAdams KL, Roark KM, Lawson K, Gauthier SA, Reagan KL, Hewes RS: A **Drosophila gain-of-function screen for candidate genes involved in steroid-dependent**

neuroendocrine cell remodeling. *Genetics* 2008, **178**:883-901.
67. Grewal SS: **Insulin/TOR signaling in growth and homeostasis: A view from the fly world.** *Int J Biochem Cell Biol* 2009, **41**:1006-1010.
68. Jünger MA, Rintelen F, Stocker H, Wasserman JD, Végh M, Radimerski T, Greenberg ME, Hafen E: **The Drosophila forkhead transcription factor FOXO mediates the reduction in cell number associated with reduced insulin signaling**. *J Biol* 2003, **2**:20.
69. Scott RC, Schuldiner O, Neufeld TP: **Role and regulation of starvation-induced autophagy in the Drosophila fat body.** *Dev Cell* 2004, **7**:167-178.
70. Jacinto E, Hall MN**: Tor signalling in bugs, brain and brawn**. *Nat Rev Mol Cell Biol* 2003, **4**:117-126.
71. Teleman AA, Hietakangas V, Sayadian AC, Cohen SM: **Nutritional control of protein biosynthetic capacity by insulin via Myc in Drosophila**. *Cell Metab* 2008, **7**:21-32.
72. Colombani J, Bianchini L, Layalle S, Pondeville E, Dauphin-Villeman C, Antoniewski C, Carré C, Noselli S, Léopold P: **Antagonistic actions of ecdysone and insulins determine final size in Drosophila.** *Sciences* 2005, **310:**667-670.
73. Guertin DA, Guntur KV, Bell GW, Thoreen CC, Sabatini DM: **Functional genomics identifies TOR-regulated genes that control growth and division**. *Curr Biol* 2006, 16(10):958-970.
74. Yan N, Macdonald PM: **Genetic interactions of Drosophila melanogaster arrest reveal roles for translational repressor Bruno in accumulation of Gurken and activity of Delta**. *Genetics* 2004, **168**:1433-1442.
75. McNeill H, Craig GM, Bateman JM: **Regulation of neurogenesis and epidermal growth factor receptor signaling by the insulin receptor/target of rapamycin pathway in Drosophila.** *genetics* 2008, **179**:843-853.
76. Sen A, Shetty C, Jhaveri D, Rodrigues V: **Distinct types of glial cells populate the Drosophila antenna**. *BMC Dev Biol* 2005, 5:25.
77. Okano H, Hayashi S, Tanimura T, Sawamoto K, Yoshikawa S, Watanabe J, Iwasak iM, Hirose S, Mikoshiba K, Montell C: **Regulation of Drosophila neural development by a putative secreted protein.** *Differentiation* 1992, **52**:1-11.
78. Mayer U, Nusslein-Volhard C: **A group of genes required for pattern formation in the ventral ectoderm of the Drosophila embryo**. *Genes Dev* 1988, **2**:1496-1511.
79. Sun H, Oakley B: **Development of anterior gustatory epithelia in the palate and tongue requires epidermal growth factor receptor**. *Dev Biol* 2002, **242**:31-43.
80. Kauffmann RC, Li S, Gallagher PA, Zhang J, Carthew RW: **Ras1 signaling and transcriptional competence in the R7 cell of Drosophila**. *Genes Dev* 1996, **10**:2167-2178.
81. Xu J, Xin S, Du W: **Drosophila Chk2 is required for DNA damage-mediated cell cycle arrest and apoptosis.** *FEBS lett* 2001, **508**:394-398.
82. Hayashi T, Xu C, Carthew RW: **Coordinate control of synaptic-layer specificity and rhodopsins in photoreceptor neurons.** *Nature* 2008, **456**:795-799.
83. Griffiths RC, Benito-Sipos J, Fenton JC, Torroja L, Hidalgo A: **Two distinct mechanisms segregate Prospero in the longitudinal glia underlying the timing of interactions with axons.** *Neuron Glia Biol* 2007, **3**:75-88.
84. Dudoit S, Fridlyand J: **A prediction-based resampling method for estimating the number of clusters in a dataset.** *genome Biol* 2002, **3:**36.31-21.
85. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**:531-537.
86. Ashburner M, Ball CA, Blake JA, Botstein D, Butle rH, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25:**25-29.
87. Yousef MS, Matthews BW: **Structural basis of Prospero-DNA interaction: implications for transcription regulation in developing cells**. *Structure* 2005, **13:**601-607.

Figures

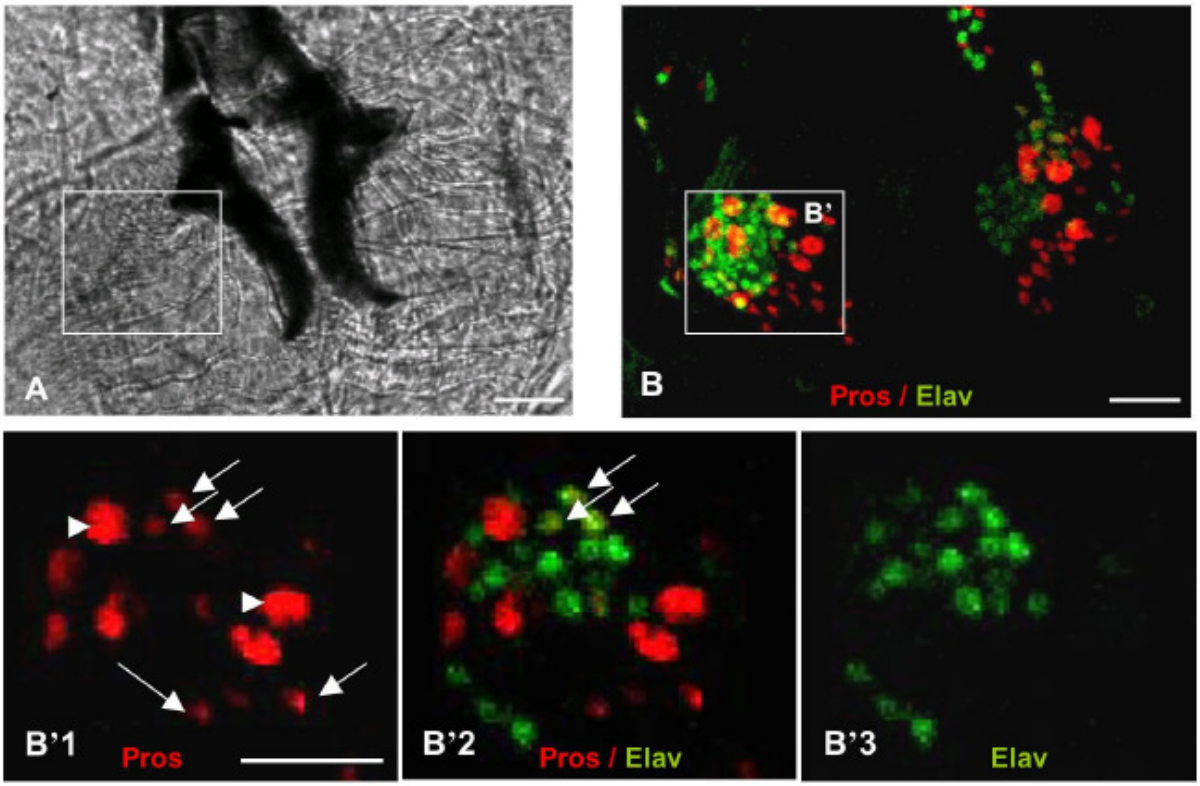


Fig 1: AMC region from third instar larvae observed by optical microscopy
(A) Bright-field view of the larval AMC region (dorsal view, anterior down), the hooks appear in dark. Cells that constitute the AMC are located on either sides of the hooks. (B) 3D reconstruction of AMC (TO +DO), labelled with Pros (red) and Elav (green) that labels neuronal cells. (B'1-3) Zoomed view of a confocal section of the framed region in B showing respectively the Pros (B'1), Pros/Elav (B'2) and Elav (B'3) staining: Anti-Prospero antiboby labels two types of Pros expressing cells (Pros+): large (arrowheads in B'1) and small cells (arrows in B'1). Some of the small Pros+ cells express Elav (B'2). Scale bars represent 10µm.

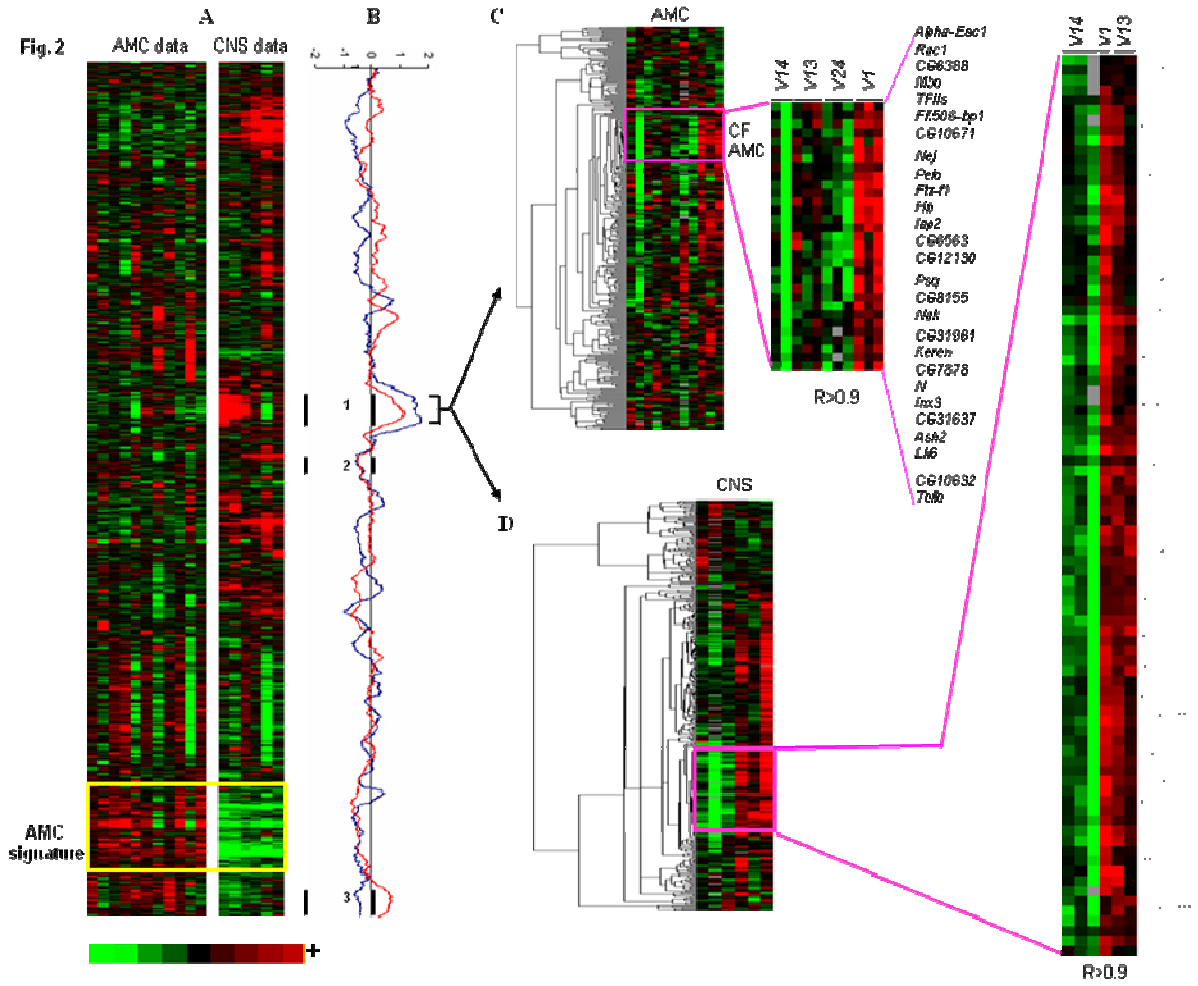


Fig. 2 Gene expression analysis

A Hierarchical clustering of 5950 genes for a total of 17 samples relative to the larval AMC and CNS of the different *prosV* mutants. Each row represents a gene and each column a sample. For each organ, the samples at the top of the image are classified according to the severity of their phenotypes (from wild type to the most severe phenotype: *V14, V13, V24* and *V1*). Each cell in the matrix corresponds to the expression level of one gene in a sample, red indicates a high level of mRNA expression compared to the median value, green a low level, and black for a level close to the median (see color scale at the bottom of the image). The Yellow frames represent the AMC tissue specific signature and therefore contain the genes that are differentially expressed between AMC and the CNS independently of Pros expression. The genes of this cluster are clearly overexpressed in the AMC (in red for all samples including *V1*) while the same genes are underexpressed for all alleles in the CNS (in green). **B** Discriminating scores (DS) smoothed in a window of 100 genes, calculated between *V1* and other *prosV* in the AMC (in red), and between *V1* and *V14* in the CNS (in blue), among the gene clusters. In the AMC, three peaks, annotated 1, 2 and 3 (black bars), appear to be enriched in differentially expressed genes, they have been associated respectively to "cell fate commitment", "proteasome complex" and "signal transduction" ontologies . **C** Hierarchical clustering of the 306 genes present in peak 1 in the AMC. Pink frame zooms on a set of highly correlated (r>0.9) genes that are differentially expressed between *V1* and all other alleles in the AMC. These genes referenced on the right according to the *Drosophila* nomenclature (see also Table 3). The dendrogram on the left represents correlation distances between the profiles of the studied genes. **D** Same as **C** in CNS samples. The Pink framed region contains 86 genes which are referenced on the right according to the *Drosophila* nomenclature.

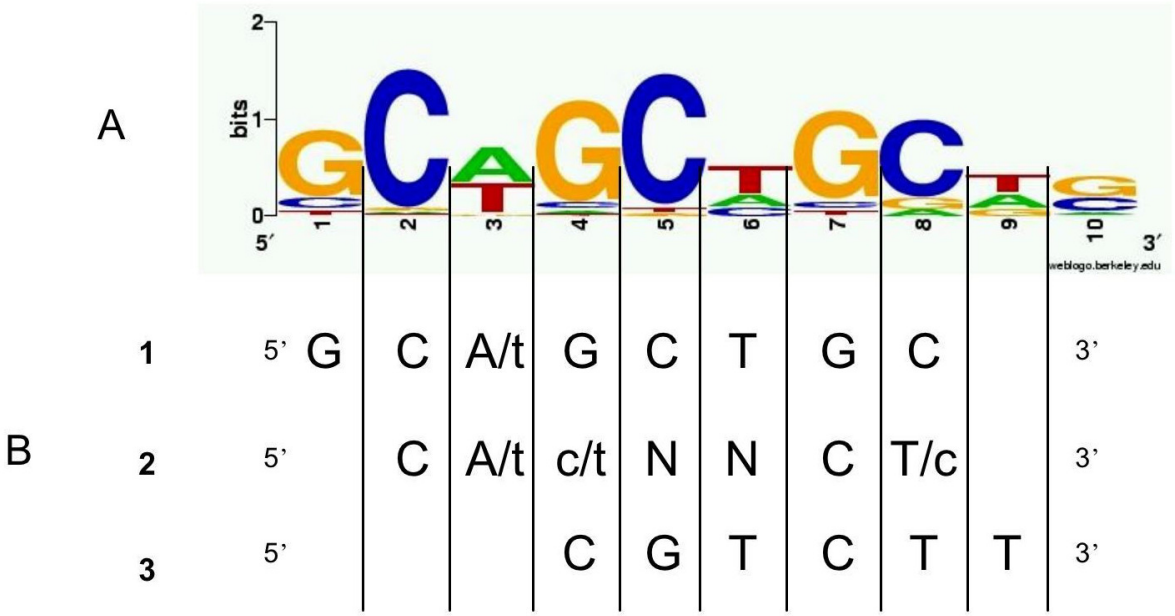


Fig.3: Putative binding site for Prospero

(A) Palindromic sequence present in the promoting region of 28 genes shared by the AMC and CNS clusters. (B) Aligned DNA sequences proposed for the Homeo-Prospero domain targeting: 1/ in this study. 2/ by Hassan et al. [31] 3/ by Cook et al. [32] and redrawn from Yousef and Matthews [87]. As it can be seen our putatif motif fits only partly that of Hassan et al. [31] and the motif proposed by Cook et al. [32].

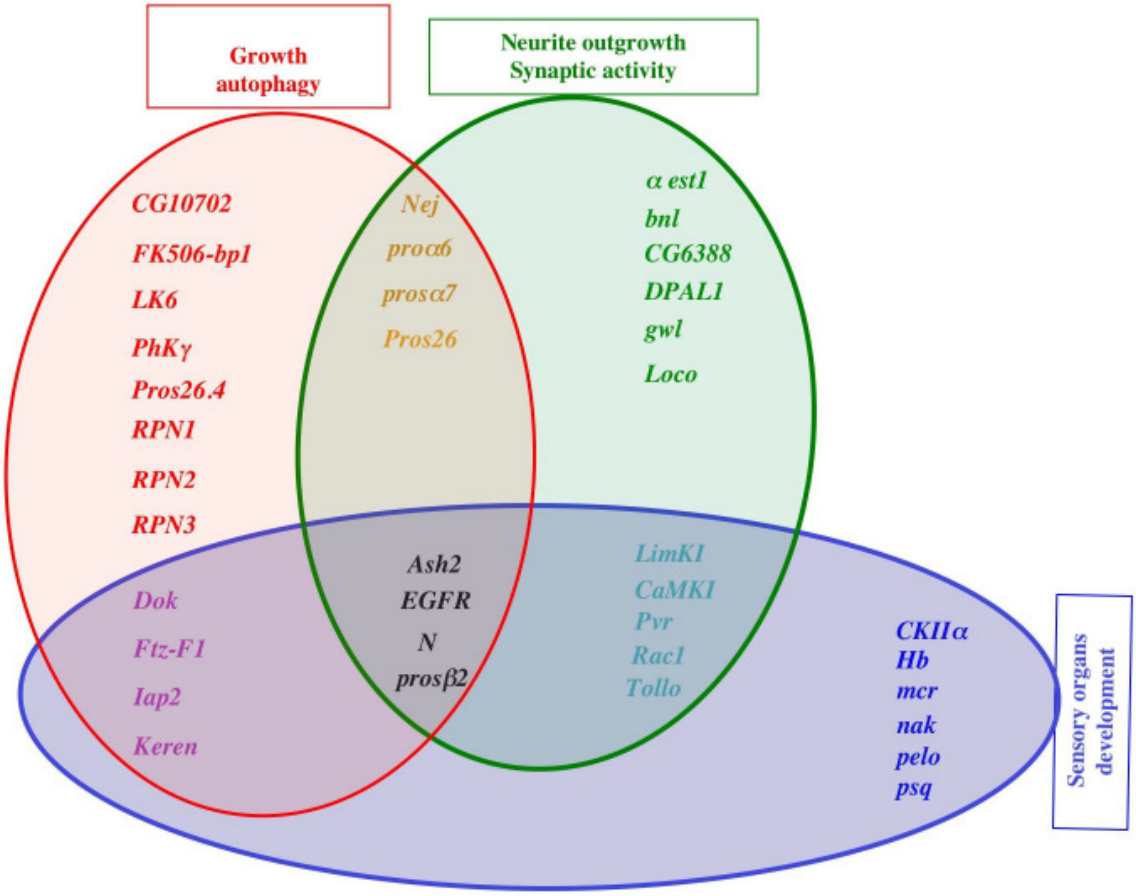


Fig. 4: Schematic representation of the overlapping function attributed to the AMC putative Pros target genes.

The functional categories were established using a manual annotation (the criteria used for this annotation are indicated in the text, see also for further phenotypic description and corresponding references Table S4, S5 and S6 in supplemental data). The three functional groups identified are represented by three distinct colored sets. The genes located at the intersection between two sets can assume both functions. It should be noted that the genes indicated in black (*EGFR, Notch, Ash2 and pros*β2) belong to the three functional groups: neurite outgrowth, sensory organ development, and growth/autophagy.

Tables

Table 1: Overview of the phenotypes associated with the different *prosV* alleles.
(redrawn from Guenin et al., 2007).
In the *prosV1* (*V1*) allele, the full length *PGal4* transposon is inserted upstream of the *pros* coding region, while in *prosV14* (*V14*), the transposon has been removed restoring the wild type phenotype; *prosV24*(*V24*) and *prosV13* (*V13*) are excision alleles and contain variable sizes of the *PGal4* transposon; The peak of developmental lethality, taste response of late homozygous 2nd instar larva, Pros expression level in larvae and axonal misrouting are indicated for each *pros* allele. The larval taste response was measured towards 0.1 *M* sucrose and 0.3 *M* NaCl concentration that are known to respectively attract or repulse wild type *Drosophila*. *V1* mutants were indifferent to both substances (altered taste response), *V24* showed an intermediate response: they were repulsed by NaCl but remained indifferent to sucrose and *V13* and *V14* present a normal taste response to both substances. The Pros expression pattern is indicated by comparison to the *V14* wild type: In the AMC, *V1* showed no Pros expression but for the other alleles, Pros expression pattern was similar to *V14*. In the CNS, all mutants alleles showed a distinct altered expression pattern as compared to the wild type (further descriptions of the Pros pattern are found in the text).

	Associated phenotypes				
	Stage of lethality	**Larval taste response**	**Pros expression in**		**Axonal routing in AMC**
			AMC	**CNS**	
V14	Viable	Normal	Normal	Normal	Normal
V13	Young adult < 2 days old	Normal	Normal	altered	Normal
V24	pupal	Intermediate	Normal	altered	Normal
V1	larva	Altered	absent	altered	Misrouting

Table 2: Pros expressing cells in the larval AMC.
We have quantified the number of Pros protein expressing cells (Pros+) and neuronal cells (Elav+) in the third instar larval AMC of wild type (*V14*) and *V1* mutants. We distinguish two types of Pros+ cells on the basis of their size: large and small cells. Some small Pros+ cells express Elav markers (under square brackets) and are probably differentiated neurons. In *V1* mutants, no more Pros protein is detected in the larval AMC, but the number of neurons remains unchanged.

Cell types	**Large cells**	**Small cells**	
alleles	**Pros+**	**Pros+**	**Elav+**
V14	8±1	40.9±4.1 [10.7±2.8]	65.8±1.4
V1	**0** ***	**0** *** [**0** ***]	62.3±0.9

Table 3: List of the 64 genes identified as putative pros targets in this study and their corresponding manual annotation

The 64 genes found highly correlated in the peak 1, 2 and 3 are grouped according to their respective GO annotation class (column on the left). The most significant classes of genes enriched in our list are "Cell fate commitment AMC", "proteasome complex" and "signal transduction". The p value indicates the probability for a given onthology to be associated at random to this cluster. The first 28 genes indicated in bold and by an asterisk share a common DNA motif (**CAGCTG**) in their promoter and were also found to be differentially expressed between *V1* and *V14 CNS*. The last column on the right specifies the known biological function (manual annotation) of these genes in the *Drosophila* larvae (see also supplemental data table S4-S6). This manual annotation allowed the attribution of biological function to 37 genes. The genes indicated in grey have either never been studied in larvae or their respective functions are currently unknown. By contrast to the GO annotation (mostly deduced from embryos), the use of manual annotation indicates that the dysfunction of *pros* leads in larvae to the missregulation of genes that mostly deal with neurite outgrowth (indicated in green), growth and autophagy (in red) and sensory organ formation (mainly olfactory, indicated in blue). All genes were found to be overexpressed in the mutant V1 AMC except for the 9 genes associated with the Proteasome complex GO annotation

GO annotation	Genes	symbol	Biological function in larvae (manual annotation)
Cell fate commitment (GO:0045165, p= 0.0003)	*Est1**	CG1031	Sensory neuron morphogenesis
	Art3*	CG6563	Not studied in larvae
	Ash2*	CG6677	Neurite outgrowth, synapse formation, growth, sensory organ development
	CG10632*	CG10632	Unknown
	CG10671*	CG10671	Unknown
	CG3021*	CG3021	Unknown
	CG31637*	CG31637	Unknown
	CG31961*	CG31961	Unknown
	CG31731*	CG31731	Unknown
	CG6388	CG6388	Neurite outgrowth
	CG7878*	CG7878	Unknown
	CG8155*	CG8155	Unknown
	DPAL1*	CG12130	Neuropeptide biosynthesis
	FK506-bp1*	CG6226	Autophagy ; growth
	Ftz-F1*	CG4059	Autophagy, sensory organ formation, olfaction
	Hb*	CG9786	labial segment formation including sense organ
	Iap2*	CG8293	Autophagy, sensory organ development
	Inx3*	CG1448	Not studied in larvae
	Keren*	CG32179	Autophagy, sensory organ development
	Mbo*	CG6819	Tracheal system development
	Nak*	CG10637	Not studied in larvae
	Nej *	CG15319	Synaptic transmission, autophagy
	Notch*	CG3936	Neurite outgrowth, nutrient sensing/growth, sense organ formation, olfaction
	Pelo*	CG3959	Not studied in larvae
	Psq*	CG2368	Sensory organ development, olfaction
	Rac1*	CG2248	Neurite outgrowth, sensory organ development
	Tollo*	CG6890	Synaptogenesis, wing development, immune response
	TFIIs*	CG3710	Not studied in larvae
	*LK6**	CG17342	Autophagy growth, nutrient sensor mechanism,
Proteasome complex (GO :0000502, $p<10^{-5}$,)	*Pros 26,4*	CG5289	Neuronal remodelling, Autophagy
	Pros 2	CG3329	Neuronal remodelling, Synaptic transmission, autophagy, sensory organ formation
	Pros 26	CG4097	Neuronal remodelling, Synaptic transmission, autophagy
	Pros 6	CG18495	Neuronal remodelling, Synaptic transmission, autophagy
	Pros 7	CG1519	Neuronal remodelling, Synaptic transmission, autophagy
	ProsMA5	CG10938	Not studied in larvae
	RPN1	CG7762	Neuronal remodelling, Autophagy
	RPN2	CG11888	Neuronal remodelling, Autophagy

	RPN5	CG1100	Neuronal remodelling, Autophagy
Signal transduction (GO :0004871, p=0.0008)	*Bnl*	CG4608	neurite outgrowth
	CaMKI	CG1495	synaptic transmission
	CG10011	CG10011	Unknown
	CG10702	CG10702	autophagy
	CG1088	CG10882	Unknown
	CG31714	CG31714	Unknown
	CG4839	CG4839	Unknown
	CG5790	CG5790	Unknown
	CG7536	CG7536	Unknown
	CG7800	CG7800	Unknown
	CKII □	CG17520	Sensory organ development
	Dok	CG2079	Sensory organ development
	EGFR	CG10079	Neurite outgrowth, synapse formation, growth, autophagy sensory organ development, olfaction
	feo	CG11207	Mitotic spindle organisation
	Gek	CG4012	Actin polymerisation
	Gwl	CG7719	Neurite outgrowth, synaptic transmission, mitotic cell cycle
	InaC	CG6518	Not studied in larvae
	Kdelr		Not studied in larvae
	LimK1	CG1848	Neurite outgrowth, synaptic transmission
	Lok	CG10895	Cell cycle, DNA damage checkpoint
	Loco	CG5248	Not studied in larvae
	Mcr	CG7586	olfaction
	PhK□	CG1830	Not studied in larvae
	Pvr	CG8222	Hemocyte formation, dorsal closure, macrochaete formation
	Rh7	CG5638	Not studied in larvae
	Toll-6	CG7250	Not studied in larvae

Table 4: Validation of microarray data using real time PCR, comparison of results obtained with both methods.

The relative expression level (*V1 /V14*) of selected genes was measured using the Q- PCR or microarray analysis data; Our results were consistent with the microarray data except for the *hb (hunchback)* gene found to be overexpressed in the CNS but not in the AMC. The values in gray correspond to the genes found differentially expressed between V1 and V14 in the CNS but not in the AMC using microarray analysis. Accordingly, no significant variation was found for these genes in the AMC, using Q-PCR.

	Relative expression level *V1/V14*			
	AMC		CNS	
Gene Primers	Microarray	Q-PCR	Microarray	Q-PCR
caps F 5'GCAGCCTGGATGAAGGTTTA 3' R 5'ATGGCGCAGCCATAGTAGTC 3'	1.38	0.63	**3.8**	**2.38**
Cdk4 F 5' TACAACAGCACCGTGGACAT 3' R 5' GGTCCAGCTGATTCTTTTCG 3'	0.95	1.3	**4.99**	**2.5**
hb F 5' CCTTCCAGTGCGACAAATG 3' R 5' ATCCGCACAACGGTACTGA 3'	**6.71**	**0.85**	**6.38**	**1.6**
Iap2 F 5'AAGGACTGGCCGAATCCCAACATC 3' R 5' CGTTGCACCAAACACACTTC 3'	**3.69**	**2.16**	**6.48**	**1.9**
nak F 5'AGGAAGCATCACAGCAAAAT 3' R 5'GCACCAGGAGCAGCTGTAAC 3'	**1.75**	**1.36**	**0.97**	**1.95**
nej F 5'AATGGATCCAACGGATATCTCT 3' R 5'CTGATCCGACCAGCCACTAT 3'	**3.26**	**1.63**	**3.79**	**3.75**
Notch F 5'AACACCGTTCGCGGAACTGATACCG 3' R 5'GGTTTTGCCATTGAGTTGTG 3'	**2.9**	**1.76**	**8.96**	**2.52**

4. TRANSCRIPTOME + CGH

Histologic progression of marginal zone lymphoma:
A proliferation signature induced by multiple secondary genetic events

Lise Willems* and Delphine Roland* and Mahatsangy Raharijaona*, Josette Briere, Benoit Ballester, Samuel Quentin, Roxane Legaie, Pauline Brice, Eric De Kerviler, Christian Gisselbrecht, Françoise Berger, Pascale Felman, Thierry Molina, Bertrand Coiffier, Jean Soulier, Rémi Houlgatte, Catherine Thieblemont

Les lymphomes de la zone marginale (MZL) représentent entre 7% et 15% de tous les lymphomes. Ils regroupent trois sous-type de lymphomes se différenciant par le site d'envahissement : les lymphomes extraganglionnaires développés à partir du tissu lymphoïde associé aux muqueuses (MALT), les lymphomes non-MALT : les lymphomes spléniques et les lymphomes ganglionnaires. Ces lymphomes de la zone marginale sont en général indolents, mais dans 10%-20% des cas, ils sont associés à une progresion histologique (Histologic Progression, HP), évoluant alors vers un lymphome plus agressif (HP-MZL). Aucun critère n'a été formellement adopté pour classer ces HP-MZL.

L'objectif de cette étude était de déterminer les altérations génétiques impliquées dans la progression des MZL non-MALT. Pour cela, nous avons intégré une approche transcriptome avec une analyse de changements de nombre de copies (CNC). Nous avons inclus dans l'étude 66 patients non-traités. Selon des critères morphologiques, 52 patients avaient un MZL non-MALT, et 14 avaient un MZL transformé, présentant une forme disséminée de la maladie. L'analyse transcriptome a été réalisée avec 32 MZL spléniques (5 réplicats, 3 triplicats) et 8 HP-MZL sur des puces nylon contenant 9216 gènes. D'autres lymphomes transformés (lymphomes folliculaires FL, un des lymphomes transformés les mieux décrits, et lymphomes lymphocytiques SLL) ont été ajoutés comme contrôle dans notre étude afin de visualiser des signatures déjà décrites, correspondant à des fonctions précises ou spécifiques d'un tissu ou un sous-type cellulaire. L'analyse des CNC a été réalisée avec 20 MZL spléniques et 6 HP-MZL sur des puces Agilent 105K. Je ne présenterai que les grandes lignes de ce travail, cette étude étant toujours en cours d'analyse.

Dans cette étude j'ai participé à l'analyse transcriptome (identification des clusters discriminant les MZL transformés, des clusters de survie, annotation fonctionnelle). Je n'étais pas impliqué dans la partie analyse CGH.
Nous avons montré que la progression est associée à une signature de prolifération cellulaire commune avec la transformation des lymphomes folliculaires. D'autres signatures spécifiques de la transformation des MZL ont été également identifiées. Elles incluent des oncogènes, des gènes impliqués dans la machinerie cellulaire et la motilité cellulaire, et dans la transduction du signal. Les altérations chromosomiques les plus fréquentes ont été des gains de 3/3q et 18q et des délétions dans 7q. Elles ont cependant été identifiées dans les deux groupes MZL et HP-MZL. Des CNC non-récurrentes ont pu être détectés dans les HP-MZL, soit seules, soit en combinaison complexe dans les régions 4p, 6p et 8q englobant plusieurs gènes. Parmi eux, les oncogènes IRF4 et Myc ont été identifiés.

L'intégration de données génétiques et de transcriptome ont montré que la progression histologique des MZL est liée à l'altération de la prolifération cellulaire. Malgré les altérations chromosomiques typiques des MZL, des CNC secondaires ont pu être détectés dans les HP-MZL. Ces CNC touchent des gènes contrôlant la prolifération cellulaire. Ces résultats mettent en évidence que la dérégulation de différentes voies conduit les cellules MZL vers un plus fort taux de prolifération, caractéristique de la progression histologique.

Histologic progression of marginal zone lymphoma:
A proliferation signature induced by multiple secondary genetic events

Short title for running head (n=50): Integrated genomic analyses in HP of non-MALT MZL

Authors
Lise Willems* and Delphine Roland* and Mahatsangy Raharijaona*, Josette Briere, Benoit Ballester, Samuel Quentin, Roxane Legaie, Pauline Brice, Eric De Kerviler, Christian Gisselbrecht, Françoise Berger, Pascale Felman, Thierry Molina, Bertrand Coiffier, Jean Soulier, Rémi Houlgatte, Catherine Thieblemont

From :

1. Hôpital Saint-Louis (SLS) APHP, INSERM U 728 - Institut Universitaire d'Hématologie - Service d'hémato-oncologie, Service d'anatomie pathologique, Service de Radiologie - Paris, France
2. INSERM U836 Equipe 7 - Faculte de Medecine, Universite J. Fourier, Grenoble, France
3. INSERM U915 - Institut du thorax, Faculté de Médecine, Université de Nantes, Nantes, France
4. England
5. Genome Rearrangements and Cancer Group, INSERM U728 - Institut Universitaire d'Hématologie, Paris 7 University, Hôpital Saint-Louis (SLS) APHP, Paris, France
6. Hospices Civils de Lyon, Centre Hospitalier Lyon Sud (CHLS), Service d'hématologie clinique, Service d'anatomie pathologique, Pierre Bénite, F-69495 France ; Université Lyon 1, Equipe d'Accueil 3737, Faculté Lyon-Sud, Oullins, F-69600, France.
7. Service d'anatomie pathologie – Hôtel-Dieu (HD) Université Paris V, Paris, France

*LW, DR and MH have contributed equally to this work as first co-authors

Correspondance
Catherine Thieblemont
Service d'hémato-oncologie
Hôpital Saint-Louis (SLS) – 1, Avenue C. Vellefaux, Paris, France
Tel 33 (0)1 42 49 98 37
e-mail : catherine.thieblemont@sls.aphp.fr

Financial support
CT was supported by the Ligue Contre le Cancer (Comités du Rhône, de la Saône et Loire, de la Drôme et de l'Ardèche. BB was supported by a Ph.D fellowship from the Association pour la Recherche contre le Cancer (ARC).

Abstract

Patients with marginal zone lymphoma (MZL) usually have an indolent outcome, except for 10-20% of them that evolve into lymphoma with a histologic progression (HP-MZL). In this study we wished to determine the genetic alterations implicated in non-MALT HP-MZL, using an integrated analysis of transcriptomic data completed by a genome-wide copy-number changes (CNC) detection. We included 66 untreated patients, classified after features reviews as MZL for 52 cases and HP-MZL for 14 cases, with 4 matched cases. Gene expression profiling identified specifically in HP-MZL a proliferation signature and another signature regrouping oncogenes, genes involved in cell machinery and in signal. In addition, HP-MZL underexpressed genes such as p53. The most common genetic imbalances were the typical gains of 3/3q and 18q, and deletions in 7q, detected in both groups, SMZL and HP-SMZL. Array-CGH analysis showed no recurrent CNC specific of all HP-SMZL. Analysing each HP-SMZL and matched pairs, secondary CNC changes were identified either alone or in complexe combination in 4p, in 6p, and in 8q comprising several genes. Among them were located Cyclin G associated kinase, IRF4, and Myc respectively. These results highlight that deregulation of different pathways drive the MZL cells toward a higher proliferative rate charateristic of HP.

Introduction

Marginal Zone Lymphomas (MZL) represent between 7% and 15% of all lymphomas in adults and have been considered as a distinct clinicopathological entity since 19911. The International Lymphoma Study Group has individualized them into three distinct subtypes of MZL depending of the location : (1) the extranodal MZL of MALT type, and the non-MALT types with the (2) splenic and the (3) nodal MZL 2,3. The hallmark of the clinical presentation of these lymphomas is the indolent outcome with a median overall survival greater than 5 years4. In spite of this usually indolent outcome, histological progression (Rhodes, Yu et al.) may occur in 10-20% of patients with MZL at the time of recurrence, and is associated to a shorter median survival of approximately 2 years5. No criteria have been yet formally adopted to range HP, even in the more recent OMS classification6, although therapeutic impact is important.

HP in B cell lymphomas has been well described in the previous years before the global analyses era - the “omic era”-, particularly in follicular lymphomas7. Deregulation of several oncogenes have been associated with this feature such as mutation of P538,9, deletion of p1610, rearrangements of myc11 and bcl612. Few studies of gene expression profiling have also been reported in transformed follicular lymphomas13,14 identifying genes specifically linked to cell proliferation (C-MET, FGFR3, LT-β R, PDGF-Rβ) and particularly p38β MAPK13 and deregulation of MYC pathways12.

The goal of our study was to determine the genetic alterations implicated in the HP of non-MALT MZL using transcriptomic analysis, and to link the transcriptomic signatures specific to HP at the genetic level with a genome-wide copy-number-changes (CNC) analysis. Results showed that HP was related to a cell proliferation signature. Beside chromosomal alterations typical to MZL, secondary CNC could be detected in HP-MZL cases, without any recurrency in all HP-MZL, but they were identified either alone or in complex combination in HP-MZL cases. These CNC included genes

associated with cell proliferation such as Cyclin G associated Kinase, IRF4, and Myc.

Materials and Methods

Selection of patients. Fresh-frozen tumor biopsies and clinical data were obtained from 66 untreated patients of 3 institutions (SLS, HD, CHLS, France) after complete morphological analysis, including cytological, immunological, cytogenetic (conventional cytogenetic and fluorescent in situ hybridization (FISH)) and/or molecular analysis, to assess the diagnosis of non-MALT MZL. All patients had signed informed consent for biopsy analysis.

All MZL patients (n=52) had a non-MALT MZL classified as splenic MZL or nodal MZL (Table 1). All transformed MZL patients (n=14) presented a disseminated disease with splenic and/or nodal and/or extranodal involvement. Biopsies were performed on involved adenopathies in most of the patients except two cases with transformed disease in the spleen. In 4 patients, a non-transformed splenic sample and its transformed nodal counterpart were available. An additional group comprising 4 transformed follicular lymphomas and one transformed small lymphocytic lymphoma were also analysed by gene expression profiling as supplementary samples of transformed lymphoma.

HP was assessed by the presence of large cells, the presence of cohesive sheets of larges cells, and an elevated Ki67. Clinical and biological characteristics of the patients are shown in Table 2. In forty percent of the non-MALT MZL patients, treatment was delayed for a median time of 2.4 years after diagnosis. In the sixty percent of patients who required therapy at diagnosis, patients were treated either by splenectomy alone or by splenectomy combined with chemotherapy in 65% of the cases. All patients with transformed lymphoma were treated with anthracyclin-based combination chemotherapy (CHOP or ACVB15).

CGH-array procedures. The genome-wide CNC were analysed in 20 splenic MZL (SMZL), in 6 HP-SMZL cases including 2 matched cases using array comparative genomic hybridization (array-CGH) (Agilent Technologies, 105K. http://www.agilent.com). Three DNA from normal tissue (blood) of these cases were also analysed. Hybridization and scanning were performed according to the manufacturer procedures. Samples were labelled with Cyanin-3 and reference DNA was labelled with Cyanin-5.

Microarray procedures. Transcriptomic analysis of 32 SMZL and 8 HP-MZL, including 2 matched cases, was performed with a nylon DNA microarray of 9,216 human genes. Isolation of total RNA, cDNA synthesis with radioactive-labelling, hybridization and washing of the cDNA microarray membranes were performed as previously described[16].

Transcriptomic data analysis.

All sample measurements were rescaled using Lowess algorithm. From the initial 6692 clones, 2,478 were significantly expressed in a majority of the samples and were kept for further analyses. Average-linkage hierarchical clustering was applied to the MZL and HP-MZL samples (n=40 samples, n=53 hybridizations) to investigate relationships between all MZL samples and relationships between genes. Clusterings were obtained with Cluster and viewed with Treeview[17]. HP-FL and HP-SLL samples were added to the analysis as controls. This allowed us to visualize patterns of gene expression already detected in our previous work[1] and corresponding to precise functions (proliferation, early response,

differenciation, intracellular signaling), cell or tissue subtypes (T cells, stroma, lymph node / APC signature). Specific associations between the clusters identified in this study and those of other lymphoma profiling studies were explored using Fisher's exact test at 0.1% risk.

Genes discriminating histologic progression (HP) were determined with a t statistic at 0.1% risk with Bonferroni correction. The survival predictor was constructed using genes that discriminated dead and live patients, as previously described[18]. Functional characterization of the discriminating genes was obtained using the GoMiner software (http://www.discover.nci.nih.gov/gominer)[19]. Over- or under-representation of Gene Ontology (http://www.geneontology.org)[20] categories were determined for these discriminating genes. Significance of over- or under-representation was calculated using Fisher's exact test at 0.1% risk.

CGHarray data analysis. Data from all microarray channels were rescaled using Lowess algorithm and ratios sample to reference were calculated. Visual analyses first allowed us to detect aberrations (apparent log-ratio shift away from baseline) of multiple adjacent probes. A more systematic search was performed by calculating the probability of copy number change (CNC) for each probe in each case. Loss was considered significant when ratio was below 0.7 (loss) and gain when ratio was above 1.3. These thresholds correspond to a probability of CNC of 0.001. Individual CNC correspond to a single significant probe and CNC region when mean signal of 20 successive probes was below 0.7 or above 1.3. The frequency of occurrence for each detected CNC was calculated separately in MZL and HP-MZL.

We then searched for CNC regions specific of MZL or of HP-MZL. CNC regions were selected when significant for 3 samples of one type and none of the other type. For the 2 patients for which both MZL and HP-MZL samples were available, CNC regions were selected if significant in one type and not in the other.

Validation of the transcriptomic data and the genomic data. Quantitative RT-PCR for *HMMR*, was performed using Taqman technology (ABI Prism 7700 sequence detector system, Perkin Elmer/Applied Biosystems, Foster City, CA, USA) on samples isolated from 23 patients (23 MZL patients, including 18 splenic MZL and 9 HP-MZL counter). All PCR primers and Taqman probes were designed with Primer Express software (version 1.0: Perkin Elmer/Applied biosystems) using sequence available at the NCBI database. Primers, Taqman probes and PCR protocols are provided in the table 3A. **Quantitative real-time PCR** reactions were performed using the ABI Prism 7900HT Sequence Detection System (Applied Biosystems) with all samples run in duplicate. Primers and probes targeting introns of the *GAK, IRF4, MYC* and *MYO3A* genes were designed with the PrimerExpress 3.0 software (Applied Biosystems). Thermocycling for each reaction was carried out in a final volume of 20 µl containing 10 ng of genomic DNA. Primers and probes final concentrations were reported in the table 3A. After 2 minutes of uracil-N-glycosylase treatment at 50°C, followed by 10 minutes of polymerase activation, the cycling conditions of 40 cycles consisted of denaturation at 95°C for 15 seconds and annealing/elongation at 60°C for 1 minute. Gene quantities were determined using standard curves, constructed by serial dilutions of a pool of normal human genomic DNA. Gene copy numbers were reported as ratios of quantities of the target gene to *MYO3A* used as the

reference. This gene was selected as the least variant gene in all tumors.

Results

Categorization of gene-expression signatures. The overall expression patterns of samples and genes were first analyzed using hierarchical clustering to group the 53 analyses of the 40 samples according to the gene expression profiles of the 2,478 expressed genes (Figure 1). Examination of the data on the vertical axis highlighted clusters of correlated genes defining gene-expression signatures. These signatures were correlated to previously identified signatures[18] allowing us to rapidly categorize most of the genes clusters in either biological processes, or cell types.

Transformation discriminating genes defined specific pathways for histologic progression in MZL. We identified genes whose expression was consistently different between MZL samples and HP-MZL samples. Some of these genes could be related to two previously identified signatures[18], the proliferation signature and the stroma signature shown in Figure 1. The proliferation signature included *PCNA* that codes for a proliferation marker used in clinical practice, genes involved in glycolysis (*GAPD, LDHA*), in cell cycle (*CDC10*, *CDK4*), and cell proliferation (*TNFRSF13B, MCL1*) (Figure 2). The stroma signature included genes such as *MIG*, *STAT1*, *cathepsin B* and *C*, *CREG*, *IGFR2R*. These proliferation and stroma genes were over-expressed in all HP-MZL samples compared to MZL samples, and were also overexpressed in the HP-FL and HP-SLL control samples.

We identified down-regulated genes in the HP samples. Among them were *p53* which is well known to have an important role in cancer progression; *PMSB1* involved in the proteasome, *KLF2* and *PFC*. *KLF2* expression in a T-cell line induced reversible quiescence resulting in marked inhibition of proliferation and reduced expression of activation markers[23]. *PFC* is a plasma protein interacting with the complement pathway, and MAP kinase signalling such as *MAP2K4*, and *MAPK12*.

Several genes regrouped in a unique cluster of unknown function were specifically overexpressed in the HP-MZL. This cluster regrouped several oncogenes (*DEK*, *DJ-1*), genes related to cell machinery and cell motility (*HMMR*, *PRPF4*, *PPOX*) and genes involved in signal transduction (*ZNF146*, *FLJ10618*, *PBXL2*). None of the other HP lymphoma samples (HP-FL or HP-SLL) over-expressed these genes. Six MZL samples (see Figure 2 on the left) over-expressed these particular genes, but they did not over-express the genes related to the proliferation and the stroma signatures. Morphological characteristics and clinical characteristics (sex, age, PS, stage, presence of anemia, monoclonal component, elevated level of LDH and B2 microglobulin, immune event, IPI score, treatment, complete response after the first treatment, progression-free survival, overall survival) of these 6 MZL samples were not significantly different from the other splenic MZL samples (Mann-Whitney test, data not shown). None of these patients had undergone histological transformation with a median follow-up of 5 years (range: 3.1 – 8.6 years).

Copy numbers changes. Array-CGH analysis showed abnormalities at the chromosomal level in 14 (70%) patients with MZL and HP-MZL. The most common imbalances detected were gains of 3/3q (n=11/61, 18 %) and 18q (n=13/61, 21%), and deletions in 7q (n=18/61, 17%). These imbalances were detected in both groups, MZL and HP-MZL (Figure 3).

By calculating the probability of CNC, we searched for gains or loss specific of MZL or HP-MZL (Figure 4A). No recurrent CNC was specific of all MZL cases nor of all HP-MZL cases. Only 1 CNC was present in 4 of the 6 HP-SMZL cases, 3 in 3 cases, and 29 in 2 cases. These CNC could also be seen in MZL cases. HP-MZL cases exhibited HP-MZL specific CN gains. One was located in 8q involving the oncogene *MYC*. Others were located in chr5, chr8, and chr14. They respectively involved genes of G protein GRS14, SLC34A1, caogulation factor XII in chr5, gene of a zinc Finger Protein in chr8. (figure 4A).

Taking advantage of the matched pairs, MZL and HP-MZL samples from the same patient, we searched for secondary CNC changes (Figure 4B). In the 2 pairs, we could detect recurrent CNC described as follow. In one pair, we could detect five CN gains: one in chr6, two in chr9, one in chr11, and one in chr13. The one located in 6p included 5 genes, comprising *IRF4 (*Figure 4B.*)*. In the other pair, one CN gain specific to HP-MZL was located in 4p and includes 9 genes. Among them, was Cyclin G associated kinase. This CNC was also present in 2 other HP-MZL cases.

These 3 genes, *MYC, cyclin G associated kinase,* and *IRF4,* are known to be involved in proliferation.

Validation.

Quantitative Fluorescent RT-PCR validation of microarray data. Using quantitative real-time PCR, we validated one gene overexpressed in HP-MZL in 28 patients (Table 4). The relative amount of *HMMR* was significantly increased in MZL cases - all splenic MZL (n=19)-, compared to HP-MZL patients (n=8) ($p< 0.0001$) (Table 3).

Validation of 3 candidate genes identified in the CNC by real-time PCR.

Real-time quantitative PCR precisely mirrored the gene copy number for all three candidate genes (*GAK*, *IRF4* and *MYC*) relative to the reference gene MYO3A in panels of tumors with MZL or HP-MZL histology, confirming their amplification in HP-MZL (table 5B).

Discussion

In this study, we reported a comprehensive view of histologic progression of non-MALT MZL with an integrated genetic and transcriptomic analyses.

Histological lymphoma transformation is a complex process that will provide the cell with an advantage in terms of aggressiveness and survival. Some evidence from the analyses performed in follicular lymphoma transformation, the best studied subtype of transformed lymphoma, suggests that increased resistance to apoptosis with aberrant expression of proteins such as BCL2, enhanced cell growth promotion and proliferation with cell cycle deregulation may yield a protective advantage during transformation and/or progression of low grade lymphoma to high grade lymphoma[24].

We performed a transcriptomic analysis of the HP – MZL, which originate from a different cell type than follicular lymphoma. This study revealed that processes common with transformation of follicular lymphoma could be observed, on the basis of two lines of evidence: 1/ many of the genes identified were common with those listed in previous reports of gene expression analyses[25,26] and 2/ dysregulated expression of these common genes was observed in HP-MZL samples and the HP-FL and HP-SLL samples hybridized as controls. Higher levels of genes belonging to the designed proliferation signature were observed in HP-MZL samples than in MZL

samples. Genes involved in cell cycle (*CDC10, CDK4*) and genes associated with cell growth and proliferation (*GAPDH, PCNA, LDHA*) were over-expressed in all the HP-MZL and in all the samples of other analysed HP-lymphoma subtypes, HP-FL and HP-SLL. This is consistent with the gene-expression profiling studies previously reported in transformed follicular lymphoma[26]. Interestingly we also found a large number of genes belonging to the stroma signature which were dysregulated in HP-MZL and in the other HP- subtypes. This observation suggests a different reaction in the stroma when proliferation is more aggressive than in the low grade lymphomas, and stress the fact that the microenvironnement plays a major role in different types of lymphoma, as it has previously been reported in follicular lymphoma[27]. Abnormalities of the *p53* tumor suppressor gene such as mutations or deletions have been described as having a pivotal role in oncogenesis, particularly in HP of follicular lymphoma[8,28]. Consistently with this earlier report, we observed in all the transformed samples, including HP-MZL and other HP-subtypes, a low level of p53 transcript associated with other down-regulated genes such as *KLF2* and *PFC*.

Besides these common processes, a specific cluster could be identified. This cluster regrouped genes such as *LIMK1*, *ZNF146*, *FLJ10618*, *PBXL2*, *HMMR*, *NPM1*, *PPOX* and *DEK* oncogene but no specific function could be associated with this cluster. These genes were up-regulated in all the HP-MZL, but down-regulated in all the other HP- lymphomas and in the majority of non-HP MZL lymphomas. The 6 MZL samples in which these genes were not down-regulated did not over-express the previously described genes related to HP. They did not exhibit any HP features, neither morphologically nor clinically. Moreover their outcome did not differ from that of other MZL patients. These data suggest that the alteration of these specific genes is perhaps one of the preliminary steps of MZL HP but is not sufficient for full-blown HP to occur. Until genes related to proliferation and stroma signatures become further altered, no HP will be observed. We hypothesize that HP is a common process for indolent lymphomas involving genes related to proliferation and inhibition of apoptosis, as well as microenvironment-related genes. In the case of MZL, a specific process could be engaged for HP to occur, once these more general processes involving cell proliferation and inhibition of apoptosis are activated.

Ruiz-Ballesteros E. et al. reported gene expression analysis of 27 splenic MZL samples[29] and identified that NF-κB-related genes contained the largest group of genes whose expression was associated with changes in survival probability, compared to the other functional clusters considered: apoptosis, BCR-signaling, cell cycle and stress response. None of the genes was individually reported as being independent prognostic factor for overall survival. Only two genes, *PKCA* and *ADAM17*, were related to the progression-free interval probability as independent value.

In conclusion, these results highlight that deregulation of different pathways drive the MZL cells toward a higher proliferative rate charateristic of histologic progression.

References

1. Jaffe ES, Harris NL, Stein H, Vardiman JW. World Health Organization Classification of Tumours: Pathology and genetics of tumours of haematopoietic and lymphoid tissues. Lyon: IARC; 2001.
2. Harris N, Jaffe E, Stein H, et al. A revised European-American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group. Blood. 1994;84:1361-1392.
3. Jaffe E, Harris N, Stein H, Vardiman J. World Health organization Classification of tumours. Tumours of haematopoietic and lymphoid tissues. Lyon: IARC; 2001.
4. Berger F, Felman P, Thieblemont C, al. e. Non-MALT marginal zone B-cell lymphomas : a description of clinical presentation and outcome in 124 patients. Blood. 2000;95:1950-1956
5. Thieblemont C, Felman P, Callet-Bauchu E, et al. Splenic marginal-zone lymphoma: a distinct clinical and pathological entity. Lancet Oncol. 2003;4:95-103.
6. Jaffe E, Harris N, Stein H, Vardiman J. World Health organization Classification of tumours. Tumours of haematopoietic and lymphoid tissues. Lyon: IARC; 2008.
7. Swerdlow S, Campo E, Harris N, et al. Follicular lymphoma. In: Press W, ed. WHO classification of tumours of haematopoietic and lymphoid tissue. Lyon: IARC; 2008:220-226.
8. Lo Coco F, Gaidano G, Louie D, Offit K, Chaganti R, Dalla-Favera R. p53 mutations are associated with histologic transformation of follicular lymphoma. Blood. 1993;82:2289-2295.
9. Sander C, Yano T, Clark H, et al. p53 mutation is associated with progression in follicular lymphomas. Blood. 1993;82:1994-2004.
10. Elenitoba-Johnson K, Gascoyne R, Lim M, Chhanabai M, Jaffe E, Raffeld M. Homozygous deletions at chromosome 9p21 involving p16 and p15 are associated with histologic progression in follicle center lymphoma. Blood. 1998;91:4677-4685.
11. Yano T, Jaffe E, Longo D, Raffeld M. MYC rearrangements in histologically progressed follicular lymphomas. Blood. 1992;80:758-767.
12. Lossos I, Levy R. Higher-grade transformation of follicle center lymphoma is associated with somatic mutation of the 5' noncoding regulatory region of the BCL-6 gene. Blood. 2000;96:635-639.
13. Elenitoba-Johnson K, Jenson S, Abbott R, et al. Involvement of multiple signaling pathways in follicular lymphoma transformation: p38-mitogen-activated protein kinase as a target for therapy. Proc Natl Acad Sci U S A. 2003;100:7259-7264.
14. Lossos I, Alizadeh A, Diehn M, et al. Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes. Proc Natl Acad Sci U S A. 2002;99:8886-8891.
15. Coiffier B, Bryon P, Ffrench M, et al. Intensive chemotherapy in aggressive lymphomas: updated results of LNH-80 protocol and prognostic factors affecting response and survival. Blood. 1987;70:1394-1399.
16. Thieblemont C, Nasser V, Felman P, et al. Small lymphocytic lymphoma (SLL), marginal zone B-cell lymphoma (MZL), mantle cell lymphoma (MCL) exhibit distinct gene-expression profiles allowing molecular diagnosis. Blood. 2004.
17. Eisen M, Spellman P, Brown P, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95:14863-14868.

18. Thieblemont C, Nasser V, Felman P, et al. Small lymphocytic lymphoma, marginal zone lymphoma, and mantle cell lymphoma exhibit distinct genomic profiles allowing molecular diagnosis. Blood. 2004;103:2727-2737.
19. Zeeberg B, Feng W, Wang G, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 2003;4:R28.
20. Ashburner M, Ball C, Blake J, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25-29.
21. Kaplan E, Meier P. Non-parametric estimation from incomplete observations. J Am Stat Assoc. 1958:457-481.
22. Peto R, Pike M, Armitage P. Design and analysis of randomized clinical trials requiring prolonged observations of each patients:II. Analysis and examples. Br J Cancer. 1997;35:1-39.
23. Buckley AF, Kuo CT, Leiden JM. Transcription factor LKLF is sufficient to program T cell quiescence via a c-Myc-dependent pathway. Nature Immun. 2001;2:698-704.
24. Lossos I, Levy R. Higher grade transformation of follicular lymphoma: phenotypic tumor progression associated with diverse genetic lesions. Semin Cancer Biol. 2003;13:191-202.
25. Lossos I, Alizadeh A, Diehn M, et al. Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes. Proc Natl Acad Sci U S A. 2002;99:8886-8891.
26. Elenitoba-Johnson K, Jenson S, Abbott R, et al. Involvement of multiple signaling pathways in follicular lymphoma transformation: p38-mitogen-activated protein kinase as a target for therapy. Proc Natl Acad Sci U S A. 2003;100:7259-7264.
27. Dave S, Wright G, B T, et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. N Engl J Med. 2004;351:2159-2169.
28. Sander C, Yano T, Clark H, et al. p53 mutation is associated with progression in follicular lymphomas. Blood;82:1994-2004.
29. Ruiz-Ballesteros E, Mollejo M, Rodriguez A, et al. Splenic marginal zone lymphoma: proposal of new diagnostic and prognostic markers identified after tissue and cDNA microarray analysis. Blood. 2005;106:1831-1838.

Tables and figures

Table 1. **Phenotypic characteristics of the 47 patients whose biopsy has been hybridized on microarrays.** Blood samples were used for microarray hybridisation when frozen material (spleen biopsy [n=9] for splenic MZL or node biopsy [n=1])) for transformed FL) was not available for RNA extraction. All blood samples had more than 50% of lymphoma cells.

Table 2. Clinical characteristics and treatment of the 47 patients

Table 3. Validation by Real-time RT PCR validation and quantitative PCR validation of the gene of interest from the transcriptomic analysis (HMMR). A. Primers and probe used to validate HMMR B. Results. We observed a significantly different expression between splenic MZL cases and HP-MZL cases when considering HMMR expression.

Figure 1. Hierarchical clustering of gene expression data. A. Measurements of gene expression from 59 microarray analyses of the 47 lymphoma samples are depicted. Each column represents the genomic profile of the 2,478 expressed genes for one tumor, and each row represents the relative level of expression of one cDNA clone.

Figure 2. Transformation discriminating genes

Figure 3. Chromosomal imbalances in MZL and HP-MZL. Frequency of Copy Number Changes in MZL and HP-MZL. Frequencies of DNA copy number imbalances in 14 MZL (top, in grey) and in 6 HP-MZL (bottom, in black) were plotted in against their chromosome position. Losses and gains corresponded to mean signal of 20 successive probes was below 0.7 or above 1.3 respectively in individual sample. The most common imbalances detected were gains of 3/3q (n=11/61, 18 %) and 18q (n=13/61, 21%), and deletions in 7q (n=18/61, 17%). These imbalances were detected in both groups, MZL and HP-MZL

Figure 4. CNC specific of MZL or HP-MZL. No recurrent CNC was specific of all MZL cases nor of all HP-MZL cases.. **B. Recurrent CNC detected first in a matched pair of MZL and HP-MZL from the same patients**

Figures 5. A. Primers and probes sequences of the genes quantified by real-time PCR
GAK, *IRF4* and *MYC* are the genes of interest in HP-MZL and *MYO3A* is the reference gene which was unaltered in all tumors. **B. Gene copy number assessment of candidate genes by quantitative real-time PCR.** This analysis substantiated the gene copy numbers observed by array-CGH for all three candidate genes *GAK*, *IRF4* and *MYC* relative to the reference gene *MYO3A*. Hatched histograms represent the mean value in MZL tumors while grey histograms represent the mean value in transformed MZL tumors.

Table 1.

	Total	MZL	Transformed		
			MZL	FL	SLL
	n= 47	n=34	n=8	n=4	n=1
Samples					
Spleen	25 (53)	23 (68)	2 (25)	0	0
Lymph node	12 (26)	2 (6)	6 (75)	3 (75)	1 (100)
Blood	10 (21)	9 (26)	0	1 (25)	0
Immunophenotype					
CD19/CD20 +	47 (100)	34 (100)	8 (100)	4 (100)	1 (100)
CD5 +	6 (13)	5 (15)	0	0	1 (100)
CD23 +	1 (2)	0	0	0	1
CD43 +	1 (2)	0	0	0	1
CD10 +	4 (8)	0	0	4	0
Cytogenetic features					
del 13q	0	0	0	0	0
+12	0	0	0	0	1 (100)
del 11q	0	0	0	0	0
+3 or +3q	9 (19)	8 (23	1/6 (16)	0	0
del 7q	7 (15)	6 (18)	1/6(16)	0	0
t(11;14)	2 (4)	2 (6)	0/6	0	0
P53 neg (FISH)	8 (17)	3 (9)	2/6 (33)	3 (75)	0
***Ig*V_H mutation status**					
Mutated (<98%)	32	24 (73)	7 (100)	2 (100)	-
Unmutated (≥98%)	9	9 (26)	0	0	-
Not done	5	1 (6)	1	2	1
Percentage of large Cells	-	-	65	90	90
Median (%)			10 - 90	80-100	-
Range (%)					
Percentage of Ki67 staining					
Median (%)	-	-	29% when diffuse	59	40
Range (%)			3-50	8-80	-

Table 2.

	Total	MZL	Transformed		
			MZL	FL	SLL
	n=47	n=34	n=8	n=4	n=1
Sex					
Male	20	14	2	4	0
Female	27	20	6	0	1
Age					
Median, years	60	63	58.5	55.5	73
Range, years	41 – 85	41-85	50-78	45-60	-
PS (ECOG)					
0-1	34	28	4	1	1
2-4	12	5	4	3	0
Not available	1	1	0	0	0
Stage					
I-II	1	1	0		0
III-IV	46	33	8	4	1
Abnormal LDH	18	9	5	4	0
Not available	7	7	0	0	0
β2 microglobulin ≥ 3 mg/l	23	15	5	3	0
Not available	7	7	0	0	0
Hemoglobin < 10 G/dl	12	8	3	1	0
Not available	2	2	0	0	0
Monoclonal component	14	11	3	0	0
Not available	5	5	0	0	0
Immune event	10	8	2	0	0
Not available	2	2	0	0	0
IPI					
Low	3	2	1	0	0
Low-Intermediate	7	6	0	1	0
High-Intermediate	19	13	4	1	1
High	11	6	3	2	0
Not available	7	7	0	0	0
Treatment					
No treatment	14	14	0	0	0
Splenectomy	4	4	0	0	0
Monochemotherapy*	9	9	0	0	0
Splenectomy + chemo.	5	4	2	0	0
Polychemotherapy*	15	3	7	4	1

Monochemotherapy Chlorambucil or Fludarbine
Polychemotherapy CHOP or CHOP like (ACVB)

Table 3.

A.

Gene name		Sequences (5'→3')	C
HMMR	Forward primer		nM
	Reverse primer		
	Probe		nM

B.

Genes discriminating for transformation	MZL patients n = 18	Transformed MZL pts n = 9	p (Mann-Whitney)
HMMR, median (range) standard deviation	0.08 (0.006-0.33) 0.095	2.56 (0.8-9.65) 3.08	**0.00001**

Figure 1. Hierarchical clustering of gene expression data

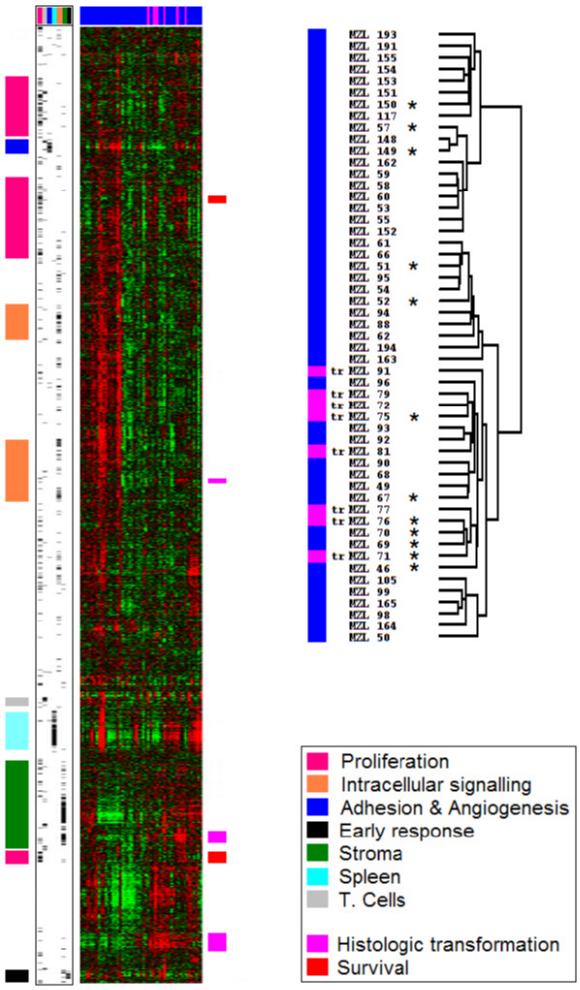


Figure 2.

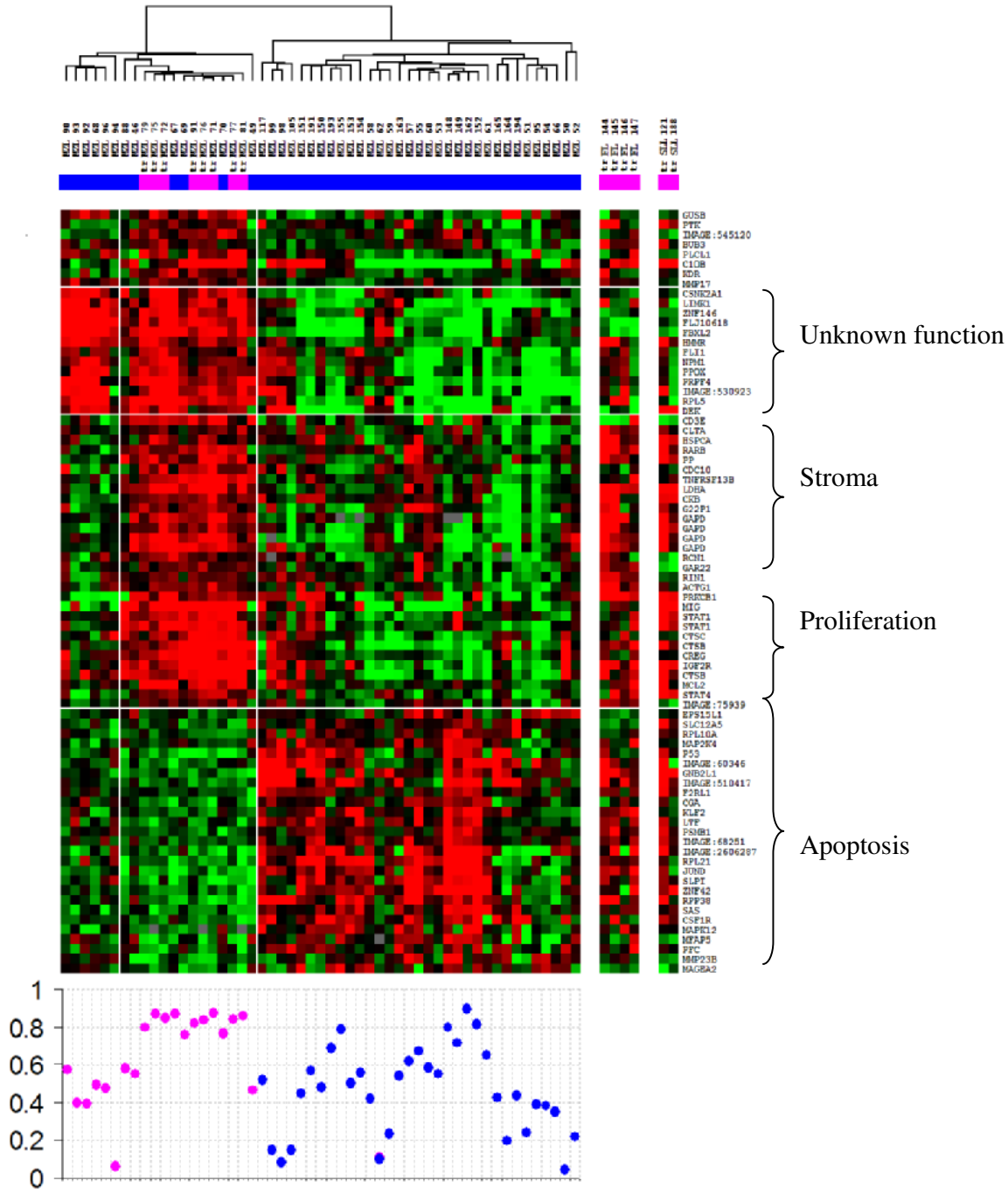

Unknown function
Stroma
Proliferation
Apoptosis
1
0.8
0.6
0.4
0.2
0

Figure3.

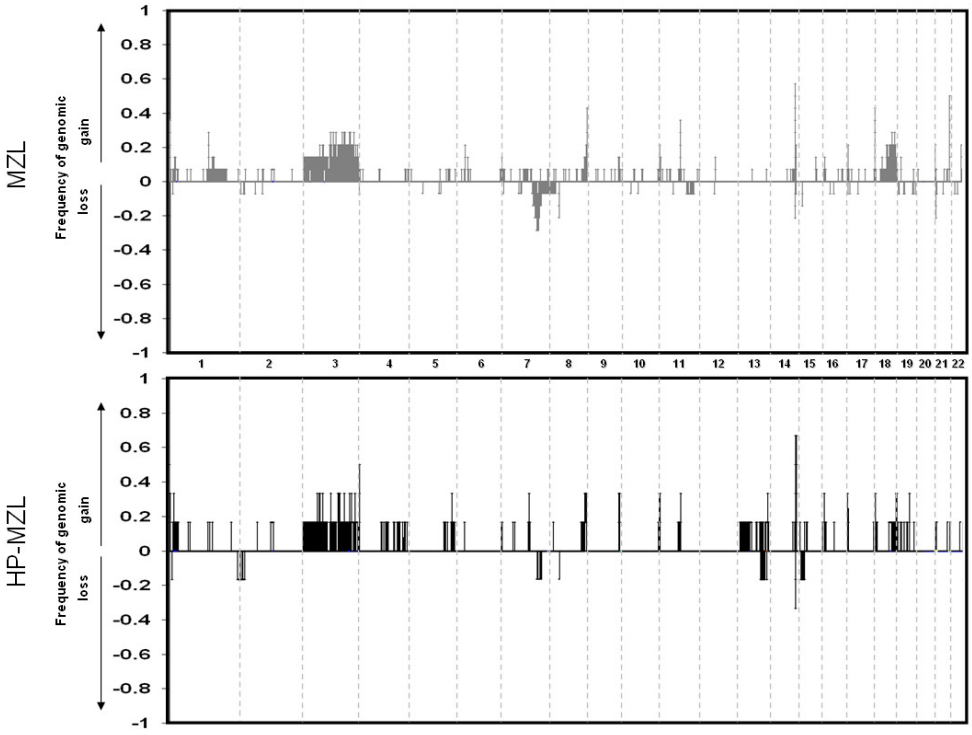
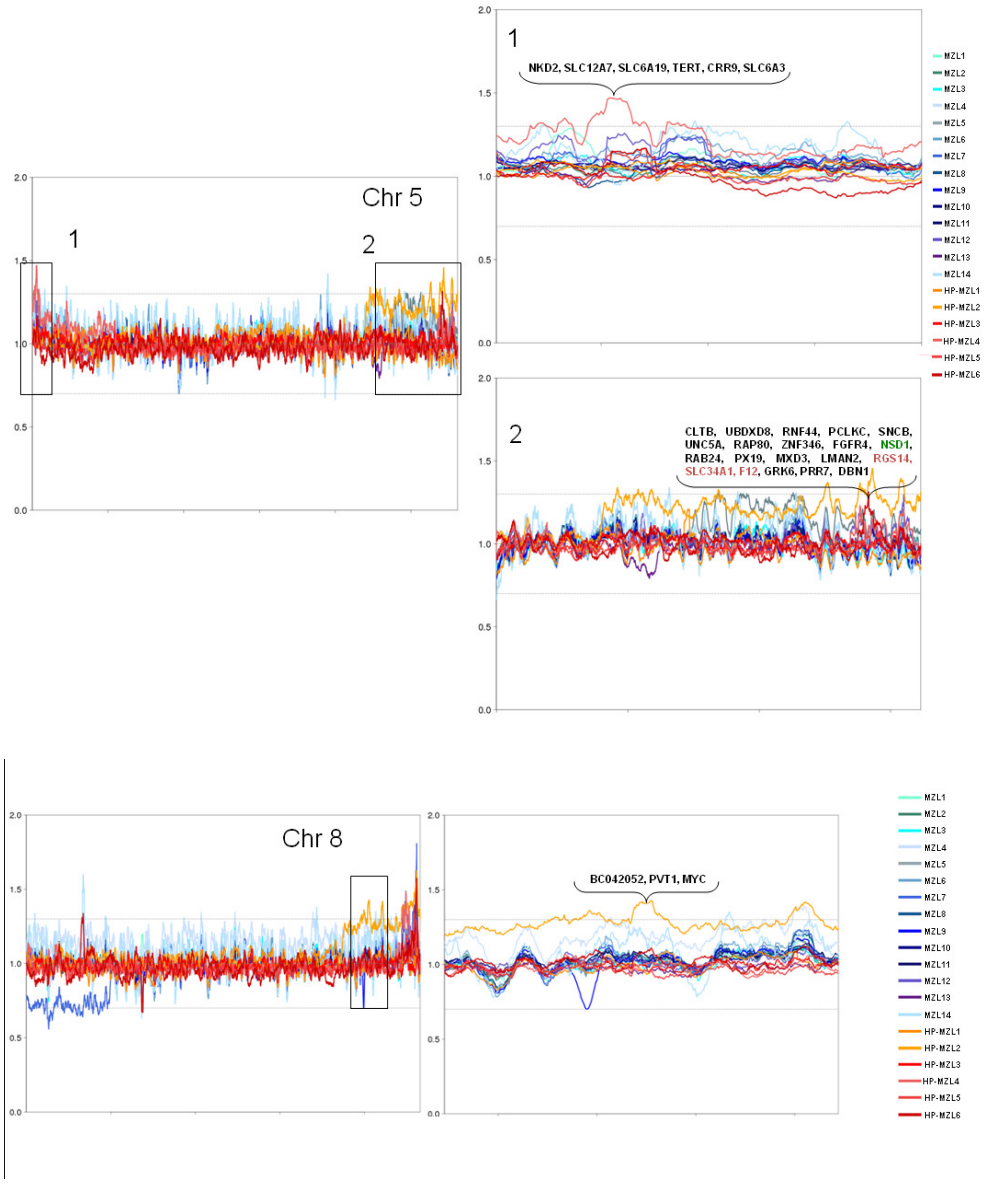
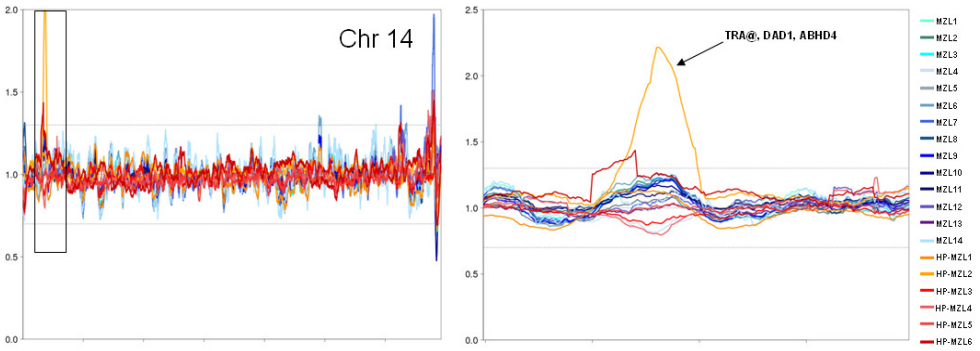

MZL
HP-MZL
Frequency of genomic
gain
loss
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

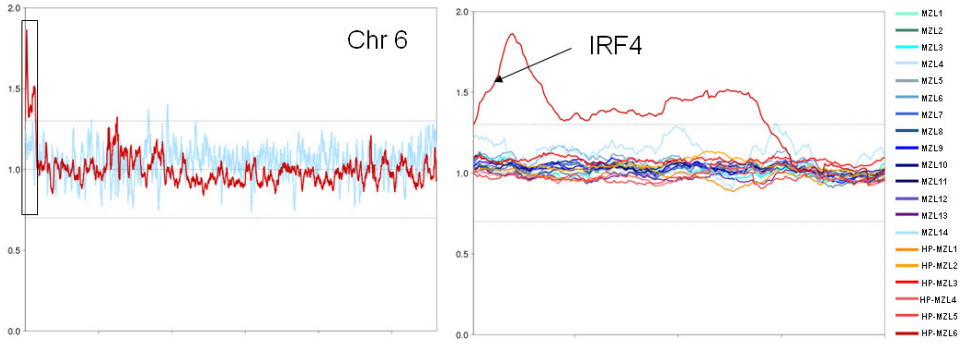
Figure 4.

A.


Chr 5
1
2
NKD2, SLC12A7, SLC6A19, TERT, CRR9, SLC6A3
CLTB, UBDXD8, RNF44, PCLKC, SNCB, UNC5A, RAP80, ZNF346, FGFR4, NSD1, RAB24, PX19, MXD3, LMAN2, RGS14, SLC34A1, F12, GRK6, PRR7, DBN1
Chr 8
BC042052, PVT1, MYC
MZL1
MZL2
MZL3
MZL4
MZL5
MZL6
MZL7
MZL8
MZL9
MZL10
MZL11
MZL12
MZL13
MZL14
HP-MZL1
HP-MZL2
HP-MZL3
HP-MZL4
HP-MZL5
HP-MZL6


Chr 14
TRA@, DAD1, ABHD4
MZL1
MZL2
MZL3
MZL4
MZL5
MZL6
MZL7
MZL8
MZL9
MZL10
MZL11
MZL12
MZL13
MZL14
HP-MZL1
HP-MZL2
HP-MZL3
HP-MZL4
HP-MZL5
HP-MZL6

B.


Chr 6
IRF4
MZL1
MZL2
MZL3
MZL4
MZL5
MZL6
MZL7
MZL8
MZL9
MZL10
MZL11
MZL12
MZL13
MZL14
HP-MZL1
HP-MZL2
HP-MZL3
HP-MZL4
HP-MZL5
HP-MZL6

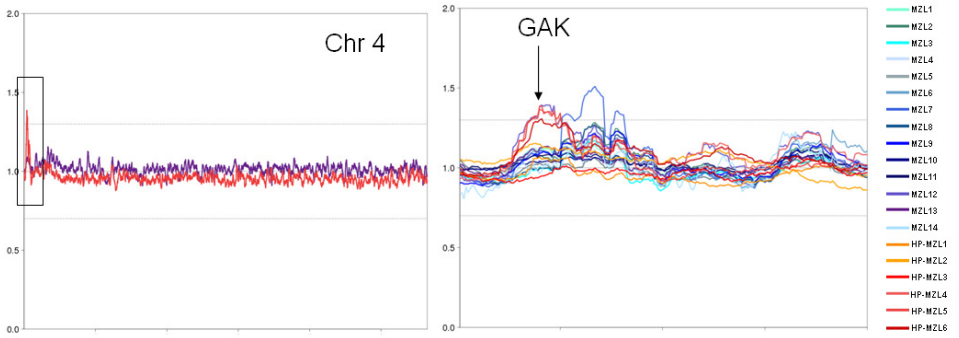
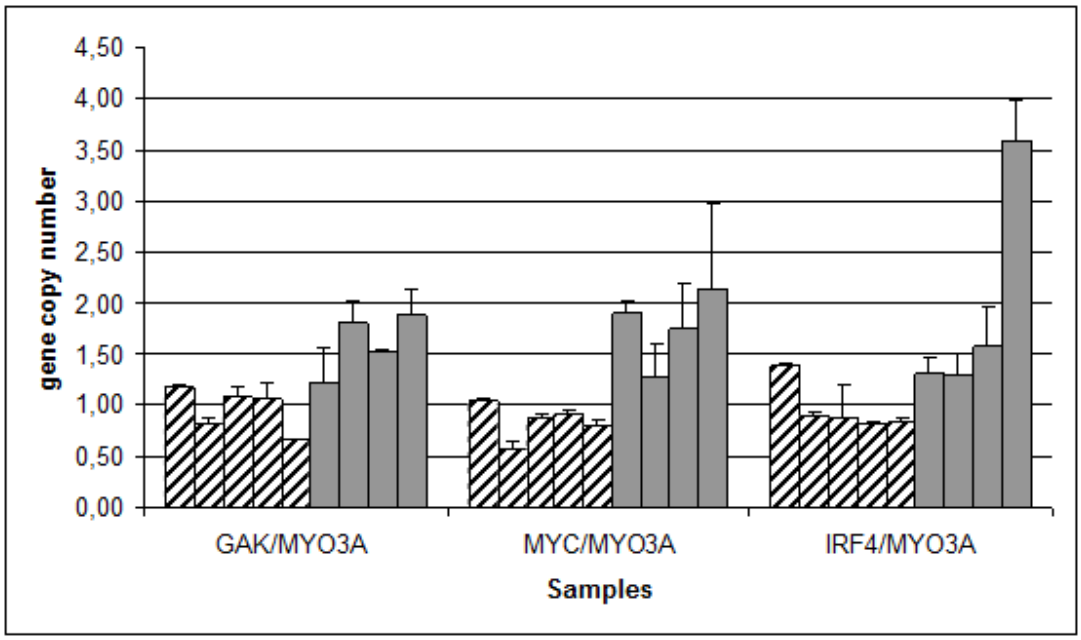

Chr 4
GAK
MZL1
MZL2
MZL3
MZL4
MZL5
MZL6
MZL7
MZL8
MZL9
MZL10
MZL11
MZL12
MZL13
MZL14
HP-MZL1
HP-MZL2
HP-MZL3
HP-MZL4
HP-MZL5
HP-MZL6

Figure 5.

A.

Gene name		Sequences (5'→3')	C
GAK	Forward primer	GGGCACCCATACCGTACCT	300 nM
	Reverse primer	CTCCCGTAGCCTCCATTCTG	
	Probe	6-FAMCGTGAGTTGTCAGTTTAGAGA-MGB	200 nM
IRF4	Forward primer	GTCACCACGTTGACGTTACATATTT	500 nM
	Reverse primer	TTTCGACATCCCAGGATCATTT	
	Probe	6FAM-CCTGAGTTACGTGGATGT-MGB	150 nM
MYC	Forward primer	CCATTTTCATTGGCAGCTTATTT	400 nM
	Reverse primer	GCAAACATGGGCAGTCTAAGG	
	Probe	6FAM-CTTGGGCTTTAGCGTTT-MGB	400 nM
MYO3A	Forward primer	TCTCCTCTGAAGATGGGACCTAGT	400 nM
	Reverse primer	TCCCCTTTTTATCCTCCATTGA	
	Probe	6FAM-TCTCTGTTGGGCGGTGG-MGB	300 nM

B.


4,50
4,00
3,50
3,00
2,50
2,00
1,50
1,00
0,50
0,00
gene copy number
GAK/MYO3A
MYC/MYO3A
IRF4/MYO3A
Samples

5. TRANSCRIPTOME + MOTIF+ChIP-chip

PGC-1-RELATED COACTIVATOR MODULATES MITOCHONDRIAL-NUCLEAR CROSSTALK THROUGH ENDOGENOUS NITRIC OXIDE

Mahatsangy Raharijaona [1,2*], Soazig Le Pennec[3,4*], Julie Poirier[3,4], Delphine Mirebeau-Prunier[3,4,5], Clothilde Rouxel[3,4], Caroline Jacques[3,4], Jean-Fred Fontaine[3,4], Yves Malthiery[3,4,5], Rémi Houlgatte [1,2], Frédérique Savagner[1,3,4,5]

[*] These authors have contributed equally to this article.

From:

1 : INSERM UMR915, l'Institut du thorax, Nantes, F-44035

2 : Université de Nantes, Nantes, F-44035

3 : INSERM UMR694, Angers, F-49033

4 : Université d'Angers, Angers, F-49033

5 : CHU Angers, Laboratoire de Biochimie, Angers, F-49033

Running Title: PRC modulates mitochondrial-nuclear crosstalk

Key words: Mitochondrial-nuclear crosstalk, PGC related coactivator, cell cycle

Address correspondence to: Frédérique Savagner, Inserm UMR 694, Laboratoire de Biochimie, CHU, 4 rue Larrey, 49033 Angers, France, Tel: +33 241 35 33 14, Fax: +33 241 35 40 17, Mail: frederique.savagner@univ-angers.fr.

Plusieurs fonctions cellulaires essentielles de la mitochondrie dépendent des interactions fonctionnelles entre le génome mitochondrial et le génome nucléaire. Les mécanismes gouvernant la coordination des facteurs de transcription nucléaires impliqués dans la biogénèse mitochondriale a été en partie expliqué par la découverte de la famille des coactivateurs PGC-1 (peroxisome proliferator-activated receptor gamma, coactivator 1). Trois membres de cette famille - PGC1α, PGC1β et PRC (PGC-1 related coactivator) – régulent plusieurs fonctions, dont

le métabolisme glucidique, la thermogénèse adaptative, l'oxydation des acides gras, le métabolisme mitochondrial, via des interactions avec des facteurs de transcription nucléaires. La biogénèse mitochondriale est contrôlée par PGC principalement par des interactions avec des facteurs respiratoires nucléaires NRF-1 et NRF-2. Alors que PGC-1α est exprimé dans des organes et tissus aux grands besoins énergétiques (foie, le cœur, les muscles squelettiques…), l'expression de PRC dépend du cycle cellulaire. Il joue un rôle d'intégrateur des voies de signalisation dirigeant la fonction respiratoire mitochondriale et la croissance cellulaire. PRC et PGC-1α ciblent quelques facteurs de transcription communs, comme NRF-1. Des interactions fortes ont été montrées entre PRC et NRF-2, ERRα ou CREB. PRC contrôle la croissance cellulaire par des mécanismes qui semblent être indépendants de son effet sur la fonction mitochondriale. PRC a également été décrit comme un régulateur positif de la chaîne respiratoire. Cependant le rôle précis de PRC dans la biogénèse mitochondriale et croissance cellulaire doit être élucidé.

Une dérégulation du statut fonctionnel mitochondrial peut avoir un effet rétroactif sur la machinerie transcriptionnelle nucléaire. Il a été suggéré que ce signal rétrograde influence la communication bidirectionnelle entre le noyau et la mitochondrie par une voie dépendent de PGC-1. D'autre part, les fonctions mitochondriales peuvent être régulées par des mécanismes complexes impliquant l'oxyde d'azote (NO). La réponse aiguë au NO, connu pour induire la production de radicaux libres oxygénés, est une inhibition réversible de la chaîne respiratoire. A plus long terme, associé avec cGMP, il déclenche la biogénèse mitochondriale par l'intermédiaire de la voie PGC-1α.

PRC semble être impliqué dans la réponse précoce aux signaux rétrogrades.

Dans l'oncocytome thyroïdien, qui constitue un modèle de tumeurs riches en mitochondries, un fort taux d'induction de PRC est observé, sans atteintes des fonctions mitochondriales (ref 15). Ces effets sont confirmés dans une lignée cellulaire dérivée d'un oncocytome thyroïdien oxyphile, XTC.UC1, s'avérant ainsi être un bon modèle pour l'étude de la coordination entre le génome mitochondrial et le génome nucléaire.

Afin d'apprécier l'importance du dialogue noyau-mitochondrie par l'intermédiaire de PRC, nous avons cherché les processus cellulaires qu'il régule dans la lignée cellulaire XTC.UC1.

Pour cela, des cellules XTC.UC1 ont été synchronisées par privation de sérum. J'ai participé à l'analyse du transcriptome de ces cellules à la suite de l'inhibition de PRC par un siRNA à différents temps après induction du cycle cellulaire. (Puce à protéine : MAPK)

Nous avons montré que le monoxyde d'azote influence l'expression de PRC au niveau transcriptionnel. Concernant le métabolisme mitochondrial énergétique, nous avons observé que PRC contrôle les complexes de la chaîne respiratoire et l'efficacité du couplage afin de maintenir l'homéostasie mitochondriale. En utilisant le SiRNA contre PRC, nous avons identifié aussi bien des activations que des répressions transcriptionnelles plus ou moins précoces dépendantes de PRC. Ceci englobe néanmoins des effets indirects du coactivateur PRC. Nous montrons que PRC intervient dans la régulation du cycle cellulaire et du métabolisme mitochondrial par l'intermédiaire de facteurs de transcription spécifiques. PRC intervient dans la voie de signalisation de ASK1 (apoptosis signal-regulating kinase 1) sensible au statut redox de la cellule et impliqué dans l'apoptose, la prolofération ou la différenciation des cellules. Nous postulons le fait que PRC assure une adaptation rapide à l'environnement cellulaire par la régulation de quelques gènes cibles comme COX15 du complexe IV de la chaîne respiratoire. Il régule séparément les sous-unités des complexes de la chaîne respiratoire, menant à une modulation rapide des fonctions mitochondriales en intégrant la régulation du complexe IV avec l'efficacité du couplage et les processus de détoxification des espèces réactives oxygénées.

PGC-1-RELATED COACTIVATOR MODULATES MITOCHONDRIAL-NUCLEAR CROSSTALK THROUGH ENDOGENOUS NITRIC OXIDE

Mahatsangy Raharijaona [1,2*], Soazig Le Pennec[3,4*], Julie Poirier[3,4], Delphine Mirebeau-Prunier[3,4,5], Clothilde Rouxel[3,4], Caroline Jacques[3,4], Jean-Fred Fontaine[3,4], Yves Malthiery[3,4,5], Rémi Houlgatte [1,2], Frédérique Savagner[1,3,4,5]

[*] These authors have contributed equally to this article.

From:

1 : INSERM UMR915, l'Institut du thorax, Nantes, F-44035

2 : Université de Nantes, Nantes, F-44035

3 : INSERM UMR694, Angers, F-49033

4 : Université d'Angers, Angers, F-49033

5 : CHU Angers, Laboratoire de Biochimie, Angers, F-49033

Running Title: PRC modulates mitochondrial-nuclear crosstalk

Key words: Mitochondrial-nuclear crosstalk, PGC related coactivator, cell cycle

Address correspondence to: Frédérique Savagner, Inserm UMR 694, Laboratoire de Biochimie, CHU, 4 rue Larrey, 49033 Angers, France, Tel: +33 241 35 33 14, Fax: +33 241 35 40 17, Mail: frederique.savagner@univ-angers.fr.

Abstract

Background
The PGC-1 related coactivator (PRC), which shares structural and functional features with PGC-1α, is believed to regulate several metabolic pathways as well as mitochondrial biogenesis. Its involvement in the early programming of cell proliferation suggests the existence of finely regulated crosstalk between mitochondrial functions and the cell cycle status.
Methodology/principal findings
PRC-regulated pathways were explored in a cell-line model derived from mitochondrial-rich tumours with an essentially oxidative metabolism and specifically high PRC expression. The functional status of mitochondria was compared to the results of microarray analysis under conditions of temporal PRC inhibition. To specify the fine PRC regulation, the expression levels of the genes and proteins involved in the oxidative phosphorylation process were studied by real time quantitative PCR and western blotting. As in earlier studies on PGC-1α, we investigated the role of nitric oxide in PRC-regulated mitochondrial biogenesis and determined its action in the control of the phosphorylation status of the mitogen-activated protein kinase pathway.
Conclusion/significance
We found that nitric oxide rapidly influences PRC expression at the transcriptional level. Focusing on mitochondrial energetic metabolism, we observed that PRC differentially controls respiratory chain complexes and coupling efficiency in a time-dependent manner to maintain mitochondrial homeostasis. Our results highlight the key role of PRC in the rapid modulation of metabolic functions in response to the status of the cell cycle.

Introduction

Several essential cellular functions of mitochondria depend on a high degree of functional interaction between the nuclear and mitochondrial genomes. Of the hundred structural subunits that make up the oxidative phosphorylation (OXPHOS) complexes, 13 are encoded by the mitochondrial genome. The mechanisms governing the coordination of the multiple transcription factors involved in mitochondrial biogenesis have been partly explained by the discovery of the PGC-1 coactivator family [1]. Three members of this family – PGC1α, PGC1β and PRC – regulate several functions, including adaptative thermogenesis, glucidic metabolism, fatty acid oxidation and mitochondrial metabolism, *via* functional interactions with various transcriptional factors. Mitochondrial biogenesis is controlled by PGC mainly through interactions with nuclear respiratory factors, NRF-1 and NRF-2. In a cell-selective manner, the efficiency of the oxidative phosphorylation process may also be regulated by PGC through the transcriptional control of uncoupling proteins (UCPs) [2].
PGC-1α is expressed in organs and tissues with high energetic needs, such as heart, liver, skeletal muscle, brown adipose tissue, brain and kidney. In contrast, the expression of the PGC-1 related coactivator (PRC) depends on the cell cycle, and plays a role in the integration of pathways directing the mitochondrial respiratory function and cell growth [3]. PRC shares key structural motifs with PGC-1α, interacting with and transactivating the promoters of NRF-1 target genes in a similar manner [4,5]. Other transcription factors targeted by PGC-1α have been studied as potential targets for

PRC. Weak interactions between PRC, on one hand, and PPARγ, TRβ and RAR, on the other, reflect the divergence between PRC and PGC-1α; however, strong interactions between PRC and factors such as NRF-2, ERRα and CREB have recently been identified [6,7]. The nuclear transcription factors involved in mitochondrial biogenesis are not exclusive to this process but contribute to integrating the expression of mitochondrial proteins with other cellular functions. Thus, PRC has been shown to control cell growth through mechanisms that seem to be independent of its effects on mitochondrial function [3]. PRC has also been described as a positive regulator of respiratory chain expression; however, the precise role played by PRC in the complex functions involved in mitochondrial biogenesis and cell growth remains to be elucidated.

Deregulation in the functional status of mitochondria can have a feedback effect on nuclear transcriptional machinery [8,9]. It has been suggested that this retrograde signalling influences the bidirectional communication between the nucleus and mitochondria through a PGC-1 dependent pathway [10]. Mitochondrial functions are regulated by complex mechanisms in which nitric oxide (NO) is a key factor. The mitochondrial effect of NO is bi-phasic depending on its production level. Thus, NO is known to induce the production of reactive oxygen species (ROS) and trigger redox signalling [11]. It directly binds the haem-copper oxidases of complex IV and leads to reversible inhibition of the respiratory chain in an acute response. In a long term effect, a nitric oxide-cGMP-dependent pathway has been shown to control mitochondrial biogenesis through the PGC-1α pathway [12]. The overexpression of NO, cGMP, or eNOS (endothelial nitric oxide synthase) was found to dramatically increase the numbers of mitochondria in a range of cell lines. This overexpression, related to a high production of PGC-1α, leads to the formation of functionally efficient mitochondria in terms of oxidative phosphorylation. Thus, NO may either bind to iron-sulphur loci in proteins such as soluble guanylyl cyclase and to haem-copper oxidases in a time- and concentration-dependent manner. Interestingly, constitutive NO synthase – like eNOS – generates a low concentration of NO, which acts mainly as a second messenger to maintain cell homeostasis [13]. Moreover, eNOS was suspected of being one of the target genes of the retrograde signalling [14]. Since PRC presents the characteristics of an early gene product, it is likely to be involved in the rapid response to retrograde signals [7].

We have demonstrated in thyroid oncocytoma, which constitute a model of mitochondrial-rich tumours, a high rate of PRC induction, with no disruption of mitochondrial functions [15]. The PRC-induced effects were confirmed in the XTC.UC1 thyroid cell line, which thus offered a useful model for investigating the coordination of the nuclear and mitochondrial genomes [16,17]. In these mitochondrial-rich cells, we found an increased expression of eNOS, clearly produced by follicular thyroid cells [18,19]. To further appreciate the importance of the nuclear-mitochondrial crosstalk mediated by PRC, we searched for the cellular programs regulated by PRC in the oxidative XTC.UC1 human oncocytic thyroid cell line, using genome-wide expression profiles of PRC SiRNA *versus* those of negative control cells. We compared the pattern of PRC inhibition with the activation of mitogen-activated protein kinase (MAPK). We investigated the effect of NO on the induction of PRC in the oncocytic cells compared to a non-oxidative BCPAP cell line derived from a papillary thyroid carcinoma. We focussed on the

various functional features of PRC compared to those of PGC-1α, especially with respect to the rapid response required to maintain mitochondrial and cell homeostasis.

Results

PRC-mediated mitochondrial biogenesis depends on the level of nitric oxide production

At T0, the endogenous NO was significantly higher (4.2 fold on average) in XTC.UC1 cells than in BCPAP cells (Figure 1A). Since identical results were obtained with the nitric oxide donor SNAP at concentrations of 50 µM and 100 µM, we considered only those of SNAP at 100 µM, this concentration being likely to produce 100 nM equivalent NO during several hours [20]. From T0 to T72, daily SNAP treatment induced comparable NO concentrations in the two cell lines, with a 1.6 fold increase in XTC.UC1 cells and a 1.7 fold increase in BCPAP cells. At T96, the use of the DAF-2/DA dye as an NO probe showed that the NO level decreased sharply for XTC.UC1 cells. This was associated with a 12-fold increase of ROS production, detected with a dihydroethidium probe, for XTC.UC1 cells at T96, whereas ROS production was stable at T48 and T72 (data not shown). The drastic decrease in the NO level measured by the DAF2 dye at T96 could be associated with the complete inhibition of complex IV activity and to the production of peroxynitrite. Dysfunctions in the respiratory chain could have induced an oxidative stress that, combined with the high NO level, may have produced peroxynitrite. However, contrary to probes measuring oxidative stress, the DAF2 probe was unable to detect peroxynitrite formation [21]. For BCPAP cells, NO production continuously increased at T96 up to 2.2 fold from T0. The profile of nitrotyrosine-modified proteins showed that the 100 µM SNAP treatment induced a significant increase in protein nitration in XTC.UC1 cells from T48 whereas protein nitration had not increased significantly in BCPAP cells at T96 (Figure 1B).

Regarding the mRNA expression of genes involved in mitochondrial biogenesis (NRF-1, ND5 and PRC) and mitochondrial ROS detoxification (SOD2), we showed that the 100 µM SNAP treatment induced nuclear gene expression from T48, whereas mitochondrial gene ND5 expression was significantly induced from T96 in XTC.UC1 cells (Figure 2). No significant induction of genes involved in mitochondrial biogenesis was observed in BCPAP cells up to T96 of SNAP treatment. However, the SOD2 expression level had increased significantly at T96 compared to the T0 level.

Measurement of mitochondrial complex IV activity at T0 and T48 with the 100 µM SNAP treatment, expressed in terms of citrate synthase, showed that complex IV activity decreased significantly in XTC.UC1 cells (T0: 0.32 ± 0.04 nmol/min/mg protein, T48: 0.22 ± 0.02; $P \leq 0.05$) whereas there was no difference in complex IV activity in BCPAP cells (T0: 0.15 ± 0.04 nmol/min/mg protein; T48: 0.18 ± 0.03). The decrease in complex IV activity may be associated with an NO effect interacting with the haem-copper complex and leading to its inhibition.

PRC expression is regulated at the transcriptional level

In XTC.UC1 cells synchronized by serum starvation for two days, we compared the PRC mRNA expression level during 48h of 20% serum and PRC SiRNA and/or SNAP treatment (Figure 3A). The expression profile during 20% serum treatment was considered as a control. After 48h of continuous SNAP treatment, there was a 2.2-fold increase of PRC induction compared to the 48h control. SiRNA treatment reduced

PRC expression by 70% at 48h while SiRNA and SNAP used in combination led to PRC re-expression up to 1.4 times the 48h control value. Moreover, PRC mRNA expression was rapidly re-induced to a significant level after 6h of treatment with SNAP+SiRNA (5.6 ± 0.6 relative PRC mRNA) compared to the level after 6h of treatment with SiRNA alone (3.5 ± 0.4 relative PRC mRNA, $P \leq 0.05$).

The expression of PRC and that of three genes involved in mitochondrial biogenesis (NRF-1, TFAM and COX5B) was significantly dowregulated when the NO/cGMP pathway was inhibited by the protein kinase G (PKG) inhibitor KT5823 at 1 µM (Figure 3B). This showed the involvement of the NO/cGMP pathway in the regulation of PRC expression, which decreased from 2.2 ± 0.1 to 0.8 ± 0.1 relative PRC mRNA with KT5823 treatment ($P \leq 0.05$).

The microarray profile of kinetic PRC SiRNA reveals nuclear-mitochondrial crosstalk

We examined the gene expression pattern of PRC SiRNA and control cells during the 48h SiRNA treatment (Figure 4A). We had previously verified the level of PRC expression during the kinetics of inhibition by comparing PRC expression in SiRNA to a negative control by quantitative RT-PCR. PRC expression was inhibited from 42% (0h and 12h) to 70% at 24h and 74% at 48h. The absence of difference in the level of PRC inhibition between 0h and 12h may be explained by the dependence of PRC expression on cell cycle induction at 12h even though PRC SiRNA had begun to act.

Eight clusters of genes were differentially expressed between PRC SiRNA and the negative control during 48h of 20% serum induction (Figure 4A). Gene ontology identified two clusters, 2 and 8, in which mitochondrial functions were enriched, and two clusters, 5 and 6, in which the cell cycle functions were enriched (Table 1). The comparison of the significant differential gene expression between PRC SiRNA and the negative control, in each cluster and at each time of treatment (Figure 4B), showed that some genes were negatively regulated by PRC at a given time and positively at another. This suggests that PRC regulates gene transcription through interactions with transcriptional factors in either a positive or a negative manner.

We searched for transcriptional factors involved in the PRC-regulating pathway in clusters 2 and 6 where ontological terms were enriched in mitochondrial metabolism and the cell mitotic cycle, respectively. Cluster 2 showed that PRC-positive regulation operated through NRF-1, NRF-2 and ERR1 interactions, whereas PRC-negative regulation acted mainly through SREBP1, CREB1 and MAZ interactions (Figure 5A). Cluster 6 showed that PRC-positive regulation depended on SP1, CREB1 and YY1 interactions whereas PRC-negative regulation depended on NRF-2 and MYC interactions (Figure 5B). The frequency of these motifs, ranging from 22-49% in the two gene clusters, was relevant of their involvement in the regulation of these genes compared to the relatively low frequency (2-15%) for the other genes of the microarray.

Oxidative phosphorylation is finely regulated by PRC

Functional analysis of the respiratory chain complexes showed that complexes I and III were rapidly down-regulated when PRC inhibition was effective (T24 of SiRNA treatment referred to T12). This was associated with increased complex IV activity and better coupling between respiration and ATP synthesis (Figure 6A). At T48, the activity of each of the four complexes decreased; although the coupling

was more efficient compared to that at T24, the ATP level was significantly lower at T48 than at T12.

Regarding the expression of selected genes coding for complex IV subunits, the mitochondrial-encoded gene COII was down-regulated at T24, together with a decrease in mtDNA copy number (Figure 6B). The expression of nuclear-encoded complex IV subunit genes was more variable. There was a large variation in the expression of COX15 during PRC inhibition whereas the expression of COX5B decreased proportionally to that of COII. The expression of COX15 at T24 increased by 1.8 fold compared to that at T12, followed by a sharp decrease to 8% of the control value at T48. This may have been related to the great difference in complex IV activity observed between T24 and T48. Taken together, these results showed that complexes I, III and IV were primarily regulated by PRC, in particular with regard to the expression of COX15, which may be associated with the increased activity of the complex IV in the first 12h of PRC invalidation. Although these changes may have caused a functional default in the respiratory chain inducing oxidative stress, the expression of SOD2, a gene coding for a protein involved in superoxide detoxification, had not significantly decreased even at T48. However, the extension of the SiRNA treatment beyond T48 led to significantly down-regulated results (data not shown). The protein expression of several subunits from each OXPHOS complex decreased significantly at T48 for the three complexes I, II and IV, but the decrease was not significant for complexes III and V (Figure 6C). However, the gene and protein expressions of the different OXPHOS complexes were correlated so that global down-regulation was observed with PRC inhibition at T48 compared to T12.

PRC regulates phosphatase/kinase activity

A phosphospecific antibody microarray for the MAPK signalling pathway was set up to compare the phosphorylation status of direct as well as indirect MAPK targets so as to determine whether PRC was required for the phosphorylation of MAPK targets. This antibody array included 93 highly specific and well-characterized phosphospecific antibodies for proteins in the MAPK pathway, each with six replicates (the raw data are shown in Supplementary Table S1). The paired antibodies for the same, but unphosphorylated, target sites were also included in the array to allow determination of the relative level of phosphorylation. Using a cutoff ratio of 0.8, we identified 11 sites that were hypophosphorylated in PRC SiRNA cells at T48 compared with control cells. These sites were involved in the regulation of the activity of seven proteins: ASK1 (P83 and P966), TP53 (P18 and P46), ATF-2 (P69-51), c-JUN (P63) SAP/JNK (P183), ELK-1 (P383 and P389), ESR1 (P167) and Stahmin-1 (P15).

Discussion

PRC, a member of the family of PGC1 transcriptional coactivators, is expressed more abundantly in proliferative cells than in growth-arrested cells. Several studies have demonstrated the important modulatory role played by this family of coactivators in maintaining optimal tissue energetic needs [22]. In the mitochondrial-rich tumour model we investigated, PGC-1α expression remained very low, being induced neither by serum nor in compensation to PRC repression.

We showed that the nuclear and mitochondrial genes of the respiratory chain were up-regulated in accordance with PRC induction in a time-dependent manner. We also observed an induction of SOD2

expression that had been previously shown to be regulated by PGC-1α [23]. The rapid action of the PKG inhibitor on both PRC expression and the nuclear genes controlling respiratory chain biosyntheses showed that mitochondrial biogenesis required the induction of the NO/PRC pathway in the XTC.UC1 cell model. The increase in PRC expression observed with combined PRC invalidation and SNAP treatment may be explained by the role of PRC in the rapid response to change in the status of the cell cycle [7]. Unlike the case of PGC-1α, an activating process by phosphorylation has not so far been described for PRC. Here we observed that an activating process by nitration may rapidly modulate PRC expression. This mechanism has also been described in the regulation of several transcription factors, such as CREB, c-Myc, Jun or Fos [24,25].

In XTC.UC1 cells, as well as in mitochondrial-rich tumours, we previously reported that the high cellular level of NO was related to the induction of eNOS synthesis in follicular thyroid cells [26]. The activation of eNOS by the PI3K-AKT pathway has been recently implicated in tumour maintenance [27]. This finding may be particularly relevant in the context of mitochondrial-rich tumours characterized by the absence of mutations usually described in thyroid tumours of follicular origin [28]. In contrast, BCPAP cells were unable to activate the PRC pathway, even when NO increased through SNAP treatment. Independently of the mitochondrial quantity and the specific growth media, there was a great difference in the mitochondrial expression profile and functions: while XTC.UC1 cells displayed an oxidative metabolism, BCPAP cells were mainly engaged in a glycolytic process [19]. This suggested that defects of the mitochondrial function in BCPAP cells may have induced a retrograde pathway modifying the status of cell phosphorylation [29]. Thus, the action of PRC in controlling the redox status should be taken into account in relation to cell transformation and cancer.

Our microarray data on PRC inhibition may be compared to a recent study using a PRC ShRNA approach [3]. The use of SiRNA and ShRNA allowed us to distinguish between the different temporal roles played by PRC, with SiRNA revealing the short term effects and ShRNA indicating the chronic action of PRC inhibition. However in both studies, a 50% decrease of PRC mRNA was sufficient to demonstrate the specific regulatory functions of this coactivator with complete inhibition of its protein expression. The effects on the OXPHOS process were severe probably because of the lack of functional compensation by other members of this family of coactivators, as suspected following studies on PGC-1α and PGC-1β null mice [30,31]. Contrary to reports on ShRNA PRC knockdown [3], we identified a decrease in mtDNA content that may be primarily related to the decrease in OXPHOS activity. We postulate that long term PRC inhibition by ShRNA could induce compensatory mechanisms to maintain mitochondrial functions used for other cellular metabolisms. Using SiRNA, we have defined the positive as well as the negative regulations mediated by PRC through specific transcription factors. These may represent the direct and indirect transcriptional effects of the PRC coactivator. Thus, the PRC factor should be seen as a physiological integrator of energetic metabolism in other biological processes.

Focusing on clusters ontologies, we showed that negative PRC regulation could be exerted on cytoskeleton and organelle

biogenesis. This should be compared with the surprising regulation of mitochondrial complex IV activity during SiRNA treatment. Indeed, the analysis of the expression of some mitochondrial and nuclear-encoded genes of complex IV showed that COX15 expression varied in a manner opposite to that of other genes. This differential expression identified on microarray data was confirmed by real-time quantitative PRC analysis. The role of COX15 in the assembly of human complex IV has recently been emphasized [32]. COX15 was also found to be an essential component of the catalytic centre of the complex. This PRC regulation seems to be independent of a direct NO effect on the haem-copper of complex IV as there is a correlation between the activity of the complex and the expression of COX15 mRNA. We postulate that PRC ensures a rapid adaptation to the cellular environment through the regulation of a few target genes such as COX15. It separately regulates the expression of OXPHOS subunits, leading to a rapid modulation of mitochondrial functionality by integrating the regulation of complex IV activity with coupling efficiency and ROS detoxification processes.

We associated PRC functions with some PGC-1α regulated functions, such as chromatin remodelling, RNA splicing, translation and angiogenesis [33,34]. Regulation of the cell cycle by PRC could be compared to the regulation of PGC-1α and PGC-1β by nutritional and hormonal signals as well as by circadian pacemakers [35]. As has recently been suggested, the rapid regulation of complexes I and III, as well as the specific modulation of complex IV activity, by the ubiquitous PRC member could support the basal energetic cellular needs when cycling progression from the G1 phase [36]. The effects of PRC on the M-phase of the cell cycle have not been described so far. We suggest that PRC coordinates cell-cycle phases with mitochondrial metabolism through both positive and negative interactions with CREB1 and NRF-2 transcription factors. Previous studies have clearly associated CREB1 and NRF-2 with the regulation of genes involved in energetic metabolism as well as the cell cycle [37,38]. Even though these results are supported by the frequency of representation for motifs associated with these transcription factors in specific clusters, they need to be confirmed by results obtained with the co-immunoprecipitation technique.

Some pathways, such as those of the phosphatase/kinase activities and vesicle organization, also seem to be specifically related to PRC [3]. Interestingly, the MAPK profile of PRC-inhibited cells showed a significant decline in the phosphorylation level of proteins involved in the ASK1 signalling pathway. ASK1, a serine/threonine protein kinase, is activated by the exposure of cells to various stimuli, including tumour necrosis factor-α, Fas ligand and hydrogen peroxide [39]. It lies upstream of a major redox-sensitive pathway leading not only to the induction of apoptosis but also to cell proliferation and differentiation [40]. The majority of proteins involved in the control of the cell cycle have been shown to be redox-sensitive with functions in the G1 and G2/M phases [41]. Since the G2/M phase is in a more reduced state than the G1 phase, oxidant-sensitive proteins may be temporally regulated by the oscillation of the intracellular redox environment [42]. We postulate that PRC may represent the redox sensor enabling the initiation of the retrograde signalling pathway, especially in a cellular model presenting a balance between the production of nitric oxide and superoxide. Interestingly, eNOS activity has also been shown to be

redox-sensitive [43]. Further studies will be needed to clarify the temporal role of PRC in the integrative regulation of cell metabolism.
In conclusion, we have identified several new pathways regulated by the PRC coactivator. Focusing on OXPHOS, we showed that the PRC-regulation of this process differed from that of PGC-1α. The nuclear OXPHOS genes were tightly controlled by PGC-1α but less so by PRC. Nevertheless, other genes such as COX15 were found to be more specifically regulated by PRC than by PGC-1α. This precision control of mitochondrial energy metabolism should be placed in the context of the complex regulation of the cell cycle. Here we observed that the eNOS/PRC signalling pathway is able to modulate the cell cycle by regulating the intracellular redox status. The role of PRC acting in complement with other PGC-1 factors should therefore be fully explored, especially in the case of metabolic diseases.

Materiel and Methods

Cell lines and SNAP treatment

Two human thyroid cell lines, XTC.UC1 and B-CPAP, were used. The XTC.UC1 cell line was established from an oncocytic cell thyroid carcinoma. The growth medium consisted of Dulbecco's modified Eagle medium (DMEM) supplemented with 10% foetal calf serum (Seromed, Biochrom AG, Berlin, Germany), 100 U/mL penicillin, 100 mg/mL streptomycin, 0.25 mg/mL fungizone, and 10 mU/mL thyrotropin (Paez, Lin et al.) (Sigma-Aldrich, Saint Louis, MO, USA). The B-CPAP cell line was established from a human papillary thyroid carcinoma cultured in RPMI-1640 medium with 10% foetal calf serum, 100 U/mL penicillin, 100 mg/mL streptomycin, 0.25 mg/mL fungizone. Except for the TSH and the foetal calf serum, all the products were obtained from Gibco BRL (Life Technologies, Paisley, United Kingdom).
The two cell lines were treated once a day for four days with the nitric oxide donor SNAP (S-nitroso-N-acetyl-D,L-penicillamine, EMD, San Diego, CA, USA) at a final concentration of either 50 or 100 µM in the selected medium. Except for the microarray analysis, which was performed in quadruplicate (N=4), all the assays were performed in quintuplicate (N=5). To control the mediating effect of NO/cGMP on PRC expression, cells were pre-treated with 1 µM of the protein kinase G inhibitor KT5823 (EMD, San Diego, CA, USA) 30 minutes before SNAP treatment (100 µM) during 24h.

Flow cytometry analysis

XTC.UC1 cells were exposed at 37°C for 20 minutes to 5 µg/mL JC-1 (5,5',6,6'-tetrachloro-1,1',3,3'-tetraethylbenzimidazolcarbocyanine) or for 30 minutes at pH 7 to 10 µM DAF-2/DA (4,5-diaminofluorescein diacetate) before trypsin treatment. The two dyes were purchased from EMD, San Diego, CA, USA. The cells were then harvested, washed twice with PBS, and analyzed on a FACScan flow cytometer (Becton Dickinson, Franklin Lakes, NJ, USA). Cultured cells, untreated with dyes, were used as a negative control to determine the FACS gating. JC-1 emits red fluorescence (aggregates) when sequestered in the mitochondrial membrane of healthy cells and emits green fluorescence (monomers) when released into the cytoplasmic compartment of the cell. At the depolarized membrane potential (–100 mV), the JC-1 green monomer emission peaks at about 527 nm. At the hyperpolarized membrane potential (-140 mV) the red emission of the JC-1 aggregates shifts towards 590 nm. The red/green ratio represents the modification of mitochondrial

functionality during treatment. DAF-2/DA is a cell-permeable derivative of DAF-2 presenting an emission peak at about 515 nm, and which reacts neither with stable oxidized forms of NO, such as NO_2^- and NO_3^-, nor with other reactive oxygen species, such as $O_2^{\cdot-}$, H_2O_2, and $ONOO^-$ [44].

Quantitative PCR analysis

MtDNA and cDNA quantifications were performed on a Chromo4 apparatus (Bio-Rad, Hercules, CA, USA) using SYBR Green I dye as a fluorescent signal (iQ sybrGreen supermix, Bio-Rad), according to the manufacturer's recommendations. For mtDNA quantification, primers were located at the following nucleotide positions: forward, 3254–3277 (in the ND1 gene), and reverse, 3412–3391 (in the $tRNA^{Leu}$ gene). The mtDNA copy number was expressed in terms of the copy number of a 110-bp fragment of the β-globin gene. For cDNA quantification, the expression of nine selected genes (seven nuclear genes: PRC, PGC-1α, NRF-1, COX5B, COX15, TFAM, and SOD2, and two mitochondrial genes, COII, ND5) was measured by real-time quantitative RT-PCR using the protocol previously described [45] and the primers listed in Table 2.

Respiratory chain complex activities and cellular ATP

Respiratory chain complex activities were measured in cell lysates, thermostatically maintained at 37 °C, using a Beckman spectrophotometer (DU800, Beckman Coulter, Fullerton, CA, USA). Cells were resuspended in a cell buffer containing 250 mM saccharose, 20 mM Tris, 2 mM EDTA, 1 mg/ml BSA, pH 7.2 (50 µl/10^6 cells).

To measure the activity of respiratory complexes I and II, the cells were first disrupted by freezing in liquid nitrogen, followed by rapid thawing at 37 °C. Complex I activity was immediately assayed on cell lysate (0.5×10^6 cells) in a KH_2PO_4 buffer (100 mM, pH 7.4), containing 1 mM KCN, 2 mM NaN_3, and 0.1 mM ubiquinone-1. After incubation for 5 min, the reaction was started by adding 0.3 mM NADH and the rate of disappearance of NADH was monitored at 340 nm. Rotenone (5 µM) was then added to determine the background rate. Complex II (succinate ubiquinone reductase) activity was measured as described elsewhere [46], except that the background rate was measured by the addition of thenoyltrifluoroacetone (200 µM). Complex III (after a freeze-thaw cycle), complex IV and citrate synthase (Miller, Long et al.) activities were assayed as described previously [47].

Twenty µl of CellTiter-Glo reagent (Promega, Madison, CA, USA) were added to 3.10^5 cells, growing in 200 µl of culture medium, to measure the cellular ATP level. An ATP curve was prepared as recommended by the manufacturers. Plates were agitated for 2 min and incubated for 10 min at room temperature before measuring the luminescence (Lumat 9507, Berthold Technologies, Bad Wildbad, Germany). The ATP level was expressed as ng per mg of total cell protein measured by the BCA method (Pierce, Rockford, IL, USA).

PRC SiRNA

To knock down PRC expression in the XTC.UC1 cells, three predesigned PRC SiRNAs (from Applied Biosystems, Foster city, CA, USA) were tested in comparison to a random negative control SiRNA (scramble, #4635). The PRC SiRNA (#121729) was chosen for at least 70% of PRC mRNA expression knockdown. For the study, 30 nM of this PRC siRNA were transfected in the XTC.UC1 cells using siPORT NeoFX (Applied Biosystems) as recommended by

the manufacturer. To synchronize the cell cycle for SNAP and/or PRC SiRNA treatments, the cells were serum-starved for two days before adding SiRNA during a 48h period. After 6h, 24h and 48h serum induction, XTC.UC1 cells were harvested for RNA or phospho-protein isolation. To synchronize the cell cycle for genome-wide expression analysis, the cells were serum-starved during 7h before replacing the medium with 20% SVF medium for 48 h. After 12h, 24h and 48h serum induction corresponding to 19h, 31h and 55h SiRNA treatment, XTC.UC1 cells were harvested for protein, DNA and RNA isolation. We refer to the times of 20% serum induction (T0, T12, T24 and T48) instead of the times of SiRNA treatment (T7, T19, T31 and T55) in the figures and the text.

The PRC expression level, analyzed by quantitative RT-PCR, normalized by β-actin expression, was considered relevant when PRC mRNA inhibition reached 70% of that of the negative control (scramble) with a significant effect on COX5B, the PRC-regulated gene.

Microarray analysis

cDNA microarray slides were prepared at the Transcriptome Core Facility of the University of Nantes (INSERM U915, France), using a set of 20,000 oligonucleotides with full functional characterization and content referencing (Ocimum Biosolutions, Hyderabad, India). RNA amplification, cDNA labelling and hybridization from XTC.UC1 cells were performed using the protocols recommended by the Transcriptome Core Facility (http://cardioserve.nantes.inserm.fr/ptf-puce/). Slides were analyzed with Axon GenePix 4.0 software (Axon, Union City, CA, USA) after scanning on a ScanArray Express II scanner (Packard Bioscience, Billerica, MA, USA). Data are available in the GEO database (GSE 14282). The hierarchical clustering of the genes was computed on median-gene-centred and log-transformed data using average linkage and uncentred correlation distances. Computations and visualization were performed using Cluster and TreeView programs [48]. EASE and Gene Set analyses were used to determine the statistical likelihood that *a priori* families of genes were over-represented and differentially expressed [49]. Differentially expressed genes were defined as those well detected in both groups and with absolute PLS (partial least-squares) loadings >3.0 and t-test values of $P<0.0002$ (FDR=0.001). In addition, genes with an absolute value of mean expression differences < 40% were excluded from further analyses. Gene ontology enrichments in gene lists were searched for using GOMiner [50]. The most abundant gene ontology terms, representing at least 5% of the genes in the lists, with P-values lower than 0.05, were considered for interpretation.

The over-representation of transcription-factor binding sites (TFBS) in promoter sequences was investigated with MSignDB software (GSEA, MIT, MI, USA). We added two more position-weight matrices to this collection: for the NRF1 transcriptional factor, we aligned 9 sequences of known NRF-1 binding sites [51]; for the oestrogen-related receptor alpha (ERR1), the position-weight matrix used was that described by Sladek *et al.* [52]. The percentage of representation of a given transcription factor was compared to its representation in the rest of the microarray.

Western blot analysis

Cells were mixed with a buffer containing 10 mmol/L HEPES (pH 7.9), 1.5 mmol/L $MgCl_2$, 10 mmol/L KCl, 0.5% NP-40 and a protease inhibitor cocktail, and then centrifuged at 13,000 g for 5 min. Supernatants were collected and the protein

concentration was determined by the BCA method (Pierce, Rockford, IL, USA). Samples containing 30 µg total protein were applied to 12% SDS-PAGE gel electrophoresis and hybridized with a cocktail of five antibodies against human nuclear-encoded subunits of the OXPHOS process: Complex I (NDUFB8, 20 KDa), complex II (Ip subunit, 30 KDa); complex III (Core2, 40 KDa); complex IV (COX2, 24 KDa); complex V (ATPsynthase F1α, 58 KDa), at a dilution of 1/500 (MS 601, MitoSciences, Eugene, OR, USA). The level of nitration-modified protein was measured using a high-activity rabbit polyclonal anti-nitrotyrosine antibody at a dilution of 1/250 (US Biological, Swampscott, MA, USA).
After overnight incubation with primary antibodies, membranes were washed before incubation with the corresponding horseradish peroxidase (HRP)-conjugated secondary antibody, which was detected with a chemiluminescent detection system (ECL Plus, Amersham Biosciences, Fairfield, CT, USA).

Phospho-Specific Protein Microarray Analysis

The phospho-specific protein microarray was obtained from Full Moon Biosystems (Sunnyvale, CA, USA). Protein microarray analysis was carried out using the protocol provided by the manufacturer. Briefly, 100 µg of cell lysate in 50 µl of reaction mixture were labelled with 1.43 µl biotin in 10 µg/µl N,N-dimethyformamide. The resulting biotin-labelled proteins were diluted 1:20 in a coupling solution before being applied to the array for conjugation. To prepare the antibody microarray, it was first blocked with a blocking solution for 30 min at room temperature, rinsed with Milli-Q grade water for three minutes, and then dried by centrifugation. Finally, the array was incubated with the biotin-labelled cell lysates at room temperature during four hours. After the array slide was washed thrice with 50 ml of 1x wash solution for 10 min each, the conjugated-labelled protein was detected using Cy3-streptavidin.

Data Analysis

Except for the microarray analyses, data are represented as mean values ± SEM, with *N* representing the number of experiments. Statistical analyses were performed by a one-way analysis of variance, the Mann-Whitney *U* test, and Tukey's HSD test. Differences were considered to be statistically significant at $P \leq 0.05$

Footnotes

We thank Dominique Couturier and Naïg Gueguen for technical help. We thank Marja Steenman and Kanaya Malkani for the critical reading of this paper. This work was supported by grants from the French Ministry of Research, the *Institut National de la Recherche Médicale* (INSERM), the University Hospital of Angers and the *Ligue Contre le Cancer*.

References

1. Puigserver P, Wu Z, Park CW, Graves R, Wright M, et al. (1998) A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. Cell 92: 829-839.

2. Wu Z, Puigserver P, Andersson U, Zhang C, Adelmant G, et al. (1999) Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic coactivator PGC-1. Cell 98: 115-124.

3. Vercauteren K, Gleyzer N, Scarpulla RC (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: 2307-2319.

4. Andersson U, Scarpulla RC (2001) Pgc-1-related coactivator, a novel, serum-inducible coactivator of nuclear respiratory factor 1-dependent transcription in mammalian cells. Mol Cell Biol 21: 3738-3749.

5. Gleyzer N, Vercauteren K, Scarpulla RC (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: 1354-1366.

6. Vercauteren K, Gleyzer N, Scarpulla RC (2008) PGC-1-related coactivator complexes with HCF-1 and NRF-2beta in mediating NRF-2(GABP)-dependent respiratory gene expression. J Biol Chem 283: 12102-12111.

7. Vercauteren K, Pasko RA, Gleyzer N, Marino VM, Scarpulla RC (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: 7409-7419.

8. Biswas G, Adebanjo OA, Freedman BD, Anandatheerthavarada HK, Vijayasarathy C, et al. (1999) Retrograde Ca2+ signaling in C2C12 skeletal myocytes in response to mitochondrial genetic and metabolic stress: a novel mode of inter-organelle crosstalk. EMBO J 18: 522-533.

9. Butow RA, Avadhani NG (2004) Mitochondrial signaling: the retrograde response. Mol Cell 14: 1-15.

10. Wu H, Kanatous SB, Thurmond FA, Gallardo T, Isotani E, et al. (2002) Regulation of mitochondrial biogenesis in skeletal muscle by CaMK. Science 296: 349-352.

11. Brookes PS, Shiva S, Patel RP, rley-Usmar VM (2002) Measurement of mitochondrial respiratory thresholds and the control of respiration by nitric oxide. Methods Enzymol 359: 305-319.

12. Nisoli E, Clementi E, Paolucci C, Cozzi V, Tonello C, et al. (2003) Mitochondrial biogenesis in mammals: the role of endogenous nitric oxide. Science 299: 896-899.

13. Liaudet L, Soriano FG, Szabo C (2000) Biology of nitric oxide signaling. Crit Care Med 28: N37-N52.

14. Nisoli E, Carruba MO (2006) Nitric oxide and mitochondrial biogenesis. J Cell Sci 119: 2855-2862.

15. Savagner F, Mirebeau D, Jacques C, Guyetant S, Morgan C, et al. (2003) PGC-1-related coactivator and targets are upregulated in thyroid oncocytoma. Biochem Biophys Res Commun 310: 779-784.

16. Savagner F, Franc B, Guyetant S, Rodien P, Reynier P, et al. (2001) Defective mitochondrial ATP synthesis in oxyphilic thyroid tumors. J Clin Endocrinol Metab 86: 4920-4925.

17. Savagner F, Chevrollier A, Loiseau D, Morgan C, Reynier P, et al. (2001) Mitochondrial activity in XTC.UC1 cells derived from thyroid oncocytoma. Thyroid 11: 327-333.

18. Patel A, Fenton C, Terrell R, Powers PA, Dinauer C, et al. (2002) Nitrotyrosine, inducible nitric oxide synthase (iNOS), and endothelial nitric oxide synthase (eNOS) are increased in thyroid tumors from children and adolescents. J Endocrinol Invest 25: 675-683.

19. Baris O, Savagner F, Nasser V, Loriod B, Granjeaud S, et al. (2004) Transcriptional profiling reveals coordinated up-regulation of oxidative metabolism genes in thyroid oncocytic tumors. J Clin Endocrinol Metab 89: 994-1005.

20. Poderoso JJ, Peralta JG, Lisdero CL, Carreras MC, Radisic M, et al. (1998) Nitric oxide regulates oxygen uptake and hydrogen peroxide release by the isolated beating rat heart. Am J Physiol 274: C112-C119.

21. Wardman P (2007) Fluorescent and luminescent probes for measurement of oxidative and nitrosative species in cells and tissues: progress, pitfalls, and prospects. Free Radic Biol Med 43: 995-1022.

22. Scarpulla RC (2008) Transcriptional paradigms in mammalian mitochondrial biogenesis and function. Physiol Rev 88: 611-638.

23. Borniquel S, Valle I, Cadenas S, Lamas S, Monsalve M (2006) Nitric oxide regulates mitochondrial oxidative stress protection via the transcriptional coactivator PGC-1alpha. FASEB J 20: 1889-1891.

24. Contestabile A (2008) Regulation of transcription factors by nitric oxide in neurons and in neural-derived tumor cells. Prog Neurobiol 84: 317-328.

25. Carreras MC, Poderoso JJ (2007) Mitochondrial nitric oxide in the signaling of cell integrated responses. Am J Physiol Cell Physiol 292: C1569-C1580.

26. Baris O, Mirebeau-Prunier D, Savagner F, Rodien P, Ballester B, et al. (2005) Gene profiling reveals specific oncogenic mechanisms and signaling pathways in oncocytic and papillary thyroid carcinoma. Oncogene 24: 4155-4161.

27. Lim KH, Ancrile BB, Kashatus DF, Counter CM (2008) Tumour maintenance is mediated by eNOS. Nature 452: 646-649.

28. Nikiforova MN, Lynch RA, Biddinger PW, Alexander EK, Dorn GW, et al. (2003) RAS point mutations and PAX8-PPAR gamma rearrangement in thyroid tumors: evidence for distinct molecular pathways in thyroid follicular carcinoma. J Clin Endocrinol Metab 88: 2318-2326.

29. Scarpulla RC (2006) Nuclear control of respiratory gene expression in mammalian cells. J Cell Biochem 97: 673-683.

30. Lin J, Wu PH, Tarr PT, Lindenberg KS, St-Pierre J, et al. (2004) Defects in adaptive energy metabolism with CNS-linked hyperactivity in PGC-1alpha null mice. Cell 119: 121-135.

31. Lelliott CJ, Medina-Gomez G, Petrovic N, Kis A, Feldmann HM, et al. (2006) Ablation of PGC-1beta results in defective mitochondrial activity, thermogenesis, hepatic function, and cardiac performance. PLoS Biol 4: e369.

32. Fernandez-Vizarra E, Tiranti V, Zeviani M (2008) Assembly of the oxidative phosphorylation system in humans: What we have learned by studying its defects. Biochim Biophys Acta

33. Arany Z, Foo SY, Ma Y, Ruas JL, Bommi-Reddy A, et al. (2008) HIF-independent regulation of VEGF and angiogenesis by the transcriptional coactivator PGC-1alpha. Nature 451: 1008-1012.

34. Lin JD (2009) Minireview: the PGC-1 coactivator networks: chromatin-remodeling and mitochondrial energy metabolism. Mol Endocrinol 23: 2-10.

35. Cam H, Balciunaite E, Blais A, Spektor A, Scarpulla RC, et al. (2004) A common set of gene regulatory networks links metabolism and growth inhibition. Mol Cell 16: 399-411.

36. Schieke SM, McCoy JP, Jr., Finkel T (2008) Coordination of mitochondrial bioenergetics

with G1 phase cell cycle progression. Cell Cycle 7: 1782-1787.

37. Barlow CA, Shukla A, Mossman BT, Lounsbury KM (2006) Oxidant-mediated cAMP response element binding protein activation: calcium regulation and role in apoptosis of lung epithelial cells. Am J Respir Cell Mol Biol 34: 7-14.

38. Crook MF, Olive M, Xue HH, Langenickel TH, Boehm M, et al. (2008) GA-binding protein regulates KIS gene expression, cell migration, and cell cycle progression. FASEB J 22: 225-235.

39. Chen Z, Seimiya H, Naito M, Mashima T, Kizaki A, et al. (1999) ASK1 mediates apoptotic cell death induced by genotoxic stress. Oncogene 18: 173-180.

40. Sayama K, Hanakawa Y, Shirakata Y, Yamasaki K, Sawada Y, et al. (2001) Apoptosis signal-regulating kinase 1 (ASK1) is an intracellular inducer of keratinocyte differentiation. J Biol Chem 276: 999-1004.

41. Conour JE, Graham WV, Gaskins HR (2004) A combined in vitro/bioinformatic investigation of redox regulatory mechanisms governing cell cycle progression. Physiol Genomics 18: 196-205.

42. Yamamoto K, Ichijo H, Korsmeyer SJ (1999) BCL-2 is phosphorylated and inactivated by an ASK1/Jun N-terminal protein kinase pathway normally activated at G(2)/M. Mol Cell Biol 19: 8469-8478.

43. Tanaka T, Nakamura H, Yodoi J, Bloom ET (2005) Redox regulation of the signaling pathways leading to eNOS phosphorylation. Free Radic Biol Med 38: 1231-1242.

44. Kojima H, Nakatsubo N, Kikuchi K, Kawahara S, Kirino Y, et al. (1998) Detection and imaging of nitric oxide with novel fluorescent indicators: diaminofluoresceins. Anal Chem 70: 2446-2453.

45. Fontaine JF, Mirebeau-Prunier D, Franc B, Triau S, Rodien P, et al. (2008) Microarray analysis refines classification of non-medullary thyroid tumours of uncertain malignancy. Oncogene 27: 2228-2236.

46. James AM, Wei YH, Pang CY, Murphy MP (1996) Altered mitochondrial function in fibroblasts containing MELAS or MERRF mitochondrial DNA mutation. Biochem J 318 (Pt 2): 401-407.

47. Rustin P, Munnich A, Rotig A (2004) Mitochondrial respiratory chain dysfunction caused by coenzyme Q deficiency. Methods Enzymol 382: 81-88.

48. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.

49. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116-5121.

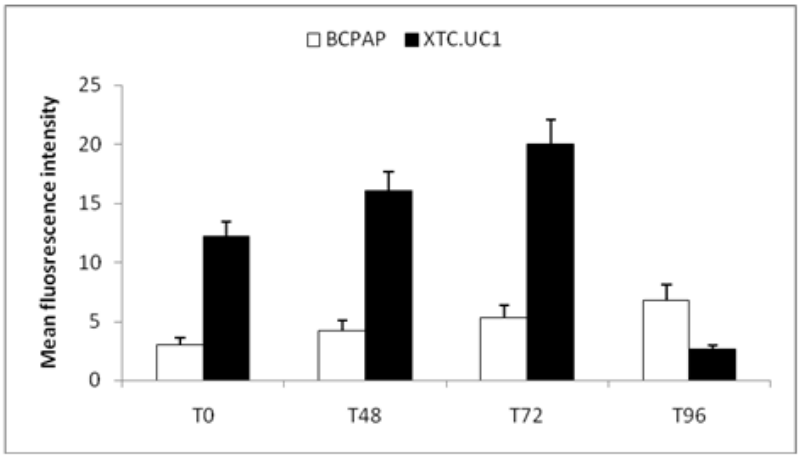
50. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4: R28.

51. Au HC, Scheffler IE (1998) Promoter analysis of the human succinate dehydrogenase iron-protein gene--both nuclear respiratory factors NRF-1 and NRF-2 are required. Eur J Biochem 251: 164-174.

52. Sladek R, Bader JA, Giguere V (1997) The orphan nuclear receptor estrogen-related receptor alpha is a transcriptional regulator of the human medium-chain acyl coenzyme A dehydrogenase gene. Mol Cell Biol 17: 5400-5409.

Figures

1A



1B

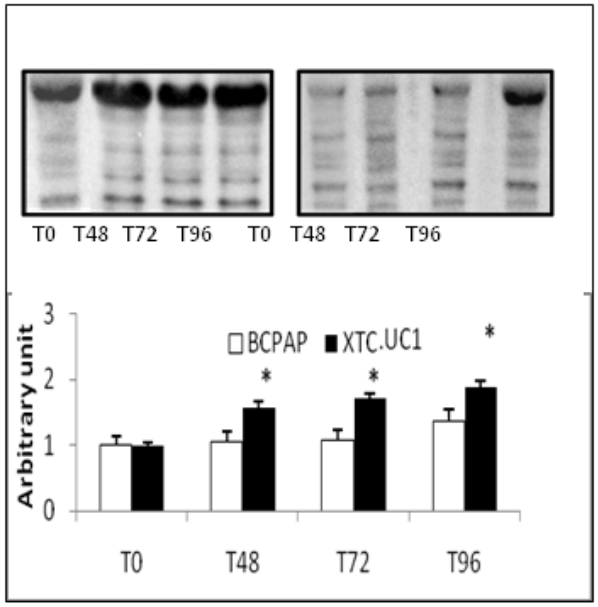


Figure 1: Effect of 100 µM SNAP treatment during 96h on XTC.UC1 and BCPAP cell lines (*N*=5 per cell line).
1A: Nitric oxide measurement with the FACScan cytometer (10 µM final DAF2/DA dye). Results are expressed in arbitrary fluorescent units as mean values ± SEM. *$P \leq 0.05$ *versus* T0. **$P \leq 0.05$ *versus* T0 and T72.
1B: Evaluation of protein nitration by Western blotting using antibody raised against nitrotyrosine-modified proteins. Immunoblots are quantified by densitometric analysis and expressed in arbitrary units (relative to α-tubulin) as mean values ± SEM. *$P \leq 0.05$ *versus* T0.

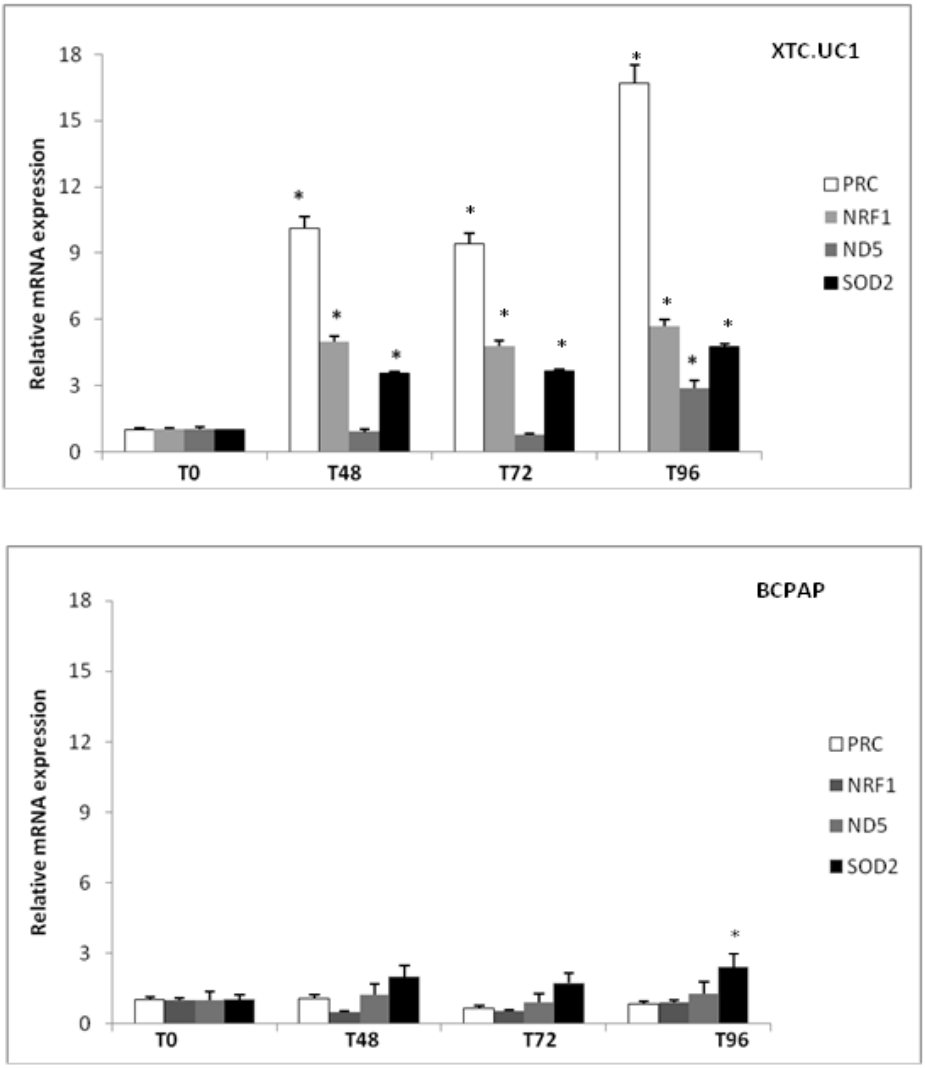
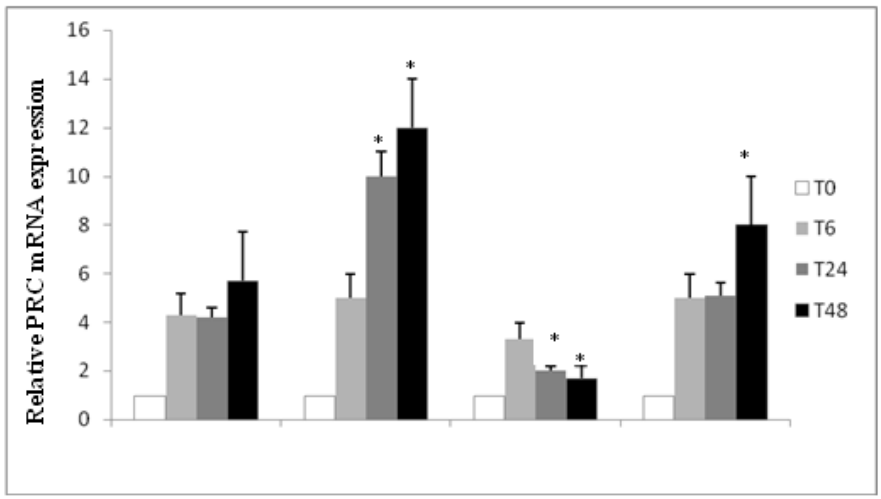


Figure 2: The expression of nuclear genes (PRC, NRF-1 and SOD2) and the mitochondrial gene (ND5) after 100 µM SNAP treatment of XTC.UC1 and BCPAP cells during 96h. Data are expressed in relative units (mRNA copies number of a specific gene/copy number of β-globin mRNA) and expressed in terms of the T0 ratio as mean values ± SEM. *$P \leq 0.05$ *versus* T0. *N*=5 per cell line and time of treatment.

3A



20 % Serum	+	+	+	+
PRC SiRNA (30 µM)	-	-	+	+
SNAP (100 µM)	-	+	-	+

3B

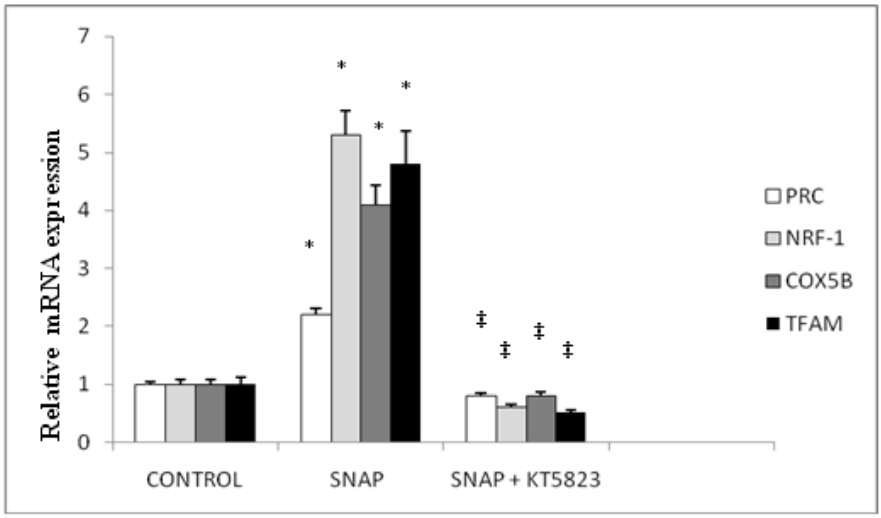


Figure 3: Effects of mRNA expression in XTC.UC1 cells during 20% serum induction and 100 µM SNAP and/or 30 nM PRC SiRNA treatment. Relative units (PRC mRNA copy number/β-globin copy number) expressed in terms of the T0 ratio as mean values ± SEM. *N*=5 per time of treatment.
3A: The variation of PRC mRNA during 48h of the different treatments. Except for T48 SiRNA, all samples differed significantly from T0 ($P \leq 0.05$). *$P \leq 0.05$ *versus* T6.

3B: The variation of NRF-1, TFAM and COX5B mRNA at T24 of treatments compared to control T0. All samples differed significantly from T0 (*, $P \leq 0.05$) or from those with 100 µM SNAP treatment (‡, $P \leq 0.05$)

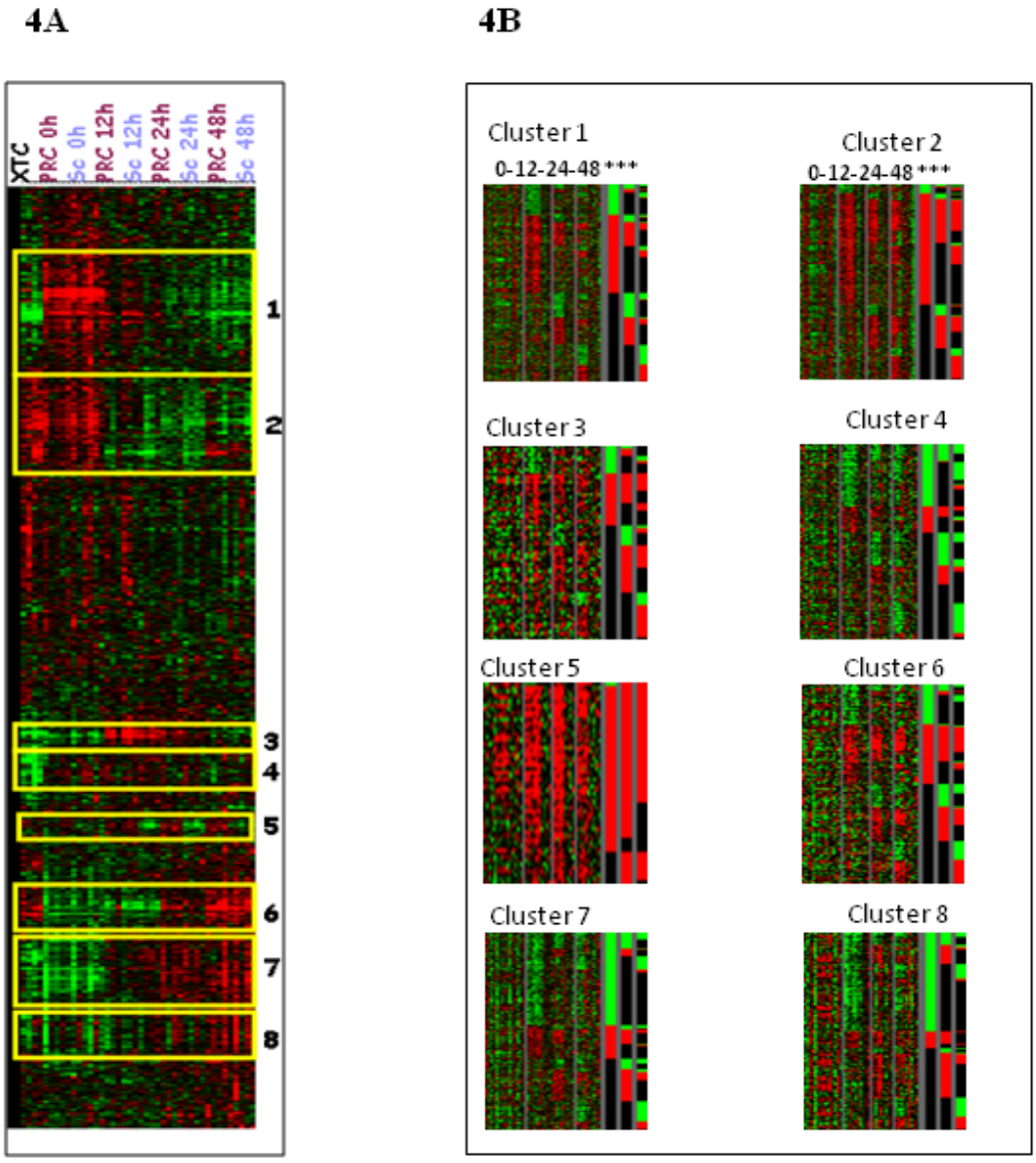


Figure 4: Microarray analysis of the XTC.UC1 cell line after transfection with either PRC SiRNA (PRC) or scramble as a negative control (Sc). The expression profiles were obtained at T12, T24 and T48 after SiRNA transfection and normalized to serum or scramble values (N=4 per time). The XTC sample represents the gene expression profile of the cell line before transfection.

4A: Global clustering of PRC SiRNA and negative control (scramble) samples normalized by serum induction. Eight clusters were selected on significant differential genes for all time of SiRNA treatment ($P \leq 0.03$).

4B: Significant over- and under-expressed genes in SiRNA samples compared to negative control (scramble) samples for each of the 8 clusters previously identified by serum centering. At the right side of each cluster (***) are represented the genes differentially expressed (over-expressed in red, and under-expressed in green) at T12, T24 and T48 ($P \leq 0.03$) in PRC SiRNA compared to matched negative controls for time. Grey bars separate times of treatment as T0, T12, T24, and T48.

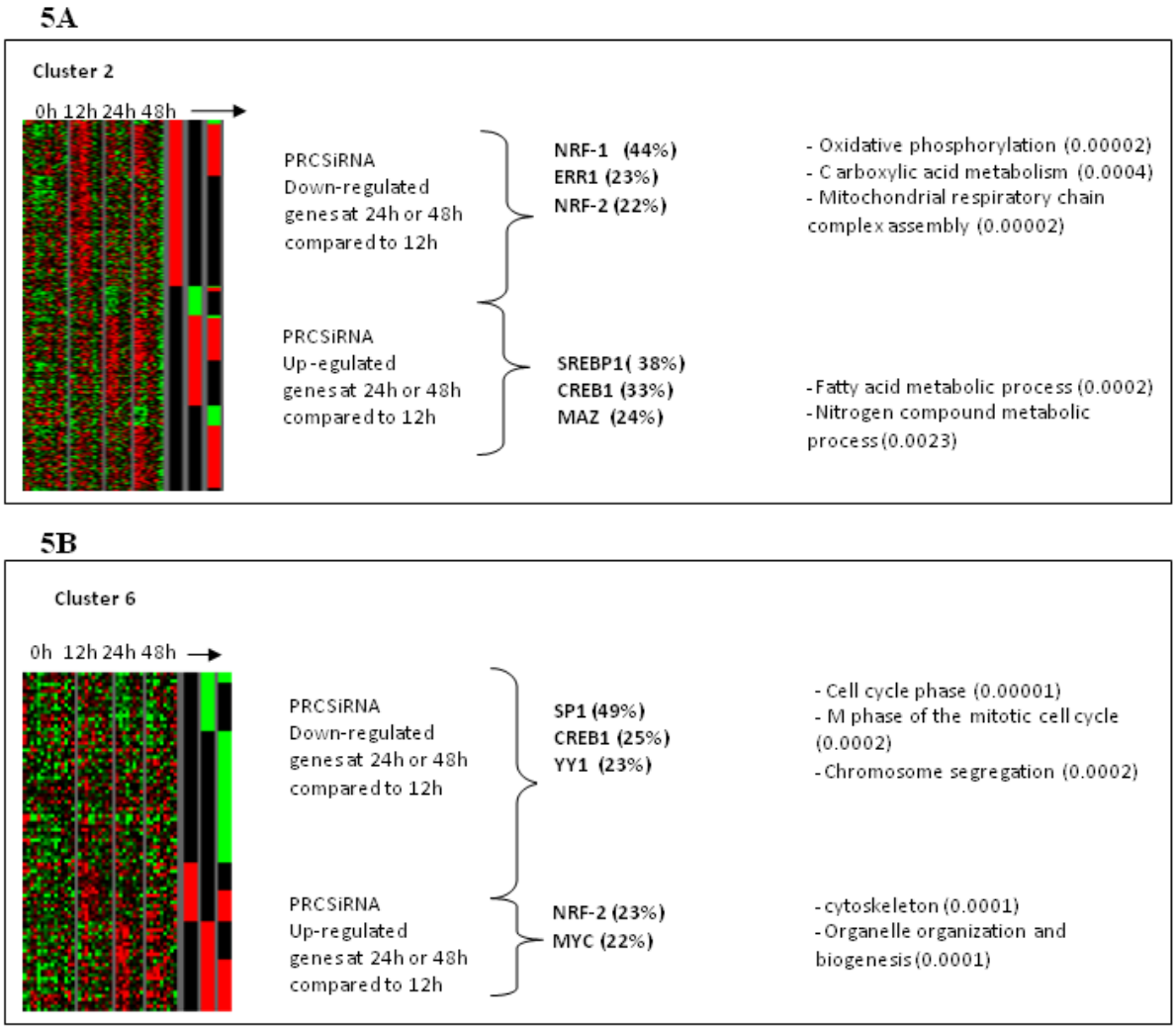


Figure 5: Focus on clusters 2 and 6 for PRC-regulated genes.
5A: PRC-sensitive under- and over-expressed genes in cluster 2
5B: PRC-sensitive under- and over-expressed genes in cluster 6
The arrow represents the differential genes in SiRNA PRC compared to negative controls at times T12, T24 and T48. Using MSignDB software (GSEA, MIT, MI, USA), the promoters of selected genes, positively or negatively regulated by PRC, were explored for transcription factor motifs. The percentage of genes in which the motif for a selected transcription factor was found is indicated in brackets. Highly significant ontologies are shown ($P \leq 0.01$, GOMiner software [50]).

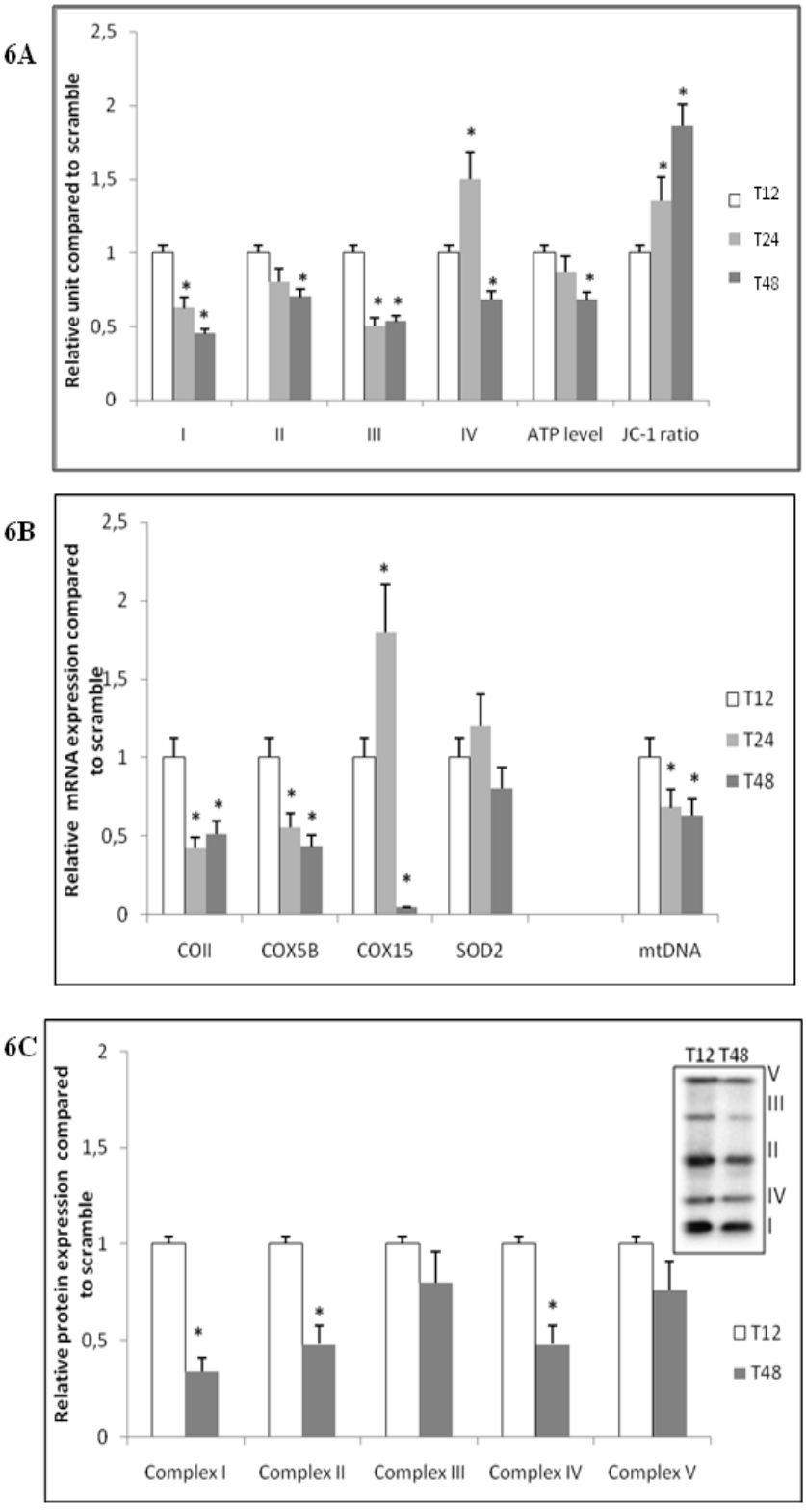


Figure 6: Functional and expression status of oxidative phosphorylation during PRC SiRNA treatment. T12 of serum induction during PRC SiRNA treatment was used as a reference to explore the efficient PRC inhibition at T24 and T48. (N=5; * $P \leq 0.05$).

6A: Enzymatic activities of four complexes of the respiratory chain (I, II, III and IV), total cell ATP and JC-1 ratios were measured on PRC SiRNA treated cells and related to control scramble cells. The JC-1 ratio represents the coupling efficiency between the respiratory chain and ATP synthesis.

6B: cDNA and mtDNA copy numbers of selected genes during PRC inhibition. The mitochondrial gene (COII) and the nuclear genes (COX5B, COX15) correspond to subunits of complex IV of the respiratory chain. Their expression, associated to mitochondrial SOD2 gene expression and mtDNA copy number were measured at T12, T24 and T48 of PRC SiRNA treatment and referred to the control scramble.

6C: Western blot analysis of subunits from each of the five complexes of OXPHOS using the MS601 antibody cocktail (Mitosciences, Eugene, OR, USA) during SiRNA treatment against PRC. Complex I (NDUFB8, 20 KDa), complex II (Ip subunit, 30 KDa); complex III (Core 2:40 KDa); complex IV (COX2, 24 KDa); complex V (ATPsynthase F1α, 58 KDa), relative to α-tubulin (65 KDa). Measurements of SiRNA treatment at T12 and T48 are referred to the negative control.

Cluster Number	Total genes	Differential gene expression PRC vs Negative control (p≤0.05) Over	Under	Principal Functional annotations compared to 5225 filtered cDNA probes	P -value
1	786	209	127	G-protein activating pathway Protein kinase and phosphatase activity Regulation of signal transduction Regulation of MAPK activity Regulation of angiogenesis	0.0013 0.0017 0.0055 0.0068 0.0116
2	538	260	65	Oxidative phosphorylation Response to stress Carboxylic acid metabolism Mitochondrial respiratory chain complex assembly Fatty acid metabolic process Nitrogen compound metabolic process Mitochondrion organization and biogenesis	0.0001 0.0001 0.0001 0.0011 0.0039 0.0049 0.0245
3	131	38	23	Chromatin remodelling Regulation of cytoskeleton organization Cell to cell signalling Cell junction assembly Transcription coactivator activity	0.0026 0.0026 0.0121 0.0218 0.0193
4	243	21	114	Ribosome biogenesis and assembly Vesicle organization and biogenesis Protein-RNA complex assembly Nucleic acid metabolic process mRNA processing	0.0001 0.0001 0.0002 0.0006 0.0022
5	57	48	3	Protein targeting to membrane DNA methyl transferase activity Spindle elongation process Cell cycle process Regulation of S phase of mitotic cycle	0.0030 0.0074 0.0107 0.0302 0.0318
6	198	91	48	Regulation of M phase of cell mitotic cycle Microtubule cytoskeleton organization Chromosome segregation Regulation of cyclin-dependent protein-kinase activity Regulation of transcription factor activity	0.0001 0.0001 0.0002 0.0016 0.0041
7	386	61	123	Response to DNA damage Response to endogenous stimulus Base excision repair process Mismatch repair process Primary metabolism process	0.0001 0.0001 0.0006 0.0006 0.0013
8	267	35	69	Mitochondrial ribosome assembly Regulation of RNA splicing SnRNA binding RNA modification Folding	0.0001 0.0003 0.0062 0.0062 0.0120

Table 1: Description and gene ontology of the genes from the 8 clusters identified and normalized on serum induction, differentially expressed between PRC SiRNA and the negative control.

COII	Forward	**5' - AAC-AAA-CGA-CCT-AAA-ACC-TG – 3'**
	Reverse	**5' - GTG-AAC-TAC-GAC-TGC-TAG-AA – 3'**
COX5B	Forward	**5' - CCA-AAG-GCA-GCT-TCA-GGC-AC – 3'**
	Reverse	**5' - CAA-GGA-AGA-CCC-TAA-TCT-AG – 3'**
COX15	Forward	**5' - CCT-CTC-GAT-GGT-AGA-TTG-GC – 3'**
	Reverse	**5' - CGG-AGG-GAG-TGC-AGT-GAC-AG – 3'**
ERR1	Forward	**5'- AAG-ACA-GCA-GCC-CCA-GTG-AA – 3'**
	Reverse	**5' - ACA-CCC-AGC-ACC-AGC-ACC-T – 3'**
ND5	Forward	**5' - GGG-GAT-TGT-GCG-GTG-TGT-G – 3'**
	Reverse	**5' - CTT-CTC-CTA-TTT-ATG-GGG-GT – 3'**
NRF-1	Forward	**5' - GGA-GTG-ATG-TCC-GCA-CAG-AA – 3'**
	Reverse	**5' - CGC-TGT-TAA-GCG-CCA-TAG-TG -3'**
PGC-1	Forward	**5' - ACT-CAA-GTG-GTG-CAG-TGA-CC – 3'**
	Reverse	**5' - CTG-GGT-ACT-GAG-ACC-ACT-GC – 3'**
PRC	Forward	**5'- GAT-CAG-AGC-AGC-GCT-GGG - 3'**
	Reverse	**5'- CAC-TAG-CAG-CTC-TCT-CCC-C – 3'**
SOD2	Forward	**5' - GCT-GCA-CCA-CAG-CAA-GCA-CC – 3'**
	Reverse	**5' - CCA-GCA-ACT-CCC-CTT-TGG-GT – 3'**
TFAM	Forward	**5' –CCG-AGG-TGG-TTT-TCA-TCT-GT- 3'**
	Reverse	**5' –CAG-GAA-GTT-CCC-TCC-AAC-GC-3'**
β-GLOBIN	Forward	**5' - GGT-GAA-CGT-GGA-TGA-AGT-TG – 3'**
	Reverse	**5' - GAG-CCA-GGC-CAT-CAC-TAA-AG – 3'**

Table 2: Primer sequences used for real-time quantitative RT-PCR

6. TRANSCRIPTOME + MOTIF+ChIP-chip (suite de 5)

Cette étude fait directement suite à l'étude précédente. Nous comptons intégrer les données de transcriptome précédemment décrit aux données de ChIP-chip que nous sommes en train de générer.

Afin de mieux déterminer l'action de PRC et de comprendre des mécanismes de régulation transcriptionnelle par cette voie, nous avons identifié les cibles du facteur NRF-1 par ChIP-chip dans les cellules XTC. Le but est d'intégrer ces données avec le transcriptome de l'étude précédente. Voici les résultats préliminaires de ce projet.

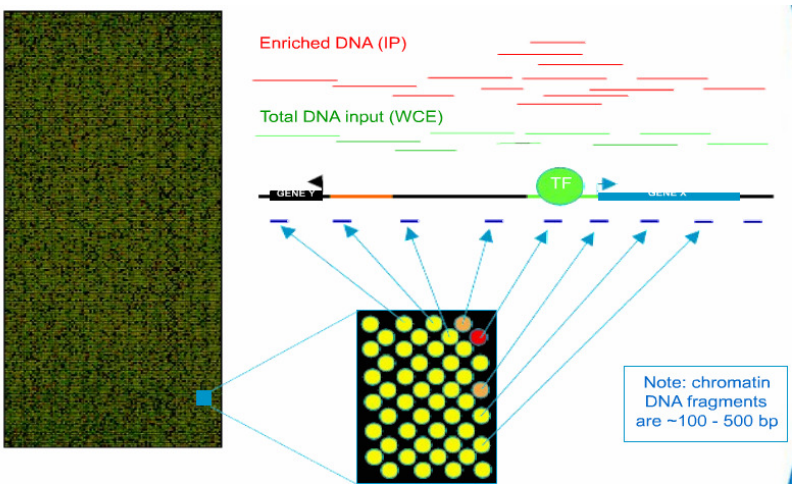
Nous avons utilisé des puces « promoteur » Agilent couvrant la région -5,5k +2,5 kb par rapport au TSS d'environ 17 000 gènes.

Des sondes positives en ChIP pour le facteur NRF-1 ont été identifiées (figure 1). Nous avons attribué une p-value pour les sondes positives, selon la méthode décrite dans la partie **3.3.2** :

Contrairement au transcriptome, les mesures dérivées des expériences en ChIP-chip apparaissent comme le mélange de deux distributions. La première correspond à la population de fragments génomiques enrichis par le ChIP (pour une sonde donnée : log-ratios IP/input positif, l'input étant de l'ADN génomique), et la deuxième au reste de la population non enrichie, et représentant le bruit de fond. L'observation de la distribution des log-ratios est donc asymétrique, avec un biais sur les valeurs positives. La distribution des log-ratios négatifs est approximativement gaussienne, alors que les positifs exhibent une queue non gaussienne. Nous avons réalisé des statistiques sur les valeurs négatives pour déterminer la significativité des valeurs positives. Une p-value est attribuée pour chacune des sondes. Nous avons sélectionné les sondes considérées comme positives ChIP avec un taux de faux-positif de 20%.

L'annotation des fonctionnelles des gènes cibles de NRF1 (enrichissement de termes d'ontologie) , une découverte de motifs et de co-occurrences de motifs, sont envisagées.

A



B

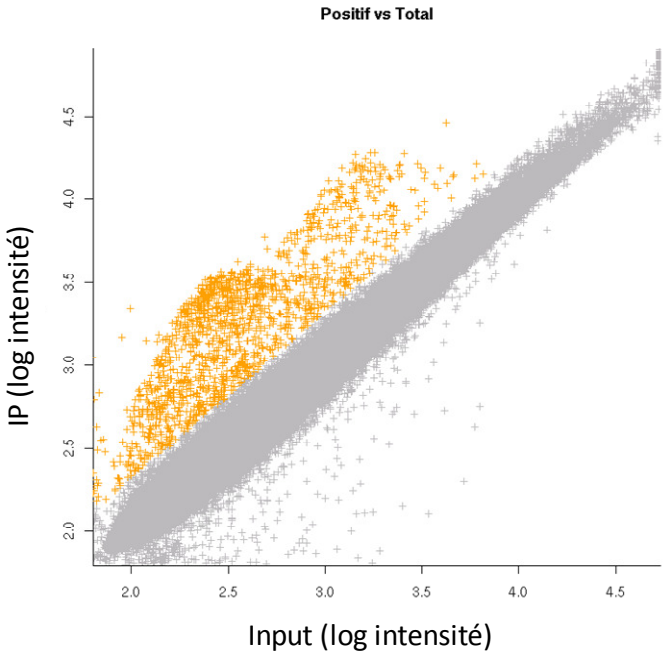


Figure 1: Analyse en ChIP-chip du facteur NRF-1. A. Principe de la quantification du signal des puces « promoteur ». B. Diagramme de dispersion des intensités des signaux dans l'IP et dans l'input. En orange sont représentées les sondes positives (FDR 20%)

.

Conclusion/Perspectives

Nous avons abordé l'étude des données d'expression à haut débit de plusieurs systèmes dans différentes situations biologiques. L'ensemble des résultats obtenus illustrent non seulement l'étendue des possibilités et des applications qu'offrent ces analyses en transcriptome par puces à ADN, mais aussi les nouvelles questions ou hypothèses que ces résultats peuvent soulever.

L'analyse des données nous a permis d'identifier des gènes et les fonctions biologiques associées à une condition physiologique particulière (Restriction calorique). Nous avons détecté des gènes et des fonctions biologiques qui sont sous l'influence d'un facteur de transcription dans l'AMC la larve de drosophile (Guenin et al.) ou d'un cofacteur dans une lignée cellulaire dérivée d'un oncocytome thyroïdien (Raharijaona et al.). Nous avons mis en évidence des signatures d'expression génique et des fonctions biologiques qui distinguaient une forme agressive d'une forme moins agressive d'un lymphome particulier (Willems, Roland, Raharijaona et al.). Nous avons également montré que différentes pathologies thyroïdiennes peuvent être caractérisées par les profils d'expression d'un nombre limité de gènes, des biomarqueurs, soulignant ainsi le fort pouvoir diagnostique et pronostique que proposent les puces à ADN (Fontaine et al.).
L'ensemble de ces études ont montré des modifications d'expression de groupes de gènes. Afin d'élucider des mécanismes moléculaires sous jacents aux modifications de l'expression génique, nous avons ajouté d'autres niveaux d'analyse du génome à ces données. Un motif correspondant à un site de liaison putatif d'un facteur de transcription est particulièrement surreprésenté dans les gènes répondant à Pros (Guenin et al.). Des combinaisons de motifs sont surpreprésentés dans un modèle cellulaire repondant au coactivateur PRC (Raharijaona et al.). Enfin, des changements en nombre de copies des oncogènes Myc et IRF4 ont été mis en évidence dans les lymphomes MZL agressifs.

Deux études ont mis en évidence la présence de motifs communs dans le promoteur d'un ensemble de gènes d'expression corrélée. Un site putatif de Pros a été proposé dans l'étude Guenin et al.. De même, dans l'étude de l'oncocytome thyroïdien, nous avons cherché des associations entre des groupes de gènes régulés par PRC et la présence d'un certain nombre de sites connus fixant des facteurs interagissant avec PRC. Bien que ces résultats proposent des pistes sur les mécanismes de régulation transcriptionnelle modulant l'expression de ces gènes, ils

ne donnent aucune preuve biologique de telles interactions. Un des objectifs de ce travail de thèse était la mise en place dans notre laboratoire de la technologie du ChIP-chip apportant ces preuves expérimentales à grande échelle. La principale difficulté est l'obtention d'une quantité suffisante de chromatine immunoprécipitée pour l'hybridation sur puces à ADN .Ceci dépend de la qualité des anticorps utilisés mais nécessite une étape d'amplification préalable à l'hybridation (WGA, Whole Genome Amplification), ceci afin de travailler sur une quantité de cellules adéquate avec le screening de plusieurs facteurs de transcription (idéalement 1 million de cellules). Nous avons rencontré des problèmes de reproductibilité, notamment lors de l'amplification de l'ADN à hybrider ou lors de la mesure des enrichissements des contrôles par PCR. Ces problèmes sont sur le point d'être résolus, et nous avons récemment commencé à générer des données exploitables dans le cas du modèle de l'oncocytome thyroïdien. Ainsi, l'étude des gènes cibles d'un ensemble de facteurs de transcription apparentés a mis en évidence des combinaisons spécifiques de facteurs associés à des profils d'expression identifiés en transcriptome. L'analyse des sites de liaison de PRC (tests avec un anticorps maison et un anticorps), NRF-1, NRF-2, ERRα et leur combinaison dans les régions régulatrices des gènes des différents clusters peut expliquer les profils d'expression observés dans la réponse à PRC. Ces études sont actuellement en cours et les résultats générés doivent être validés. Des efforts importants sont actuellement menés pour améliorer l'analyse des données en ChIP-chip, et de nouvelles méthodologies devront être testées et comparées (Song, Johnson et al. 2007). Des modules transcriptionnels (suite ordonnée de sites de liaison de facteurs de transcription) peuvent alors être mis en évidence par ChIP-chip ou par la découverte de motifs présents au voisinage des sites de liaison détectés expérimentalement. Des conservations de telles régions régulatrices dans des gènes orthologues devront confirmer leur rôle fonctionnel.

Des mises au point sur tissus sont en cours. Nous comptons appliquer le ChIP-chip en cancérologie, sur de plus petites quantités de tissus que sont les biopsies (Zhang, Zhong et al. 2009) et étudier donc ces mécanismes *in vivo*.

Des premiers résultats obtenus par analyse combinée du transcriptome PRC-dépendant et du ChIP-chip de trois facteurs de transcription (NRF1, NRF2 et ERRalpha) suggérent que pour un même cluster de fonctions régulées par le coactivateur PRC, plusieurs combinaisons de facteurs

de transcription peuvent opérer, ceci selon une cinétique qui semble être en phase avec le cycle cellulaire. L'analyse des réseaux de régulation au travers des différents clusters PRC-dépendant permet de valider à la fois de nouvelles fonctions régulées par une combinaison de facteurs de transcription mais aussi d'identifier le ou les facteurs de transcription les plus pertinents pour leur implication dans la coordination de diverses fonctions indispensables à la vie cellulaire. Ainsi nous proposons un premier schéma de régulation de la prolifération mitochondriale et cellulaire par la combinaison de PRC avec plusieurs facteurs de transcription, qui montre l'importance de facteurs tels que CREB-1 et NRF-2 dans la régulation conjointe (figure 7).

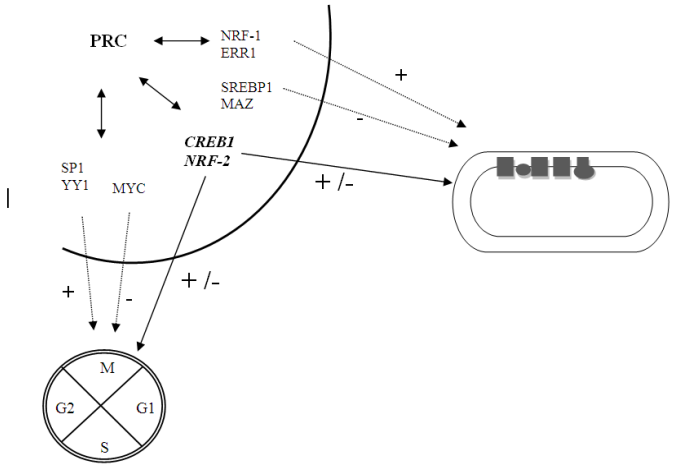


Figure7 Coordination entre cycle cellulaire et biogénèse mitochondriale. PRC coordonne la prolifération cellulaire et mitochondriale via l'interaction avec différents facteurs de transcription

Le ChIP-chip de tous les facteurs proposés devra être réalisé, mais sont déjà en cours les ChIP de CREB1 et YY1.

De même, la comparaison des cibles de Pros et des facteurs impliqués dans la voie Notch/EGFR dans l'AMC permettra notamment de déterminer dans quelle mesure les voies de signalisation Notch/ EGFR sont impliquées dans la réponse au facteur Pros.

De telles approches s'intéressent surtout à la régulation transcriptionnelle des gènes. Le profilage moléculaire des micros ARN par puce à ADN nous permettra de nous intéresser aux régulations post-transcriptionnelles. Nous explorons actuellement les microARN régulés par l'inhibition du

facteur PRC. Les premiers résultats montrent que certaines fonctions cellulaires vitales (la phosphorylation oxydative par exemple) sont régulées à la fois au niveau transcriptionnel et au niveau post transcriptionnel. Les mécanismes peuvent être relativement complexes. En plus de s'intéresser à leur régulation, il faudra mesurer leur impact sur l'expression de ces différents facteurs et de l'ensemble des gènes sensibles à PRC. Il sera sûrement nécessaire de développer de nouveaux outils capables d'intégrer les données microARN à celles du transcriptome et du ChIP-chip;

L'objectif final de l'intégration de ces données est de modéliser des réseaux biologiques du système étudié. La réalisation de réseaux biologiques à partir de l'intégration de données génomiques variées suppose une cohérence dans ces données et a de fait plus de puissance dans l'interprétation biologique. Des efforts sont actuellement en cours pour faciliter l'intégration de ces données de différentes natures et pour simplifier leur visualisation (Salari and Pollack 2009). Mais aussi de nouvelles technologies pourront être appliquées à nos modèles de pathologies Après près de quinze ans d'utilisation, les biopuces ont prouvé qu'elles étaient un outil de choix pour l'étude du génome. Cependant, des technologies de séquençage de nouvelle génération ont récemment émergé pour de telles études. Elles sont moins coûteuses que les méthodes traditionnelles de séquençage et s'affranchissent des étapes de biologie moléculaire (construction de librairies, clonage…) et des biais associés. Proposées notamment par 454/Roche, Illumina/Solexa ou SOLiD/Applied Biosystem, elles sont capables de séquencer en parallèle plusieurs millions de petits fragments d'ADN. Leurs applications sont quasiment les mêmes que celles des puces à ADN : caractérisation du transcriptome (Nielsen, Hogh et al. 2006), intéractions ADN-protéines ou des modifications d'histones (ChIP-seq (Johnson, Mortazavi et al. 2007) ; (Barski, Cuddapah et al. 2007), de la méthylation de l'ADN (Taylor, Kramer et al. 2007). L'exploitation des données reste encore difficile par rapport à celle que proposent les puces. La petite taille des séquences peut par exemple être un problème pour leur assemblage. De plus une forte proportion (15-20%) ne peut être cartographiée sur le génome sans ambigüité, alors que l'on sait précisemment ce que l'on mesure en puce puisque chaque sonde possède des annotations de gènes, d'exon, de localisation chromosomique... Ces nouvelles méthodes de séquençage risquent de concurrencer les biopuces. Peu d'études se sont pour l'instant attachées à comparer les résultats obtenus en séquençage haut débit et par puces à ADN. Euskirchen et al. (Euskirchen,

Rozowsky et al. 2007) ont mené une étude comparative ChIP-chip *vs.* ChIP-Pet. Leurs résultats suggèrent une forte concordance, entre les deux méthodes : les sites les plus enrichis sont communs entre les deux méthodes. Des différences sont observées pour les sites les moins enrichis. Une approche en ChIP-seq sera utile pour confirmer les cibles de nos facteurs et en identifier de nouvelles pour compléter nos réseaux. Des conclusions similaires ont également été tirées pour l'étude de profils d'expression (Liu, Jenssen et al. 2007). Dans notre laboratoire, nous nous sommes naturellement orientés vers une analyse du ChIP par puce à ADN, étant donné nos équipements, et notre expérience pour l'analyse des données de puces. Compte tenu de l'explosion rien n'exclut l'utilisation des techniques de séquençage à haut débit dans un futur proche.

Cette biologie intégrée pourra fournir ainsi des clés dans la description et la compréhension d'un système et sur l'origine des dysfonctionnements biologiques observés dans les pathologies. Une telle approche suscite de nombreux espoirs car la structure de ces réseaux pourra aider dans l'identification de cibles thérapeutiques encore plus pertinentes.

Abréviations

ACP Analyse en composante principale.

ADN Acide Désoxyribonucléique; ADNc: ADN complémentaire.

AMC Antenno-maxillary complex; complexe antenno-maxilaire.

ARN Acide Ribonucléique; **ARNm**: ARN messager; **ARNt**: ARN de transfert; **miRNA**: micro RNA : micro ARN; **si RNA**: small interfering RNA : petit ARN interférant; **shRNA**: short Hairpin RNA : ARN court en épingle à cheveux.

BAC Bacterial Artificial Chromosome; Chromosome artificiel bactérien.

CGH Comparative Genomic Hybridization; Hybridation Génomique Comparative.

ChIP Chromatin ImmunoPrecipitation; Immunoprécipitation de la chromatine.

CNS Central Nervous System; Système Nerveux Central.

CNV, CNC Copy-Number Variation, Variation de nombre de copies; copy number changes; changement de nombre de copies.

Cy3, Cy5 Cyanine 3, Cyanine 5.

DAG Directed acyclic graph; graphe orienté acyclique.

DamID DNA Adenine Methyl-transferase Identification.

DMH Differential Methylation Hybridation.

DS Discriminating Score ; Score discriminant

EM (algorithm) Expectation maximisation (algorithm) : (Algorithme)espérance-Maximisation.

EST Expressed Sequence Tag, Etiquette de séquences transcrites.

FDA Food and Drug Administration.

FL Follicular Lymphoma : lymphome folliculaire.

FTA Follicular Thyroïd Adenoma: Adénome folliculaire thyroïdien.

FTC Follicular Thyroïd Carcinoma: carcinome folliculaire thyroïdien.

GEO Gene expression omnibus.

GO Gene Ontology.

HGP Human Genome Project; Projet génome humain.

IP Immunoprécipitation

KNN K-Nearest-Neighbor: K plus proches voisins.

LDA Linear Discriminant Analysis; Analyse discriminante linéaire.

LOWESS Locally Weighted Scatterplot Smoothing

MAQC MicroArray Quality Control.

MeDIP Methylated DNA ImmunoPrecipitation: ImmunoPrécipitation de l'ADN méthylé.

MINDACT The Microarray in Node Negative and 0 to 3 Positive Lymph Node Disease May Avoid Chemotherapy Trial.

MZL, HP-MZL, MZL-MALT: Marginal zone lymphoma: lymphome de la zone marginale; Histologic Progression – MZL; marginal zone lymphoma of mucosa-associated lymphoid tissue: lymphome du tissu lymphoïde associé aux muqueuses.

PCR Polymerase chain reaction. Amplification/réaction en Chaîne par Polymérase.

PTC Papillary Thyroïd Carcinoma; carcinome papillaire thyroïdien.

PWM Position Weight Matrix: Matrice Poids-Position.

SAM Significant Analysis of Microarray.

SLL Small lymphocytic lymphoma: lymphome lymphocytique.

SNP Single Nucleotide Polymorphism: Polymorphisme de nucléotide unique.

SOM Self-Organizing Maps: Carte de Kohonen.

TMA Tissue MicroArray.

TSS Trascription Start Site: Site de départ de transcription.

Références Bibliographiques

Acharya, M. R., A. Sparreboom, et al. (2005). "Rational development of histone deacetylase inhibitors as anticancer agents: a review." Mol Pharmacol **68**(4): 917-32.

Adams, M. D., M. B. Soares, et al. (1993). "Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library." Nat Genet **4**(4): 373-80.

Adorjan, P., J. Distler, et al. (2002). "Tumour class prediction and discovery by microarray-based DNA methylation analysis." Nucleic Acids Res **30**(5): e21.

Al-Shahrour, F., R. Diaz-Uriarte, et al. (2004). "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes." Bioinformatics **20**(4): 578-80.

Alizadeh, A. A., M. B. Eisen, et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature **403**(6769): 503-11.

Andre, F., B. Job, et al. (2009). "Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array." Clin Cancer Res **15**(2): 441-51.

Antequera, F. and A. Bird (1993). "CpG islands." Exs **64**: 169-85.

Antequera, F. and A. Bird (1993). "Number of CpG islands and genes in human and mouse." Proc Natl Acad Sci U S A **90**(24): 11995-9.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Ballester, B., O. Ramuz, et al. (2006). "Gene expression profiling identifies molecular subgroups among nodal peripheral T-cell lymphomas." Oncogene **25**(10): 1560-70.

Bannister, A. J. and T. Kouzarides (2005). "Reversing histone methylation." Nature **436**(7054): 1103-6.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-37.

Benjamini, Y. a. H., Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testings." Journal of the RoyalStatistical Society: Series B **57**: 289-300.

Bergamaschi, A., Y. H. Kim, et al. (2006). "Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer." Genes Chromosomes Cancer **45**(11): 1033-40.

Bernstein, B. E., E. L. Humphrey, et al. (2002). "Methylation of histone H3 Lys 4 in coding regions of active genes." Proc Natl Acad Sci U S A **99**(13): 8695-700.

Bertucci, F., P. Finetti, et al. (2005). "Gene expression profiling identifies molecular subtypes of inflammatory breast cancer." Cancer Res **65**(6): 2170-8.

Bertucci, F., V. Nasser, et al. (2002). "Gene expression profiles of poor-prognosis primary breast cancer correlate with survival." Hum Mol Genet **11**(8): 863-72.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev **16**(1): 6-21.

Birrer, M. J., M. E. Johnson, et al. (2007). "Whole genome oligonucleotide-based array comparative genomic hybridization analysis identified fibroblast growth factor 1 as a prognostic marker for advanced-stage serous ovarian adenocarcinomas." J Clin Oncol **25**(16): 2281-7.

Blaveri, E., J. L. Brewer, et al. (2005). "Bladder cancer stage and outcome by array-based comparative genomic hybridization." Clin Cancer Res **11**(19 Pt 1): 7012-22.
Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-93.
Brazma, A., H. Parkinson, et al. (2003). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res **31**(1): 68-71.
Bucher, P. (1990). "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences." J Mol Biol **212**(4): 563-78.
Buck, M. J., A. B. Nobel, et al. (2005). "ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data." Genome Biol **6**(11): R97.
Bulyk, M. L., E. Gentalen, et al. (1999). "Quantifying DNA-protein interactions by double-stranded DNA arrays." Nat Biotechnol **17**(6): 573-7.
Buyse, M., S. Loi, et al. (2006). "Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer." J Natl Cancer Inst **98**(17): 1183-92.
Cardoso, F., M. Piccart-Gebhart, et al. (2007). "The MINDACT trial: the first prospective clinical validation of a genomic tool." Mol Oncol **1**(3): 246-51.
Carninci, P., J. Yasuda, et al. (2008). "Multifaceted mammalian transcriptome." Curr Opin Cell Biol **20**(3): 274-80.
Carrasco, D. R., G. Tonon, et al. (2006). "High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients." Cancer Cell **9**(4): 313-25.
Carroll, J. S., C. A. Meyer, et al. (2006). "Genome-wide analysis of estrogen receptor binding sites." Nat Genet **38**(11): 1289-97.
Chen, R. Z., U. Pettersson, et al. (1998). "DNA hypomethylation leads to elevated mutation rates." Nature **395**(6697): 89-93.
Chen, W., J. Houldsworth, et al. (2006). "Array comparative genomic hybridization reveals genomic copy number changes associated with outcome in diffuse large B-cell lymphomas." Blood **107**(6): 2477-85.
Ching, T. T., A. K. Maunakea, et al. (2005). "Epigenome analyses using BAC microarrays identify evolutionary conservation of tissue-specific methylation of SHANK3." Nat Genet **37**(6): 645-51.
Churchill, G. A. (2002). "Fundamentals of experimental design for cDNA microarrays." Nat Genet **32 Suppl**: 490-5.
Clark, T. A., A. C. Schweitzer, et al. (2007). "Discovery of tissue-specific exons using comprehensive human exon microarrays." Genome Biol **8**(4): R64.
Climent, J., P. Dimitrow, et al. (2007). "Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer." Cancer Res **67**(2): 818-26.
Crawley, J. J. and K. A. Furge (2002). "Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data." Genome Biol **3**(12): RESEARCH0075.
Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-63.
Datta, S. and S. Datta (2003). "Comparisons and validation of statistical clustering techniques for microarray gene expression data." Bioinformatics **19**(4): 459-66.
de Fraipont, F., M. El Atifi, et al. (2005). "Gene expression profiling of human adrenocortical tumors using complementary deoxyribonucleic Acid microarrays identifies several candidate genes as markers of malignancy." J Clin Endocrinol Metab **90**(3): 1819-29.

Dudoit, S. and J. Fridlyand (2002). "A prediction-based resampling method for estimating the number of clusters in a dataset." Genome Biol **3**(7): RESEARCH0036.
Dutt, A. and R. Beroukhim (2007). "Single nucleotide polymorphism array analysis of cancer." Curr Opin Oncol **19**(1): 43-9.
Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-10.
Egger, G., G. Liang, et al. (2004). "Epigenetics in human disease and prospects for epigenetic therapy." Nature **429**(6990): 457-63.
Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.
Esteller, M. (2005). "Aberrant DNA methylation as a cancer-inducing mechanism." Annu Rev Pharmacol Toxicol **45**: 629-56.
Euskirchen, G. M., J. S. Rozowsky, et al. (2007). "Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies." Genome Res **17**(6): 898-909.
Feinberg, A. P. and B. Tycko (2004). "The history of cancer epigenetics." Nat Rev Cancer **4**(2): 143-53.
Feuk, L., A. R. Carson, et al. (2006). "Structural variation in the human genome." Nat Rev Genet **7**(2): 85-97.
Frayling, T. M. (2007). "Genome-wide association studies provide new insights into type 2 diabetes aetiology." Nat Rev Genet **8**(9): 657-62.
Frey, B. J., N. Mohammad, et al. (2005). "Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs." Nat Genet **37**(9): 991-6.
Garinis, G. A., G. P. Patrinos, et al. (2002). "DNA hypermethylation: when tumour suppressor genes go silent." Hum Genet **111**(2): 115-27.
Ginsberg, S. D. (2005). "RNA amplification strategies for small sample populations." Methods **37**(3): 229-37.
Golub, T. R., D. K. Slonim, et al. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." Science **286**(5439): 531-7.
Gress, T. M., J. D. Hoheisel, et al. (1992). "Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues." Mamm Genome **3**(11): 609-19.
Gunderson, K. L., S. Kruglyak, et al. (2004). "Decoding randomly ordered DNA arrays." Genome Res **14**(5): 870-7.
Haibe-Kains, B., C. Desmedt, et al. (2008). "Comparison of prognostic gene expression signatures for breast cancer." BMC Genomics **9**: 394.
Hatchwell, E. and J. M. Greally (2007). "The potential role of epigenomic dysregulation in complex human disease." Trends Genet **23**(11): 588-95.
Herman, J. G. and S. B. Baylin (2003). "Gene silencing in cancer in association with promoter hypermethylation." N Engl J Med **349**(21): 2042-54.
Hewitt, S. M. (2006). "The application of tissue microarrays in the validation of microarray results." Methods Enzymol **410**: 400-15.
Hosack, D. A., G. Dennis, Jr., et al. (2003). "Identifying biological themes within lists of genes with EASE." Genome Biol **4**(10): R70.
Huang, H. E., S. F. Chin, et al. (2004). "A recurrent chromosome breakpoint in breast cancer at the NRG1/neuregulin 1/heregulin gene." Cancer Res **64**(19): 6840-4.

Huang, T. H., M. R. Perry, et al. (1999). "Methylation profiling of CpG islands in human breast cancer cells." Hum Mol Genet **8**(3): 459-70.
Hughes, T. R., M. Mao, et al. (2001). "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer." Nat Biotechnol **19**(4): 342-7.
Hunter, D. J., P. Kraft, et al. (2007). "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer." Nat Genet **39**(7): 870-4.
Hurowitz, E. H. and P. O. Brown (2003). "Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae." Genome Biol **5**(1): R2.
Ishkanian, A. S., C. A. Malloff, et al. (2004). "A tiling resolution DNA microarray with complete coverage of the human genome." Nat Genet **36**(3): 299-303.
Iyer, V. R., C. E. Horak, et al. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF." Nature **409**(6819): 533-8.
Jacquemont, M. L., D. Sanlaville, et al. (2006). "Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders." J Med Genet **43**(11): 843-9.
Jaenisch, R. and A. Bird (2003). "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals." Nat Genet **33 Suppl**: 245-54.
Jansen, R. C. and J. P. Nap (2001). "Genetical genomics: the added value from segregation." Trends Genet **17**(7): 388-91.
Johnson, D. S., A. Mortazavi, et al. (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-502.
Jones, P. A. and S. B. Baylin (2002). "The fundamental role of epigenetic events in cancer." Nat Rev Genet **3**(6): 415-28.
Kim, S. W., J. W. Kim, et al. (2007). "Analysis of chromosomal changes in serous ovarian carcinoma using high-resolution array comparative genomic hybridization: Potential predictive markers of chemoresistant disease." Genes Chromosomes Cancer **46**(1): 1-9.
Kim, T. H., L. O. Barrera, et al. (2005). "A high-resolution map of active promoters in the human genome." Nature **436**(7052): 876-80.
Kirmizis, A., S. M. Bartley, et al. (2004). "Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27." Genes Dev **18**(13): 1592-605.
Kononen, J., L. Bubendorf, et al. (1998). "Tissue microarrays for high-throughput molecular profiling of tumor specimens." Nat Med **4**(7): 844-7.
Kouzarides, T. (2007). "Chromatin modifications and their function." Cell **128**(4): 693-705.
Kozal, M. J., N. Shah, et al. (1996). "Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays." Nat Med **2**(7): 753-9.
Kreil, D. P. and R. R. Russell (2005). "There is no silver bullet--a guide to low-level data transforms and normalisation methods for microarray data." Brief Bioinform **6**(1): 86-97.
Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.
Lawrence, C. E. and A. A. Reilly (1990). "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences." Proteins **7**(1): 41-51.
Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.

Lenz, G., G. W. Wright, et al. (2008). "Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways." Proc Natl Acad Sci U S A **105**(36): 13520-5.
Liang, P. and A. B. Pardee (1992). "Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction." Science **257**(5072): 967-71.
Liew, C. C. and V. J. Dzau (2004). "Molecular genetics and genomics of heart failure." Nat Rev Genet **5**(11): 811-25.
Lin, B., Z. Wang, et al. (2006). "Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays." Genome Res **16**(4): 527-35.
Liu, C. G., G. A. Calin, et al. (2004). "An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues." Proc Natl Acad Sci U S A **101**(26): 9740-4.
Liu, F., T. K. Jenssen, et al. (2007). "Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates." BMC Genomics **8**: 153.
Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.
Loi, S., B. Haibe-Kains, et al. (2008). "Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen." BMC Genomics **9**: 239.
MacBeath, G. (2002). "Protein microarrays and proteomics." Nat Genet **32 Suppl**: 526-32.
MacBeath, G. and S. L. Schreiber (2000). "Printing proteins as microarrays for high-throughput function determination." Science **289**(5485): 1760-3.
Martin-Magniette, M. L., T. Mary-Huard, et al. (2008). "ChIPmix: mixture model of regressions for two-color ChIP-chip analysis." Bioinformatics **24**(16): i181-6.
Martinez, R., J. I. Martin-Subero, et al. (2009). "A microarray-based DNA methylation study of glioblastoma multiforme." Epigenetics **4**(4): 255-64.
McKusick, V. A. and F. H. Ruddle (1987). "A new discipline, a new name, a new journal [editorial]." Genomics **1**: 1-2.
Mestre-Escorihuela, C., F. Rubio-Moscardo, et al. (2007). "Homozygous deletions localize novel tumor suppressor genes in B-cell lymphomas." Blood **109**(1): 271-80.
Miller, L. D., P. M. Long, et al. (2002). "Optimal gene expression analysis by microarrays." Cancer Cell **2**(5): 353-61.
Nguyen, C., D. Rocha, et al. (1995). "Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones." Genomics **29**(1): 207-16.
Nielsen, K. L., A. L. Hogh, et al. (2006). "DeepSAGE--digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples." Nucleic Acids Res **34**(19): e133.
Odom, D. T., N. Zizlsperger, et al. (2004). "Control of pancreas and liver gene expression by HNF transcription factors." Science **303**(5662): 1378-81.
Orlando, V. and R. Paro (1993). "Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin." Cell **75**(6): 1187-98.
Paez, J. G., M. Lin, et al. (2004). "Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification." Nucleic Acids Res **32**(9): e71.
Paris, P. L. (2009). "A Whole-Genome Amplification Protocol for a Wide Variety of DNAs, Including Those from Formalin-Fixed and Paraffin-Embedded Tissue." Methods Mol Biol **556**: 89-98.

Penn, S. G., D. R. Rank, et al. (2000). "Mining the human genome using microarrays of open reading frames." Nat Genet **26**(3): 315-8.
Pinkel, D., R. Segraves, et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." Nat Genet **20**(2): 207-11.
Pollack, J. R., C. M. Perou, et al. (1999). "Genome-wide analysis of DNA copy-number changes using cDNA microarrays." Nat Genet **23**(1): 41-6.
Quackenbush, J. (2002). "Microarray data normalization and transformation." Nat Genet **32 Suppl**: 496-501.
Ramakrishnan, R., D. Dorris, et al. (2002). "An assessment of Motorola CodeLink microarray performance for gene expression profiling applications." Nucleic Acids Res **30**(7): e30.
Raychaudhuri, S., J. M. Stuart, et al. (2000). "Principal components analysis to summarize microarray experiments: application to sporulation time series." Pac Symp Biocomput: 455-66.
Reiner, A., D. Yekutieli, et al. (2003). "Identifying differentially expressed genes using false discovery rate controlling procedures." Bioinformatics **19**(3): 368-75.
Ren, B., F. Robert, et al. (2000). "Genome-wide location and function of DNA binding proteins." Science **290**(5500): 2306-9.
Rhodes, D. R., J. Yu, et al. (2004). "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression." Proc Natl Acad Sci U S A **101**(25): 9309-14.
Roth, F. P., J. D. Hughes, et al. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." Nat Biotechnol **16**(10): 939-45.
Salari, K. and J. R. Pollack (2009). "Integration of diverse microarray data types." Methods Mol Biol **556**: 205-16.
Sanoudou, D., L. A. Frieden, et al. (2004). "Molecular classification of nemaline myopathies: "nontyping" specimens exhibit unique patterns of gene expression." Neurobiol Dis **15**(3): 590-600.
Scharer, C. D., C. D. McCabe, et al. (2009). "Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells." Cancer Res **69**(2): 709-17.
Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.
Schmutz, J., J. Wheeler, et al. (2004). "Quality assessment of the human genome sequence." Nature **429**(6990): 365-8.
Sebat, J., B. Lakshmi, et al. (2007). "Strong association of de novo copy number mutations with autism." Science **316**(5823): 445-9.
Shaffer, A. L., N. C. Emre, et al. (2008). "IRF4 addiction in multiple myeloma." Nature **454**(7201): 226-31.
Shaffer, A. L., A. Rosenwald, et al. (2001). "Signatures of the immune response." Immunity **15**(3): 375-85.
Shi, L., L. H. Reid, et al. (2006). "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." Nat Biotechnol **24**(9): 1151-61.
Shin, I., J. W. Cho, et al. (2004). "Carbohydrate arrays for functional studies of carbohydrates." Comb Chem High Throughput Screen **7**(6): 565-74.

Slonim, D. K. (2002). "From patterns to pathways: gene expression data analysis comes of age." Nat Genet **32 Suppl**: 502-8.
Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Stat Appl Genet Mol Biol **3**: Article3.
Song, J. S., W. E. Johnson, et al. (2007). "Model-based analysis of two-color arrays (MA2C)." Genome Biol **8**(8): R178.
Sorlie, T., C. M. Perou, et al. (2001). "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." Proc Natl Acad Sci U S A **98**(19): 10869-74.
Staudt, L. M. and S. Dave (2005). "The biology of human lymphoid malignancies revealed by gene expression profiling." Adv Immunol **87**: 163-208.
Steenman, M., G. Lamirault, et al. (2005). "Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays." Eur J Heart Fail **7**(2): 157-65.
Steinemann, D., B. Skawran, et al. (2006). "Assessment of differentiation and progression of hepatic tumors using array-based comparative genomic hybridization." Clin Gastroenterol Hepatol **4**(10): 1283-91.
Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-5.
Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-50.
Tagawa, H., M. Suguro, et al. (2005). "Comparison of genome profiles for identification of distinct subgroups of diffuse large B-cell lymphoma." Blood **106**(5): 1770-7.
Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proc Natl Acad Sci U S A **96**(6): 2907-12.
Tan, P. K., T. J. Downey, et al. (2003). "Evaluation of gene expression measurements from commercial microarray platforms." Nucleic Acids Res **31**(19): 5676-84.
Tavazoie, S., J. D. Hughes, et al. (1999). "Systematic determination of genetic network architecture." Nat Genet **22**(3): 281-5.
Taylor, K. H., R. S. Kramer, et al. (2007). "Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing." Cancer Res **67**(18): 8511-8.
Thieblemont, C., V. Nasser, et al. (2004). "Small lymphocytic lymphoma, marginal zone B-cell lymphoma, and mantle cell lymphoma exhibit distinct gene-expression profiles allowing molecular diagnosis." Blood **103**(7): 2727-37.
Thomas, G., K. B. Jacobs, et al. (2008). "Multiple loci identified in a genome-wide association study of prostate cancer." Nat Genet **40**(3): 310-5.
Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.
van 't Veer, L. J., H. Dai, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-6.
van Helden, J., B. Andre, et al. (1998). "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies." J Mol Biol **281**(5): 827-42.
van Steensel, B. and S. Henikoff (2000). "Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase." Nat Biotechnol **18**(4): 424-8.

Velculescu, V. E., L. Zhang, et al. (1995). "Serial analysis of gene expression." Science **270**(5235): 484-7.
Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.
Wang, Y., S. M. Hewitt, et al. (2006). "Tissue microarray analysis of human FRAT1 expression and its correlation with the subcellular localisation of beta-catenin in ovarian tumours." Br J Cancer **94**(5): 686-91.
Waterman, M. S., R. Arratia, et al. (1984). "Pattern recognition in several sequences: consensus and alignment." Bull Math Biol **46**(4): 515-27.
Weber, M., J. J. Davies, et al. (2005). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells." Nat Genet **37**(8): 853-62.
Wei, S. H., C. M. Chen, et al. (2002). "Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers." Clin Cancer Res **8**(7): 2246-52.
Wessels, L. F., T. van Welsem, et al. (2002). "Molecular classification of breast carcinomas by comparative genomic hybridization: a specific somatic genetic profile for BRCA1 tumors." Cancer Res **62**(23): 7110-7.
Winkler (1920). *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche. Jena: Gustav Fischer Verlag*.
Workman, C., L. J. Jensen, et al. (2002). "A new non-linear normalization method for reducing variability in DNA microarray experiments." Genome Biol **3**(9): research0048.
Yan, P. S., C. M. Chen, et al. (2001). "Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays." Cancer Res **61**(23): 8375-80.
Yang, Y. H., S. Dudoit, et al. (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." Nucleic Acids Res **30**(4): e15.
Yang, Y. H. and T. Speed (2002). "Design issues for cDNA microarray experiments." Nat Rev Genet **3**(8): 579-88.
Yeager, M., N. Orr, et al. (2007). "Genome-wide association study of prostate cancer identifies a second risk locus at 8q24." Nat Genet **39**(5): 645-9.
Yin, J. Q., R. C. Zhao, et al. (2008). "Profiling microRNA expression with microarrays." Trends Biotechnol **26**(2): 70-6.
Zanke, B. W., C. M. Greenwood, et al. (2007). "Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24." Nat Genet **39**(8): 989-94.
Zeeberg, B. R., W. Feng, et al. (2003). "GoMiner: a resource for biological interpretation of genomic and proteomic data." Genome Biol **4**(4): R28.
Zhang, L., K. Zhong, et al. (2009). "Genome-wide analysis of histone H3 lysine 27 trimethylation by ChIP-chip in gastric cancer patients." J Gastroenterol **44**(4): 305-12.
Zhou, Y., S. M. Luoh, et al. (2003). "Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis." Cancer Res **63**(18): 5781-4.
Zhu, H., M. Bilgin, et al. (2001). "Global analysis of protein activities using proteome chips." Science **293**(5537): 2101-5.
Zhu, H., J. F. Klemic, et al. (2000). "Analysis of yeast protein kinases using protein chips." Nat Genet **26**(3): 283-9.
Zhu, H. and M. Snyder (2003). "Protein chip technology." Curr Opin Chem Biol **7**(1): 55-63.

Ziauddin, J. and D. M. Sabatini (2001). "Microarrays of cells expressing defined cDNAs." Nature **411**(6833): 107-10.

Annexes

Molecular risk stratification in advanced heart failure patients

Guillaume Lamirault; Nolwenn Le Meur; Jean-Christian Roussel ; Marie-France Le Cunff; Daniel Baron; Audrey Bihouée; Isabelle Guisle; Mahatsangy Raharijaona; Gérard Ramstein; Raluca Teusan; Catherine Chevalier; Jean-Pierre Gueffet; Jean-Noël Trochu; Jean J. Léger; Rémi Houlgatte; Marja Steenman ;

L'insuffisance cardiaque est une maladie qui demeure grave malgré de nombreux progrès thérapeutique. Une stratification du risque est impérative afin d'approcher au plus près le pronostic du patient. Elle est cruciale pour l'individualisation de la stratégie thérapeutique, en particulier dans la prise de décision pour l'implantation d'une assistance ventriculaire ou d'une transplantation cardiaque. Des modèles de prédiction ont été développés pour des insuffisances cardiaques modérées, mais restent à améliorer pour des états plus avancés.

Il a été montré que la sévérité de l'insuffisance cardiaque corrèle avec l'intensité du remodelage cardiaque se produisant lors de la progression de la pathologie. Ce remodelage est en relation avec des altérations touchant plusieurs voies et fonctions biologiques, modifiant alors le tissu et les caractéristiques du myocarde. Des études récentes ont montré que des profils d'expression de gènes pouvaient distinguer, même parmi des insuffisants cardiaques avancés, des sous-groupes de patients avec des portraits moléculaires cardiaques spécifiques

Dans cette étude, nous voulons tester l'hypothèse que des profils d'expression génique peuvent distinguer des patients en insuffisance cardiaque avec différent degré de risque de décès.

Nous avons analysé les profils d'expression du ventricule droit et gauche de cœurs explantés de 44 patients souffrant d'insuffisance cardiaque avancée et ayant subi une transplantation cardiaque ou l'implantation d'une assistance ventriculaire. Pour cela nous avons utilisé une puce dédiée maison contenant 4217 sondes pertinentes pour l'étude du transcriptome cardiaque.

Les patients ont été classés, selon leur statut clinique, en 3 groupes reflétant la sévérité de la maladie.

L'analyse du transcriptome a permis d'identifier un classifieur de 170 gènes et un autre de 129 gènes (66 gènes en communs) pour les échantillons de ventricules gauche et droit respectivement, discriminant le groupe de patients à l'insuffisance la plus sévère du groupe de patients à l'insuffisance la moins sévère. Le classifieur issu du ventricule gauche a identifié les patients du premier groupe avec une sensibilité de 92% et une spécificité de 96%, et les patients du deuxième

groupe avec une sensibilité de 88% et une spécificité de 100%. Le classifieur issu du ventricule droit a identifié les patients du premier groupe avec une sensibilité de 96% et une spécificité de 100%, et les patients du deuxième groupe avec une sensibilité de 100% et une spécificité de 100%. Des échantillons du groupe intermédiaire, qui n'a pas été utilisé pour construire le classifieur, ont pu être classés dans l'un des deux autres groupes.

L'étude des profils d'expression génique permet de détecter des patients en insuffisance cardiaque à très haut risque de décès, avec une bonne sensibilité et spécificité. Des échantillons de ventricule droit ou gauche peuvent être analysés pour la stratification du risque avec un pouvoir prédictif similaire. Ces résultats montrent que l'utilisation des profils d'expression géniques pourrait être appliquée à l'évaluation des patients en insuffisance cardiaque avancée et suggèrent une évaluation de cette approche à des stades plus précoces de la maladie.

Molecular risk stratification in advanced heart failure patients

Guillaume Lamirault, MD, PhD[a,b,c]; Nolwenn Le Meur, PhD[a,b,c]; Jean-Christian Roussel, MD[c]; Marie-France Le Cunff[a,b,c]; Daniel Baron, PhD[a,b,c]; Audrey Bihouée, MSc[a,b,c]; Isabelle Guisle[a,b,c]; Mahatsangy Raharijaona, MSc[a,b,c]; Gérard Ramstein, PhD[a,b,c,d]; Raluca Teusan, MSc[a,b,c]; Catherine Chevalier, PhD[a,b,c]; Jean-Pierre Gueffet, MD[a,b,c]; Jean-Noël Trochu, MD, PhD[a,b,c]; Jean J. Léger, PhD[a,b,c]; Rémi Houlgatte, PhD[a,b,c]; Marja Steenman, PhD[a,b,c,*]

Affiliations:

[a] Inserm, U915, Nantes, F-44000, France;

[b] Université de Nantes, Faculté de Médecine, Nantes, F-44000, France;

[c] CHU de Nantes, l'institut du thorax, CIC, Nantes, F-44000, France;

[d] Laboratoire d'Informatique de Nantes Atlantique, Nantes, F-44322, France

[*] **Correspondence to**: Dr Marja Steenman, l'institut du thorax, INSERM U915, Faculté de Médecine, 1 rue Gaston Veil, 44035 Nantes, France

Fax: +33 240412950

Tel: +33 240411122

marja.steenman@nantes.inserm.fr

ABSTRACT:
Background: Risk stratification in advanced heart failure is crucial for the individualization of therapeutic strategy, in particular for heart transplantation and ventricular assist device implantation.
Objectives: We tested the hypothesis that cardiac gene expression profiling can distinguish between heart failure patients with different risks of death.
Methods: We obtained tissue samples from both left (LV) and right (RV) ventricle of explanted hearts of 44 patients undergoing cardiac transplantation or ventricular assist device placement. Gene expression profiles were obtained using an in-house developed microarray containing 4217 cardiac-relevant genes. Based on their clinical status, patients were classified into three heart-failure-severity groups: Deteriorating (n=12), Intermediate (n=19), and Stable (n=13).
Results: Two-class statistical analysis of gene expression profiles of Deteriorating and Stable patients identified a 170-gene and a 129-gene predictor for LV and RV samples, respectively. The LV molecular predictor identified patients with stable and deteriorating status with a sensitivity of 88% and 92%, and a specificity of 100% and 96%, respectively. The RV molecular predictor identified patients with stable and deteriorating status with a sensitivity of 100% and 96%, and a specificity of 100% and 100%, respectively. The molecular prediction was reproducible across biological replicates in LV and RV samples.
Conclusions: Gene expression profiling has the potential to reproducibly detect heart failure patients at highest risk of death with high sensitivity and specificity. In addition, not only LV but also RV samples could be used for molecular risk stratification with similar predictive power.
Key words: genes, heart failure, remodeling, ventricles, microarrays

Main topics: heart failure, gene expression profiling.

INTRODUCTION

Risk stratification in advanced heart failure (Venter, Adams et al.) aims at identifying patients who will rapidly progress to refractory myocardial dysfunction or who are at high risk of sudden cardiac death. This stratification is crucial for the individualization of therapeutic strategy, in particular for the listing and prioritization of patients for heart transplantation, and identification of patients for left ventricular assist devices (LVAD). Prediction models have been mainly developed for moderate HF [1, 2], some of them being applicable to end-stage HF patients [3, 4]. However, since these models have modest predictive capacity, outcome prediction still remains to be improved in advanced HF.
It has already been shown that HF severity correlates with the intensity of the cardiac remodeling process occurring during HF progression [5]. This remodeling process is related to transcriptomal alterations affecting numerous molecular pathways and biological functions, modifying tissue and morphological characteristics of the myocardium [6]. One tool that might lead to better outcome prediction is gene expression profiling. We and others recently showed that gene expression profiling could distinguish, even in advanced HF, sub-groups of patients with specific cardiac molecular portraits [7-11].
Here we demonstrate that cardiac gene expression profiling can distinguish between heart failure patients with different risks of death.

METHODS

Cardiac samples

Cardiac tissue was obtained from explanted hearts of 44 patients with advanced HF who underwent a cardiac transplantation or a total

artificial heart placement. Pre-transplant evaluation - including coronary artery angiography and macroscopic and histological examination of the explanted hearts - confirmed the diagnosis, etiology and severity of the disease for all patients. Extensive individual clinical information can be found in the on-line data supplement Table 1.

Patients were classified into three severity groups based on their clinical status at the time of transplantation. The individual clinical status was defined based on the United Network for Organ Sharing (Garinis, Patrinos et al.) medical urgency status [12] and occurrence of hospitalizations for Acute Decompensated Heart Failure (ADHF) during the three months prior to the surgical procedure ('recent ADHF'). Deteriorating patients were characterized by UNOS-1A status. Stable patients were defined as UNOS-2 patients with no recent ADHF. The remaining patients were classified as Intermediate (UNOS-1B status or UNOS-2 status with recent ADHF).

For each of the 44 explanted hearts, two spatially distinct transmural samples were obtained from non-infarcted zones of both left ventricle (LV) and right ventricle (RV) immediately after cardiac explantation, leading to a total of 176 distinct tissue samples.

Microarrays

Microarray preparation and hybridization, and expression data acquisition and processing are described in the supplemental data methods.

Data analysis

Unsupervised hierarchical clustering was applied to the entire data set median-centered on genes, using the Pearson correlation as a similarity metric and average linkage clustering. Results were displayed using TreeView [13]. Gene clusters were selected using 10 and 0.6 as minimal gene number and minimal correlation respectively. GoMiner was used to identify functional categories that were over- or underrepresented in specific clusters compared to the list of all analyzed genes [14].

'Predictors' and 'molecular severity score'

LV- and RV-specific data were separated into distinct datasets and analyzed separately using an identical strategy:

The 'Predictor' was defined as a list of genes differentially expressed between Stable and Deteriorating patient groups. These genes were identified using 'Significance Analysis of Microarrays' (SAM) [15] with a maximum false-discovery rate of 1% .

LV and RV predictors were used to calculate a transcriptome-based 'molecular severity score' (MSS) for each sample. First, expression profiles were mean-centered and standard deviation-scaled on genes. The mean profile was calculated for Stable (Ms) and for Deteriorating (Ramakrishnan, Dorris et al.) samples. The molecular severity score (MSS) of a specific sample was defined as the normalized Euclidean squared distance (ranging from 0 to 1) between the sample and the stable mean profile and was calculated as described below:

$$MSS = \frac{Es}{Es + Ed} \text{ where } Es = \sum_{i=1}^{n} [X_i - Ms_i]^2$$

and $\text{Ed} = \sum_{i=1}^{n} [X_i - Md_i]^2$ and X = the expression profile of the specific analyzed sample and i = an index of the n genes included in the 'predictors'.

To define the significance level of the obtained MSS, an unpredictable interval in-between the Stable and Deteriorating profiles was calculated. The cut-offs for the unpredictable interval were defined as the

2.5^{th} and 97.5^{th} percentiles of a random-MSS distribution based on 104 random permutations of the expression profiles.

Leave-one-out cross validation was performed on Stable and Deteriorating samples. For both LV and RV, 50 distinct data sets were produced. Each data set was partitioned into a test set consisting of one sample and a learning set consisting of the 49 other samples. The learning set was used to calculate a MSS using the strategy described. The obtained MSS was employed to predict the MSS value of the test sample. This process was repeated so that the MSS value of each sample was predicted using an MSS estimated from all 49 other samples in the data set.

To test the diagnostic power of our classification, we calculated the sensitivity, the specificity, and the positive and negative predictive values of the molecular prediction of Stable and Deteriorating status using the cross-validation results. To test whether the obtained classification was independent of the method used, we also classified Stable and Deteriorating samples using the Prediction Analysis of Microarrays (PAM) method [16] using a previously described strategy [10, 17] (see supplemental data for detailed methods).

Reproducibility

We tested between-sample reproducibility of the MSS values of all biological duplicates. Expression data from biological duplicates were separated to generate two comparable data sets. MSS from the duplicate sets were compared using the correlation coefficient. Analyses were performed separately on the LV-specific and RV-specific data sets.

Potential biases

We tested whether between-group variations in drug treatment could have biased the predictor discovery. To avoid confounding factors, subgroups of samples from the same chamber and the same severity group were analyzed separately. For each Predictor, expression profiles of samples positive and negative for a specific drug treatment were geneby- gene compared using a student t-test with $p<0.01$. We also tested the predictive power of our predictor in etiology-based and age-based subgroups of patients using the same strategy.

RESULTS

We profiled cardiac gene expression in a cohort of 44 advanced-HF patients using a 4217-oligonucleotide microarray containing genes selected for their involvement in cardiac (patho)physiology. Based on the analysis of clinical information the 44 patients were classified into three HF-severity groups: Deteriorating (n=12), Intermediate (n=19) and Stable (n=13). After raw data extraction and consolidation, 4035 genes were validated for further analysis.

Hierarchical clustering and functional annotation

The 176 cardiac samples and the 4035 selected genes were clustered according to their expression profiles using a hierarchical clustering procedure (Figure 1). Samples were grouped in two major clusters mainly based on the expression profile of a 387-gene cluster (white bar). This patient molecular clustering was not correlated with the clinical severity classification. However, within each of the two major clusters, Stable and Deteriorating samples were preferentially classified into distinct sub-clusters ($p<0.001$ within each major cluster, χ^2 test).

Gene clusters were selected by automated analysis of the gene classification. Functional annotation revealed enrichment of genes involved in a specific biological process or tissue-type for most of the clusters. Clusters that were too small to obtain a statistically significant annotation

using GoMiner software (annotations 'natriuretic peptides' and 'cell metabolism') were functionally annotated based on literature analysis.
Several of the clusters showed marked differential expression between Stable and Deteriorating samples for LV and/or RV samples. 'Cell metabolism', 'natriuretic peptides', and 'extracellular matrix' gene clusters displayed higher expression for Deteriorating samples than for Stable samples in both LV and RV. 'Cytoskeleton' and 'cell death' gene clusters displayed higher expression for Stable samples than for Deteriorating samples in both LV and RV. Interestingly, the 'mitochondrion' gene cluster displayed higher expression for Stable samples than for Deteriorating samples in RV but not LV.

Prediction of clinical status
Two-class statistical analysis of gene expression profiles of the 24 Deteriorating and 26 Stable samples resulted in the identification of a 170-gene and 129-gene predictor for LV and RV respectively (Supplement data Table 2). Sixty-six genes were present in both LV and RV predictors.
Figure 2 shows individual MSS values calculated for the Stable and Deteriorating groups.
The overall good classification rate was 95/100, with one Stable LV sample predicted as Deteriorating and four LV samples in the unpredictable interval. All RV samples were correctly predicted.
A cross-validation strategy was employed to account for data over-fitting due to reclassification of the samples used to define the predictors. The overall good classification rate was 94/100, with one Stable sample predicted as Deteriorating and five samples in the unpredictable interval. The LV molecular predictor identified patients with stable and deteriorating status with a sensitivity of 88% and 92%, and a specificity of 100% and 96%, respectively. The RV molecular predictor identified patients with stable and deteriorating status with a sensitivity of 100% and 96%, and a specificity of 100% and 100%, respectively.
The difference in proportion of samples correctly classified for LV (45/50) and RV (49/50) samples was not statistically significant ($p=0.20$, Fisher exact test). Equivalent prediction power results were obtained when using the PAM prediction method (Supplemental data Table 3).
We also tested the predictive power of a single-gene predictor based on NPPB gene expression levels (supplemental data Figure 1). Using this predictor, a high misclassification rate was observed for deteriorating patients in both LV and RV samples.

Intermediate group analysis
In agreement with the clinical classification, Intermediate samples - which were not used for the construction of the predictors - were on average classified in-between the two other groups (Figure 3). Progression of clinical severity for the LV samples was associated with a gradual increase of the MSS mean values ($p<0.001$ for overall and all pairwise comparisons, one-way analysis of variance on ranks followed by Dunn test). Similar results were observed for RV samples. In addition, we observed that 54 of the 76 Intermediate samples (66%) exhibited MSS values outside the unpredictable interval and could have been predicted as either Stable or Deteriorating.

Effects of potential biases
The three patients groups were comparable regarding sex, age, HF etiology, and LV ejection fraction (Table 1). As expected, differences in severity levels were associated with significant inter-group variations regarding treatment with adrenergic agonists, phospho-diesterase inhibitors, beta-

blockers, and angiotensin converting enzyme inhibitors/angiotensin receptor blockers. Significant differences in expression related to these medications were found for only 0-7.0% of the genes included in the LV and RV predictors. Furthermore, significant differences in expression related to age and HF etiology were found for 1.2 to 2.9% and for 0.6 to 2.3% of the genes respectively. Removing these genes from the predictors did not modify the good classification rates of the samples (data not shown).

Biological reproducibility

We aimed to test whether our classification was reproducible across biological replicates. A significant correlation between MSS values obtained for the duplicate sets was observed (Figure 4), with a better correlation for RV samples than for LV samples.

DISCUSSION

We produced and analyzed the largest set to date of transcriptomal profiles of LV and RV samples from a cohort of 44 HF patients. Ventricular samples were analyzed using a dedicated microarray representing genes selected for their contribution to cardiovascular (patho)physiology. Replication at both the biological and the technical level, and control of experimental variations at the different steps of the study allowed detection of even subtle expression changes. We identified a set of genes of which expression changes discriminated between patients with different clinical severity levels and established that clinical deterioration of HF patients was associated with a molecular deterioration expression profile in both LV and RV. Therefore, our study confirms the potential of cardiac gene expression profiling to predict outcomes in patients with advanced HF.

Related findings in previous studies

It has previously been shown that gene expression profiling can discriminate between cardiac patients with different clinical characteristics [7-11, 18]. Etiology-related gene expression profiles have been identified in Chagas disease, and hypertrophic, dilated, viral, and ischemic cardiomyopathies [9, 10, 19]. In a recent study, Heidecker et al. identified a transcriptomic signature that could predict clinical outcome of new-onset idiopathic dilated cardiomyopathy patients [17]. Taken together, these findings offer valuable information regarding the molecular basis of HF related to distinct etiologies and they could lead to individualized therapeutic strategies in HF.

Other clinical characteristics such as age and sex have also been shown to have an effect on the transcriptomal profile of HF patients [20]. In our study, the molecular severity markers correctly classified HF patients independent of etiology or age. Because most of our patients were male, we could not validate our classification in female HF patients. We also showed that our results were unchanged when another prediction method was used [10, 17].

Potential clinical significance of findings

Prognosis evaluation for advance HF patients.

The results of our study suggest that gene expression profiling has the potential to detect HF patients at highest risk of death with high sensitivity and specificity. Prognosis evaluation is fundamental for the indication of LVAD implantation and heart transplantation in advanced HF patients. Patients depending on intravenous inotropic therapy have the worst prognosis and should benefit from urgent or elective LVAD implantation or urgent transplantation whenever possible [21]. However, risk stratification remains particularly difficult for ambulatory advanced HF patients not

depending on intravenous inotropic therapy or prolonged hospitalization, with major impairment of their functional capacities and poor survival [22, 23]. Specific risk scores are not yet available for advanced HF patients but become mandatory in the context of this growing cohort of patients [24, 25]. Therefore we propose that, in advanced heart failure, molecular prediction based on endomyocardial biopsies could be a valuable strategy for patient risk stratification.

Our results showed that the LV and RV predictors lead to a better prediction of clinical status than the NPPB predictor, in particular regarding the prediction of the deteriorating status. It has previously been shown that the NPPB mRNA level in the left ventricle and the BNP peripheral blood level are correlated [26]. The B-type Natriuretic Peptide (BNP) blood level is widely used as a clinical predictor for HF patients. However BNP blood level predictive value is still controversial in the specific condition of end-stage HF [27, 28]. A previous report showed that a lower natriuretic peptide blood level, that usually implies a better outcome, may also imply poor outcome in severe HF patients [28]. Similarly, our results show a low NPPB mRNA level for a patient sub-group in the Deteriorating status group.

Transcriptomal remodeling of the right ventricle.

Our results suggest that molecular prediction using samples taken from RV may be as powerful as molecular prediction using samples taken from LV. Most of the patients with advanced HF have severe LV dysfunction, whereas RV dysfunction intensity is variable among these patients. In addition, transcriptome remodeling of the RV in HF has been evaluated to a lesser extent than for the LV. Our data show that most of the molecular processes disturbed in the LV are also disturbed in the RV. In addition, sensitivity and specificity of prediction of both Stable and Deteriorating statuses using RV samples were at least equivalent with those obtained using LV samples. These results are in agreement with a previous study showing accurate prediction of clinical outcome of new-onset HF patients using a transcriptomic signature obtained from RV endomyocardial biopsies [17].

Prediction reproducibility

Measurement reproducibility is another crucial point when developing a predictor of HF severity. Relatively high variability of widely used biomarkers like BNP or N-terminal proBNP blood levels may be a problem for patient management [29]. Our results show that gene expression profiling is reproducible among biological replicates. Reproducibility was higher for RV samples, reinforcing the interest of RV sample utilization to develop a molecular predictor in advanced HF. A hypothesis is that regional tissue heterogeneity may be higher in LV than in RV. One cause may be the presence of infarct scars that preferentially affect the LV. However, ventricular samples analyzed in this study were obtained after careful dissection of the ventricles excluding infarct scars. We also did not observe a higher variability of MSS values obtained for LV samples in patients affected by coronary artery disease compared to other patients.

Potential limitations

Effect of medication

Therapeutic interventions, in particular medications, may induce modifications of the cardiac transcriptome [30]. We tested the hypothesis that the patient classification may be modified by angiotensin converting enzyme inhibitors/angiotensin receptor blockers, beta-blockers, and inotropic drugs. A very low number of genes included in the distinct predictors displayed differential expression associated with different drug

intake. Removing these genes from the predictors did not modify the patients' classification. Therefore, medications do not strongly modify the expression level of our predictors.

Clinical classification

We compared our molecular predictors to a 2-parameter clinical classification that has not been previously evaluated in advanced HF. Because we used samples taken at the time of cardiac transplantation, it was not possible to compare our predictors to a relevant clinical end-point like mortality or hospitalization for ADHF. We hypothesized that the use of parameters measured at the time of transplantation would better reflect the clinical phenotype at this time and decided to combine two established predictors of HF severity to classify patients. The UNOS medical urgency status has been specifically developed for advanced HF patients listed for cardiac transplantation. The UNOS-1A status at the time of listing is associated with a 1 month-mortality >30% whereas UNOS-2 patients have a 1 monthmortality <10% [4]. The mortality rate on the UNOS waiting list is more than 4 fold higher for UNOS-1A than for UNOS-2 patients [23]. To better define our group of Stable patients we combined the UNOS medical urgency status with the occurrence of ADHF episodes. Frequent rehospitalizations have been recognized as a strong predictor of HF patient mortality [24]. Other HF severity prediction scores have been developed in advanced HF [3, 4]. Comparison of one of these HF severity predictors to the UNOS medical urgency status did not reveal a higher predictive power [4]. Other predictors included the measurement of peak oxygen consumption that cannot be recorded in the most severely affected patients [3].

We analyzed expression profiles of patients with advanced HF at the time of cardiac transplantation. Further clinical studies are needed to determine whether gene expression profiling of cardiac tissue provides sensitive prognostic information for advanced ambulatory HF patients using clinical end-points like mortality or hospitalization for HF.

Acknowledgments

The authors thank the thoracic surgery and cardiology departments of the Nantes University Hospital for their participation. Funding was provided by the "Institut National de la Santé et de la Recherche Médicale" (INSERM), the "Centre National de la Recherche Scientifique" (CNRS), "Ouest Génopole", the "Association Française contre les Myopathies" (AFM), and the "Region Pays de la Loire". GL was supported by the "Fondation pour la Recherche Médicale".

Références :

1. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *JAMA* 2003;290:2581-7.

2. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation* 2006;113:1424-33.

3. Aaronson KD, Schwartz JS, Chen TM, Wong KL, Goin JE, Mancini DM. Development and prospective validation of a clinical index to predict survival in ambulatory patients referred for cardiac transplant evaluation. *Circulation* 1997;95:2660-7.

4. Smits JM, Deng MC, Hummel M, De Meester J, Schoendube F, Scheld HH, Persijn GG, Laufer G, Van Houwelingen HC. A prognostic model for predicting

waiting-list mortality for a total national cohort of adult heart-transplant candidates. *Transplantation* 2003;76:1185-9.
5. Francis GS. Pathophysiology of chronic heart failure. *Am J Med* 2001;110 Suppl 7A:37S-46S.
6. Swynghedauw B. Molecular mechanisms of myocardial remodeling. *Physiol Rev* 1999;79:215-62.
7. Blaxall BC, Tschannen-Moran BM, Milano CA, Koch WJ. Differential gene expression and genomic patient stratification following left ventricular assist device support. *J Am Coll Cardiol* 2003;41:1096-106.
8. Kaynak B, von Heydebreck A, Mebus S, Seelow D, Hennig S, Vogel J, Sperling HP, Pregla R, exi-Meskishvili V, Hetzer R, Lange PE, Vingron M, Lehrach H, Sperling S. Genome-wide array analysis of normal and malformed human hearts. *Circulation* 2003;107:2467-74.
9. Liew CC, Dzau VJ. Molecular genetics and genomics of heart failure. *Nat Rev Genet* 2004;5:811-25.
10. Kittleson MM, Ye SQ, Irizarry RA, Minhas KM, Edness G, Conte JV, Parmigiani G, Miller LW, Chen Y, Hall JL, Garcia JG, Hare JM. Identification of a gene expression profile that differentiates between ischemic and nonischemic cardiomyopathy. *Circulation* 2004;110:3444-51.
11. Steenman M, Lamirault G, Le Meur N, Le Cunff M, Escande D, Leger JJ. Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays. *Eur J Heart Fail* 2005;7:157-65.
12. Renlund DG, Taylor DO, Kfoury AG, Shaddy RS. New UNOS rules: historical background and implications for transplantation management. United Network for Organ Sharing. *J Heart Lung Transplant* 1999;18:1065-70.
13. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;95:14863-8. 14. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 2003;4:R28.
15. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001;98:5116-21. 16. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* 2002;99:6567-72.
17. Heidecker B, Kasper EK, Wittstein IS, Champion HC, Breton E, Russell SD, Kittleson MM, Baughman KL, Hare JM. Transcriptomic biomarkers for individual risk assessment in new-onset heart failure. *Circulation* 2008;118:238-46.
18. Kaab S, Barth AS, Margerie D, Dugas M, Gebauer M, Zwermann L, Merk S, Pfeufer A, Steinmeyer K, Bleich M, Kreuzer E, Steinbeck G, Nabauer M. Global gene expression in human myocardium-oligonucleotide microarray analysis of regional diversity and transcriptional regulation in heart failure. *J Mol Med* 2004;82:308-16.
19. Wittchen F, Suckau L, Witt H, Skurk C, Lassner D, Fechner H, Sipo I, Ungethum U, Ruiz P, Pauschinger M, Tschope C, Rauch U, Kuhl U, Schultheiss HP, Poller W. Genomic expression profiling of human inflammatory cardiomyopathy (DCMi) suggests novel therapeutic targets. *J Mol Med* 2007;85:257- 1.
20. Boheler KR, Volkova M, Morrell C, Garg R, Zhu Y, Margulies K, Seymour AM, Lakatta EG. Sex- and age-dependent human transcriptome variability: implications for chronic heart failure. *Proc Natl Acad Sci U S A* 2003;100:2754-9.

21. Rogers JG, Butler J, Lansman SL, Gass A, Portner PM, Pasque MK, Pierson RN, III. Chronic mechanical circulatory support for inotrope-dependent heart failure patients who are not transplant candidates: results of the INTrEPID Trial. *J Am Coll Cardiol* 2007;50:741-7.
22. Lietz K, Long JW, Kfoury AG, Slaughter MS, Silver MA, Milano CA, Rogers JG, Naka Y, Mancini D, Miller LW. Outcomes of left ventricular assist device implantation as destination therapy in the post-REMATCH era: implications for patient selection. *Circulation* 2007;116:497-505.
23. Deng MC, Smits JM, Packer M. Selecting patients for heart transplantation: which patients are too well for transplant? *Curr Opin Cardiol* 2002;17:137-44.
24. Metra M, Ponikowski P, Dickstein K, McMurray JJ, Gavazzi A, Bergh CH, Fraser AG, Jaarsma T, Pitsis A, Mohacsi P, Bohm M, Anker S, Dargie H, Brutsaert D, Komajda M. Advanced chronic heart failure: A position statement from the Study Group on Advanced Heart Failure of the Heart Failure Association of the European Society of Cardiology. *Eur J Heart Fail* 2007;9:684-94.
25. Stevenson LW, Couper G. On the fledgling field of mechanical circulatory support. *J Am Coll Cardiol* 2007;50:748-51.
26. Hystad ME, Geiran OR, Attramadal H, Spurkland A, Vege A, Simonsen S, Hall C. Regional cardiac expression and concentration of natriuretic peptides in patients with severe chronic heart failure. *Acta Physiol Scand* 2001;171:395-403.
27. Potapov EV, Hennig F, Wagner FD, Volk HD, Sodian R, Hausmann H, Lehmkuhl HB, Hetzer R. Natriuretic peptides and E-selectin as predictors of acute deterioration in patients with inotrope-dependent heart failure. *Eur J Cardiothorac Surg* 2005;27:899- 905.
28. Miller WL, Burnett JC, Jr., Hartman KA, Henle MP, Burritt MF, Jaffe AS. Lower rather than higher levels of B-type natriuretic peptides (NT-pro-BNP and BNP) predict short-term mortality in end-stage heart failure patients treated with nesiritide. *Am J Cardiol* 2005;96:837-41.
29. Bruins S, Fokkema MR, Romer JW, Dejongste MJ, van der Dijs FP, van den Ouweland JM, Muskiet FA. High intraindividual variation of B-type natriuretic peptide (BNP) and amino-terminal proBNP in patients with stable chronic heart failure. *Clin Chem* 2004;50:2052-8
30. Lowes BD, Gilbert EM, Abraham WT, Minobe WA, Larrabee P, Ferguson D, Wolfel EE, Lindenfeld J, Tsvetkova T, Robertson AD, Quaife RA, Bristow MR. Myocardial gene expression in dilated cardiomyopathy treated with beta-blocking agents. *N Engl J Med* 2002;346:1357-65.

Figures/Tables

Table 1: Clinical characteristics of HF severity patient groups

	Stable n=13	Deteriorating n=12	Intermediate n=19	p-value
Male / Female	12/1	10/2	16/3	0.747
Age, years	50 (15)	49 (9)	48 (12)	0.559
Initial cardiac disease, CAD / DCM / other	5 / 6 / 2	4 / 7 / 1	8 / 7 / 4	0.840
HF duration, months	32 (29)	24 (33)	29 (32)	0.459
Heart rate, min^{-1}	69 (13)	100 (16)	76 (16)	<0.001 * 3
Systolic arterial pressure, mmHg	102 (17)	97 (9)	103 (12)	0.509
LVEF, %	24 (11)	22 (7)	24 (7)	0.829
LVEDD, mm	73 (13)	66 (7)	65 (10)	0.254
Blood urea nitrogen, mmol/l	9.1 (5.6)	9.8 (4.2)	9.0 (4.1)	0.884
Serum Creatinine, µmol/l	107 (26)	107 (22)	101 (36)	0.642
Medications, % of patients				
ACEI / ARB	100	58	84	0.024 *
Beta-Blockers	69	0	26	<0.001 * 2
Adrenergic agonists	0	100	42	<0.001 * 2 3
Phosphodiesterase Inhibitors	0	67	0	<0.001 * 3
Aldosterone blockers	77	58	53	0.420
Statin	46	33	32	0.724
Digoxin / Digitoxin	46	25	26	0.502
UNOS medical urgency status	**2**	**1A**	**1B or 2**	
Number of recent ADHF episodes	**0**	**1.8 (0.4)**	**1.7 (0.8)**	

CAD: Coronary Artery Disease; DCM: Dilated Cardiomyopathy; LVEF: Left Ventricle Ejection Fraction; LVEDD: Left Ventricle End Diastolic Diameter; ACEI: Angiotensin Converting Enzyme Inhibitors; ARB: Angiotensin Receptor Blockers; ADHF: Acute Decompensated Heart Failure. Data are presented as 'mean (SD)' when appropriate. P-value indicates the result of a comparison between the three patient groups using Fisher's exact test or Kruskal-Wallis rank sum test. If $p<0.05$, groups were compared two-by-two. *: $p<0.05$ between Deteriorating and Stable; 2: $p<0.05$ between Intermediate and Stable; 3: $p<0.05$ between Deteriorating and Intermediate. An ADHF episode was defined as recent if it occurred during the 3 months before the heart transplantation/total artificial heart placement. HF duration was defined as the delay between onset of HF symptoms and heart transplantation/total artificial heart placement. Values for LVEF, LVEDD, blood urea nitrogen, and serum creatinine corresponded to pre-operative measurements. All patients were treated with loop diuretics (furosemide and/or bumetanide). Only medications related to HF therapy are presented. The clinical profile was determined based on the patients' medical urgency status in the UNOS classification and the occurrence of recent ADHF episodes.

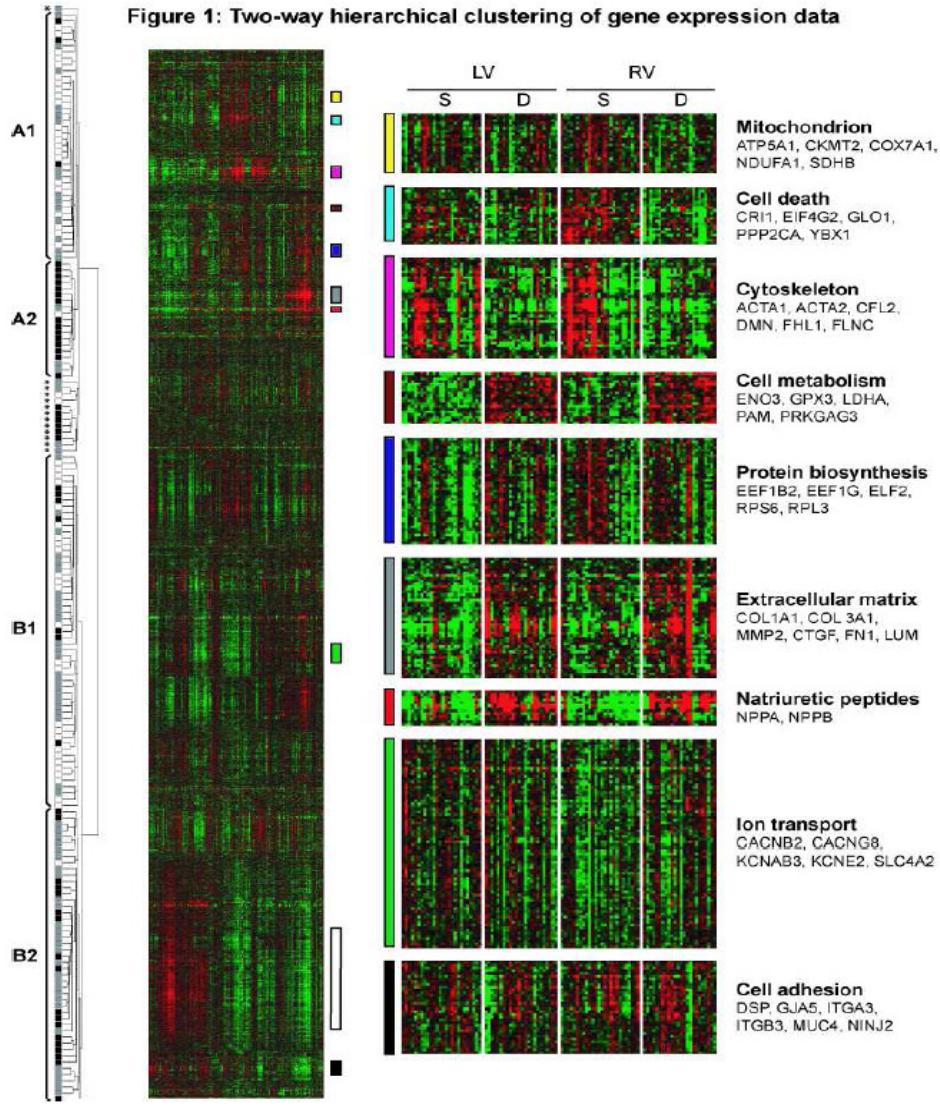


Figure 1: Two-way hierarchical clustering of gene expression data

Left: Classification tree of the samples. The dendrogram is based on similarity of the gene expression profiles of the 176 analyzed samples. Samples were separated into 4 main clusters (A1, A2 and B1, B2). Only clusters containing at least 15 samples were considered as significant. Some samples (indicated *) were not included in any cluster. White, grey, and black boxes on the left side of the dendrogram denote Stable, Intermediate, and Deteriorating clinical status respectively.

Middle: Heat map of expression values for 176 samples and 4035 genes after hierarchical clustering of both genes and samples. Each column represents the 4035-gene expression profile for one sample. Each row represents the 176-sample expression profile for one gene. Results are presented using a color code. Green and red represent lower and higher expression levels relative to the median expression level of the gene, respectively.

Right: Selected gene clusters indicated by colored bars in the middle part of the figure. Intermediate samples were removed and remaining samples were ordered based on their origin (LV: left ventricle, RV: right ventricle) and the clinical status of the patient (S: Stable, D: Deteriorating). On the right side, functional annotation of the clusters is shown. Some genes representative of the functional annotation of the cluster are indicated using their HUGO gene nomenclature committee symbol.

Figure 2: Prediction of HF severity based on gene expression profiles

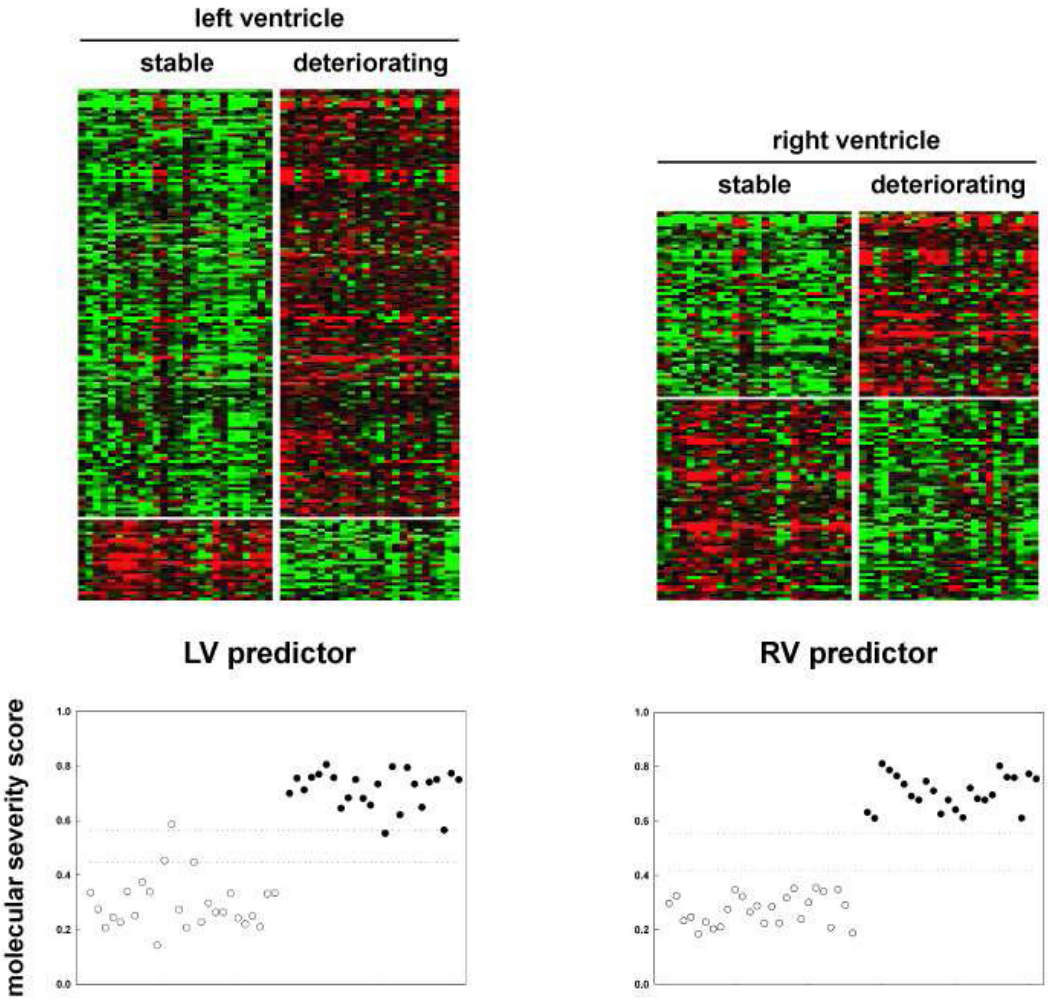


Figure 2: Prediction of HF severity based on gene expression profiles.
Top: Gene expression profiles of Stable and Deteriorating samples for the LV and RV severity predictors. Each column represents the gene expression profile for one sample. Each row represents the relative expression level for one gene. Color code as in Figure 1.
Bottom: patient classifications for the LV and RV severity predictors.
Open and filled circles correspond to Stable and Deteriorating LV samples respectively. Open and filled triangles correspond to Stable and Deteriorating RV samples respectively. Dashed lines denote upper and lower limits of the unpredictable interval.

Figure 3: Prediction of HF severity in all samples

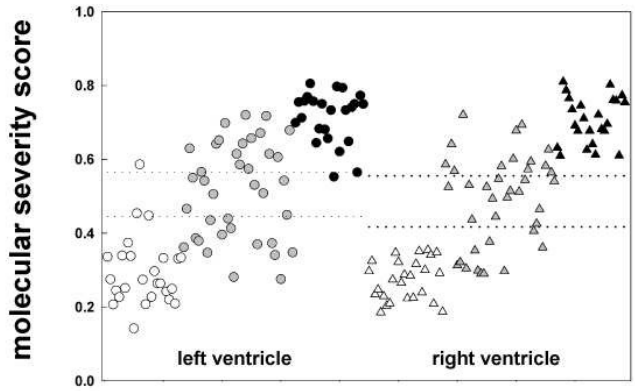


Figure 3: Prediction of HF severity in all samples.
Individual MSS values obtained for the LV and RV predictors are presented for all 176 analyzed samples. Open and black-filled circles correspond to Stable and Deteriorating LV samples respectively. Open and black-filled triangles correspond to Stable and Deteriorating RV samples respectively. Intermediate samples are shown in gray. Dashed lines denote upper and lower limits of the unpredictable interval.

Figure 4: Between-sample reproducibility

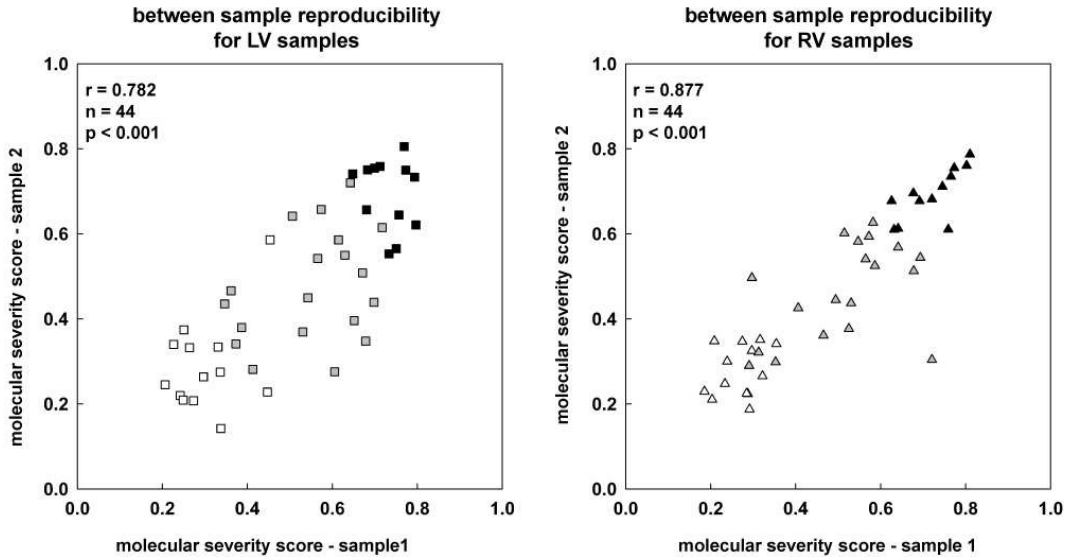


Figure 4: Between-sample reproducibility.
Between-sample reproducibility was assessed using MSS values calculated from biological replicates. Subgroup analysis based on the origin of the sample (LV or RV) is shown. The correlation coefficient was used as a between-sample reproducibility index. Squares: LV samples; Triangles: RV samples. Open symbols: Stable samples; grey-filled symbols: Intermediate samples; black-filled symbols: Deteriorating samples

Génomique des pathologies neuromusculaires

(ENMG 2008 - 16es Journées Francophones d'Electroneuromyographie", Péréon Y, Solal Ed., Marseille, 2008, p. 25-39)

Daniel Baron [1,2], Solenne Carat [1,2], Mahatsangy Raharijaona [1,2], Rémi Houlgatte [1,2,3,4]

1 : INSERM, U533, Nantes, F-44000 France.
2 : Université de Nantes, Faculté de Médecine, l'institut du thorax, Nantes, F-44000 France.
3 : CHU Nantes, l'institut du thorax, F-44000 France.
4 : remi.houlgatte@nantes.inserm.fr

1- Introduction

Les puces à ADN (1) représentent actuellement la technologie d'analyse à grande échelle la plus utilisée actuellement puisqu'elles permettent l'étude des profils d'expression de plusieurs milliers de gènes, au sein des échantillons. Cette technologie est devenue incontournable dans l'étude massive d'expression des gènes. Ce type d'étude correspond à une véritable révolution tant par la quantité de résultats générés que par la modification de pensée qu'elle offre.

Les premières publications d'étude du transcriptome par les puces à ADN datent de 1995, par l'équipe de Pat Brown (2;3), qui a démontré qu'en analysant simultanément l'ensemble des transcrits d'un tissu ou d'une cellule, il devient alors possible d'obtenir des signatures moléculaires caractéristiques de certains états cellulaires. Ces groupes de gènes co-exprimés (ou « clusters ») au travers des échantillons sous-tendent la même fonction biologique (4) au travers de plusieurs études (5), sont co-localisés (6) et co-régulés (7).

La technologie des puces à ADN est aussi celle qui a connu les développements récents les plus importants, et a amené les avancées les plus notables en cancérologie (8-13) pour la classification des tumeurs, la découverte de nouveaux marqueurs biologiques ou encore l'établissement de sous-groupes de patients à visée pronostique.

Plus récemment, l'utilisation des puces à ADN pour l'étude du transcriptome musculaire s'est généralisée (14-18). Mais il a fallu attendre 2000 pour que ne soit publiée pour la première fois par l'équipe de Hoffman (19), l'étude du transcriptome d'une pathologie musculaire : la dystrophie musculaire de Duchenne (DMD) causée par une mutation du gène de la dystrophine (20;21). Ces premiers travaux ont ouvert la voie à de nombreuses études de transcriptome des pathologies neuromusculaires.

2 - Apport du transcriptome à la compréhension des pathologies neuromusculaires

2.1 - Le muscle malade diffère du muscle sain

De nombreuses études ont utilisé les puces à ADN pour comparer le transcriptome du muscle sain à celui du muscle malade. Elles montrent que les différences d'expression observées et les fonctions biologiques associées à ces gènes retracent le remodelage cellulaire sous-jacent à la pathologie.

Par exemple, chez les patients DMD (voir Fig. 1), plusieurs études (19;22-25) ont montré que les altérations du transcriptome reflètent les événements d'inflammation, de fibrose, de nécrose des myocytes …observés dans les biopsies de patients présentant déjà les symptômes de la maladie (âge > 5 ans) et ce, malgré la variabilité individuelle des malades (26). Ces mêmes altérations du transcriptome sont aussi observées chez des enfants (âge < 2 ans) ne présentant pas encore les symptômes de la maladie (27) et la réponse inflammatoire pourrait être l'un des signes les plus précoce de la maladie (28), associé à un processus intense de régénération / dégénérescence des myofibres (26).

Des résultats semblables sont aussi observés dans la dystrophie musculaires des ceintures de type 2B (LGMD2B) avec une sur-expression de gènes impliqués dans la réponse immunitaire, le transport du calcium ou le développement musculaire en accord avec l'inflammation musculaire, l'instabilité des membranes et la régénérations des fibres (29). Dans la

Dystrophie Facio-Scapulo-Humérale (FSHD), les profils d'expression révèlent que le muscle atteint passe d'un phénotype de muscle à fibre rapides à un phénotype lent (30) avec une sur-expression de gènes impliqués dans la néo-vascularisationon progressive du muscle (31).

Dans le cas de myosites (MIs), les gènes du complexe majeur d'histocompatibilité de Classe I et II, ceux impliqués dans la réponse immunitaire (32-34) et ceux codants pour les immunoglobulines sont sur-exprimés en accord avec l'infiltration du muscle par des lymphocytes B différenciés (35).

Pour les myopathies congénitales (CMDs) de type Fukuyama (FCMD) ou avec déficit en mérosine (MDC1A) , une sur-expression de gènes codants pour des protéines de la matrice extracellulaire est observée signant une fibrose interstitielle (36) accompagnée d'une sous expression de gènes impliqués dans la contraction musculaire, signant la régénération moins rapide (ou dégénérescence) des fibres musculaires (37). Des résultats très similaires sont observés dans d'autres CMDs – myopathie myotubulaire liée à l'X (XLMTM) (38) et myopathie à déficit de némaline (Steenman, Lamirault et al.) (39) – avec en plus une sur-expression de gènes du cytosquelette et une sous-expression de gènes impliqués dans le métabolisme du glucose associés au remodelage intra et extra-cellulaire.

L'analyse du transcriptome permet de mettre aussi en évidence des gènes ou des fonctions, dont l'expression n'est pas suspectée *a priori* dans la pathologie comme par exemple la sur-expression de gènes cardiaques (19;22), l'implication de la voie des facteurs de croissance de type insulin-like, l'infiltration de cellulaire dendritiques immunitaires (19;22;23) dans le muscle DMD.

2.2 - Comparaisons des pathologies entre elles : signatures spécifiques

La comparaison des résultats obtenus pour plusieurs pathologies musculaires donne un aperçu des spécificités de chacune des pathologies ainsi que les réponses moléculaires et cellulaires communes.

La mise en évidence de signatures communes à plusieurs pathologies reflète souvent des conséquences secondaires de la pathologie. Ces signatures communes dépendent des pathologies prises en compte dans la comparaison. Par exemple, en accord avec les remodelages cellulaires observés, les gènes surexprimés dans le muscle DMD ou atteint l'alpha-sarcoglycanopathie (α-SGD) sont impliqués dans la réponse immunitaire ou codent pour des protéines de la matrice extracellulaire ; les gènes sous-exprimés sont eux impliqués dans le métabolisme énergétique et la fonction mitochondriale (19). Les gènes dérégulés à la fois dans le muscle DMD ou atteint d'une myosite infantile (JDM) reflètent plutôt les processus de dégénérescence et de régénération musculaire dans les deux maladies (40). Chez ces derniers (JDM) des marqueurs de l'inflammation et de la réponse immunitaires (immunoglobulines, récepteurs aux lymphocytes T) sont aussi retrouvés sur-exprimés chez les patients atteints d'une autre myosite (sIBM) (32;40). Plus récemment, l'inflammation et la nécrose ont été identifiées comme étant les fonctions biologiques communes dans une étude couvrant 13 pathologies distinctes (41).

La mise en évidence des signatures spécifiques de chacune des pathologies permet de une classification moléculaire des échantillons (42) au moins aussi efficacement que leur classification sur de simples critères biochimiques et génétiques (32;41). Certaines de ces signatures peuvent être masquées par les signatures communes à plusieurs pathologies (36). Leur mise en évidence peut se faire en filtrant les gènes dérégulés dans plusieurs pathologies (31;40;43).

Ces signatures spécifiques reflètent les voies de régulation transcriptionnelles impliquées dans la pathologie. Par exemple, dans leur étude sur les myopathies inflammatoires incluant 45 biopsies musculaires, Greenberg et al. (32) ont mis en évidence un profil d'expression moléculaire permettant de distinguer les muscles MIs du muscle DMD et du muscle sain. Ils montrent qu'au sein des MIs, les patients atteints de dermatomyositis (DM) sont caractérisés par la sur-expression de gènes de réponse à l'interféron (44) aussi bien chez l'adulte (32) que chez l'enfant (40).

Bien que l'analyse du transcriptome mette en évidence des différences d'expression entre des tissus peu (27) ou pas différenciés histologiquement (45), la mise en évidence, au sein d'une même pathologie, de sous-classe de sous-classes distinctes par le transcriptome semble plus difficilement réalisable

comme l'attestent certaines études (34;41). Ces résultats pourraient rendre compte de la technologie, du faible nombre d'échantillons étudiés pour une pathologie donnée, de la variabilité génétique des patients (SNP) (26;46), du type de prélèvement (39) ou d'autres facteurs comme l'âge (30). L'existence de sous classes histo-cliniques a cependant pu être validée par le transcriptome dans le cas des MIs (32), d'encéphalomyopathie mitochondriale (47) ou de NM (42), un groupe que l'on pensait homogène auparavant (39). La discrimination de telles sous-classes par le transcriptome pourrait aussi corréler avec des critères plus discrets comme le score de sévérité pathologique (29) ou le génotype (30) de patients.

2.3- Effet de molécules

L'effet de molécule -à visée thérapeutique- sur le muscle peut être aussi appréhendé par la mesure du transcriptome et devrait connaître un essor considérable dans les années à venir. Trois études ont démontré par exemple que l'insuline, l'épinéphrine ou les hormones thyroïdiennes stimulent les gènes impliqués dans le métabolisme énergétique dans le muscle sain (15;48;49) par des mécanismes distincts. Dans le muscle DMD, l'effet anabolique d'un analogue synthétique de la testostérone diminuerait la dégénérescence des fibres musculaires (50). Chez les patients MI, un traitement aux immunoglobulines entrainerait des différences substantielles de l'expression de chémokines (33).

3 – autres niveaux d'analyse en génomique : intégration des données

Bien que les études de transcriptome aient amélioré notre compréhension des mécanismes sous-jacents aux pathologies musculaires (51), d'autres niveaux d'études peuvent être envisagés (voir Fig. 2).

3.1 – transcriptome

- Méta-analyses

La comparaison de résultats (ou méta-analyse (52-54)) d'études indépendantes de transcriptome musculaire est réalisable (55) et permettrait de valider des clusters de gènes dérégulés pour une même pathologie (voir Fig. 3) ou des signatures hautement spécifiques d'une pathologie.

Bien que majoritairement composé de fibres musculaires, le muscle est un tissu complexe composé de plusieurs types cellulaires dont les proportions relatives diffèrent selon la sévérité de l'état pathologique. La méta-analyse systématique en combinant des résultats obtenus sur des lignées cellulaires ou reflétant le transcriptome musculaire dans diverses conditions physiopathologiques (atrophie, âge, effet de molécules…) permettrait non seulement d'affiner les fonctions biologiques inférées par les gènes dérégulés, mais aussi de mettre en évidence les transcrits spécifiques d'un type cellulaire (6;56) par micro-dissection virtuelle. Par exemple, la comparaison du transcriptome du muscle avec celui d'autres organes permet de mettre en évidence des gènes d'expression spécifiquement musculaire (57), voire d'expression modulée dans des pathologies musculaires (25).

Parmi les gènes d'expression dérégulée dans les pathologies, l'identification des facteurs de transcriptions est une étape indispensable vers la compréhension des mécanismes de régulation transcriptionnelle sous-jacents au remodelage cellulaire. Par exemple, les facteurs de transcription exprimés dans les cellules immunitaires sont sur-exprimés dans le muscle DMD, alors que ceux exprimés dans les fibres musculaires (voir Fig. 4) sont sous-exprimés (23). En effet, la co-expression (5) et la co-localisation (6) des transcrits au sein d'un tissu ou d'une population de cellules, implique une régulation trancriptionnelle commune (7) des gènes. L'identification de ces facteurs de transcription et la découverte (ou recherché) de motifs de liaison de ces facteurs dans les promoteurs dès gènes permet de mettre en évidence de tels modules de régulation (58-60) comme la mise en évidence de la dérégulation de MYOD et de ses cibles potentielles dans le muscle FSHD (30)

- Micro ARNs

Les Micro ARNs (miRNAs) sont des ARNs simple-brin non codants, répresseurs des mécanismes post-transcriptionnels. Plusieurs centaines de

miRNAs ont été identifié chez les mammifères, et bon nombre d'entre eux sont spécifiques d'un tissu et/ou temporellement régulés dans leur expression (61). La fonction de seulement une petite fraction d'entre ceux-ci a été décrite en détail mettant en évidence leur implication dans une variété de processus physiologiques liés au développement (62;63), en particulier dans le développement musculaire. Certains de ces miRNAs pourraient de plus être impliqués fortement dans les pathologies musculaires puisque dans une étude récente, 185 miRNAs ont été montrés comme étant spécifiquement sur ou sous-exprimés dans 10 pathologies musculaires majeures (64).

3.2 - génome

- Polymorphisme génétique

Trouver des associations de SNPs (polymorphisme simple nucléotide) avec un risque accru de pathologie représente un enjeu majeur de la génétique (65;66). Ceci implique de déterminer pour des millions de loci, quel nucléotide polymorphe est présent (déséquilibre de liaison) dans un groupe de malades comparé aux individus sains. En effet, plus de 5 millions de SNPS, ayant une fréquence dans la population supérieure à 10%, sont prédits dans le génome humain (66;67). Ces problèmes peuvent être abordés à l'aide des puces à SNPs (68;69) avec succès dans les pathologies musculaires (70) pour lesquelles l'influence de plusieurs gènes modulateurs (71) ou de variants rares (42) est souvent suspectée.

- Chromatine et régulation transcriptionnelle

Le ChIP-chip (72;73) permet d'étudier à grande échelle, les sites de fixations d'un facteur de transcription à l'ADN (voir Fig. 5). Il s'agit d'immunoprécipiter (ChIP : Chromatine ImmunoPrecipitation) le facteur de transcription étudié en même temps que les fragments d'ADN sur lesquels il est fixé. Ces segments d'ADN sont ensuite caractérisés par une puce à ADN (chip) constituées de sondes des régions promotrices, des îlots CpG ou couvrant l'ensemble du génome. Cette technique, particulièrement intéressante car ne modifiant pas le système étudié, a déjà permis de mettre en évidence le réseau de régulation transcriptionnelle qui gouverne la différenciation myogénique des cellules (74;75) et l'inhibition de ces mêmes gènes par le complexe répresseur Polycomb 2 (PCR2) dans les cellules souches embryonnaires (76). Des résultats plus récents obtenus sur du muscle cardiaque (77) montrent que cette technique est aussi applicable à tout type d'échantillon biologique.

3.3 – Protéome

Les validations de protéines peuvent se faire à petite (45) ou à une grande échelle (30). Toutefois, à l'heure actuelle, il n'existe pas d'équivalent des puces à ADN pour les protéines, en raison des contraintes technologiques importantes. Une méthode originale (78;79), le tissue microarray (TMA), pourrait néanmoins être systématiquement envisagée pour valider facilement, à moyenne échelle, les marqueurs protéiques dans tous les échantillons étudiés.

4- Conclusion

L'intégration de l'ensemble de ces données permettra de mettre en évidence les réseaux de régulation transcriptionnelle sous-jacents aux pathologies neuromusculaires (voir Fig. 6). La connaissance de ces réseaux permettra de prédire l'effet d'une molécule (ou d'un polymorphisme) sur les facteurs clés du réseau mais aussi de définir de nouvelles cibles thérapeutiques raisonnables en distinguant les causes des conséquences. Ces connaissances devraient à terme permettre une médecine personnalisée en tenant compte de la variabilité de chaque patient.

Reference List

1. **Lockhart DJ, Winzeler EA** 2000 Genomics, gene expression and DNA arrays 1. Nature 405:827-836

2. **Schena M, Shalon D, Davis RW, Brown PO** 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray 3. Science 270:467-470

3. **Shalon D, Smith SJ, Brown PO** 1996 A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6:639-645

4. **Eisen MB, Spellman PT, Brown PO, Botstein D** 1998 Cluster analysis and display of genome-wide expression patterns
1. Proc Natl Acad Sci U S A 95:14863-14868

5. **Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P** 2004 Coexpression analysis of human genes across many microarray data sets
1. Genome Res 14:1085-1094

6. **Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de RM, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO** 2000 Systematic variation in gene expression patterns in human cancer cell lines
1. Nat Genet 24:227-235

7. **Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM** 1999 Systematic determination of genetic network architecture
1. Nat Genet 22:281-285

8. **Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM** 2000 Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling
4. Nature 403:503-511

9. **Bertucci F, Houlgatte R, Benziane A, Granjeaud S, Adelaide J, Tagett R, Loriod B, Jacquemier J, Viens P, Jordan B, Birnbaum D, Nguyen C** 2000 Gene expression profiling of primary breast carcinomas using arrays of candidate genes
3. Hum Mol Genet 9:2981-2991

10. **Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V** 2000 Molecular classification of cutaneous malignant melanoma by gene expression profiling
2. Nature 406:536-540

11. **Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES** 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring
3. Science 286:531-537

12. **Hedenfalk IA, Ringner M, Trent JM, Borg A** 2002 Gene expression in inherited breast cancer. Adv Cancer Res 84:1-34

13. **Kim CJ, Reintgen DS, Yeatman TJ** 2002 The promise of microarray technology in melanoma care
1. Cancer Control 9:49-53

14. **Bodine SC, Latres E, Baumhueter S, Lai VK, Nunez L, Clarke BA, Poueymirou WT, Panaro FJ, Na E, Dharmarajan K, Pan ZQ, Valenzuela DM, DeChiara TM, Stitt TN, Yancopoulos GD, Glass DJ** 2001 Identification of ubiquitin ligases required for skeletal muscle atrophy
1. Science 294:1704-1708

15. **Clement K, Viguerie N, Diehn M, Alizadeh A, Barbe P, Thalamas C, Storey JD, Brown PO, Barsh GS, Langin D** 2002 In vivo regulation of human skeletal muscle gene expression by thyroid hormone
1. Genome Res 12:281-291

16. **Gomes MD, Lecker SH, Jagoe RT, Navon A, Goldberg AL** 2001 Atrogin-1, a muscle-specific F-box protein highly expressed during muscle atrophy
1. Proc Natl Acad Sci U S A 98:14440-14445

17. **Lee CK, Klopp RG, Weindruch R, Prolla TA** 1999 Gene expression profile of aging and its retardation by caloric restriction
7. Science 285:1390-1393

18. **Pietu G, Eveno E, Soury-Segurens B, Fayein NA, Mariage-Samson R, Matingou C, Leroy E, Dechesne C, Krieger S, Ansorge W, Reguigne-Arnould I, Cox D, Dehejia A, Polymeropoulos MH, Devignes MD, Auffray C** 1999 The genexpress IMAGE knowledge base of the human muscle transcriptome: a resource of structural, functional, and positional candidate genes for muscle physiology and pathologies
5. Genome Res 9:1313-1320

19. **Chen YW, Zhao P, Borup R, Hoffman EP** 2000 Expression profiling in the muscular dystrophies: identification of novel aspects of molecular pathophysiology
2. J Cell Biol 151:1321-1336

20. **Hoffman EP, Fischbeck KH, Brown RH, Johnson M, Medori R, Loike JD, Harris JB, Waterston R, Brooke M, Specht L, .** 1988 Characterization of dystrophin in muscle-biopsy specimens from patients with Duchenne's or Becker's muscular dystrophy
1. N Engl J Med 318:1363-1368

21. **Koenig M, Monaco AP, Kunkel LM** 1988 The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein
1. Cell 53:219-228

22. **Bakay M, Zhao P, Chen J, Hoffman EP** 2002 A web-accessible complete transcriptome of normal human and DMD muscle
2. Neuromuscul Disord 12 Suppl 1:S125-S141

23. **Haslett JN, Sanoudou D, Kho AT, Bennett RR, Greenberg SA, Kohane IS, Beggs AH, Kunkel LM** 2002 Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle
2. Proc Natl Acad Sci U S A 99:15000-15005

24. **Haslett JN, Sanoudou D, Kho AT, Han M, Bennett RR, Kohane IS, Beggs AH, Kunkel LM** 2003 Gene

expression profiling of Duchenne muscular dystrophy skeletal muscle
1. Neurogenetics 4:163-171

25. **Tkatchenko AV, Le CG, Leger JJ, Dechesne CA** 2000 Large-scale analysis of differential gene expression in the hindlimb muscles and diaphragm of mdx mouse
3. Biochim Biophys Acta 1500:17-30

26. **Noguchi S, Tsukahara T, Fujita M, Kurokawa R, Tachikawa M, Toda T, Tsujimoto A, Arahata K, Nishino I** 2003 cDNA microarray analysis of individual Duchenne muscular dystrophy patients
1. Hum Mol Genet 12:595-600

27. **Pescatori M, Broccolini A, Minetti C, Bertini E, Bruno C, D'amico A, Bernardini C, Mirabella M, Silvestri G, Giglio V, Modoni A, Pedemonte M, Tasca G, Galluzzi G, Mercuri E, Tonali PA, Ricci E** 2007 Gene expression profiling in the early phases of DMD: a constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression
1. FASEB J 21:1210-1226

28. **Chen YW, Nagaraju K, Bakay M, McIntyre O, Rawat R, Shi R, Hoffman EP** 2005 Early onset of inflammation and later involvement of TGFbeta in Duchenne muscular dystrophy
1. Neurology 65:826-834

29. **Campanaro S, Romualdi C, Fanin M, Celegato B, Pacchioni B, Trevisan S, Laveder P, De PC, Pegoraro E, Hayashi YK, Valle G, Angelini C, Lanfranchi G** 2002 Gene expression profiling in dysferlinopathies using a dedicated muscle microarray
1. Hum Mol Genet 11:3283-3298

30. **Celegato B, Capitanio D, Pescatori M, Romualdi C, Pacchioni B, Cagnin S, Vigano A, Colantoni L, Begum S, Ricci E, Wait R, Lanfranchi G, Gelfi C** 2006 Parallel protein and transcript profiles of FSHD patient muscles correlate to the D4*Z4* arrangement and reveal a common impairment of slow to fast fibre differentiation and a general deregulation of MyoD-dependent genes
1. Proteomics 6:5303-5321

31. **Osborne RJ, Welle S, Venance SL, Thornton CA, Tawil R** 2007 Expression profile of FSHD supports a link between retinal vasculopathy and muscular dystrophy
1. Neurology 68:569-577

32. **Greenberg SA, Sanoudou D, Haslett JN, Kohane IS, Kunkel LM, Beggs AH, Amato AA** 2002 Molecular profiles of inflammatory myopathies
7. Neurology 59:1170-1182

33. **Raju R, Dalakas MC** 2005 Gene expression profile in the muscles of patients with inflammatory myopathies: effect of therapy with IVIg and biological validation of clinically relevant genes
8. Brain 128:1887-1896

34. **Zhou X, Dimachkie MM, Xiong M, Tan FK, Arnett FC** 2004 cDNA microarrays reveal distinct gene expression clusters in idiopathic inflammatory myopathies
1. Med Sci Monit 10:BR191-BR197

35. **Greenberg SA, Bradshaw EM, Pinkus JL, Pinkus GS, Burleson T, Due B, Bregoli L, O'Connor KC, Amato AA** 2005 Plasma cells in muscle in inclusion body myositis and polymyositis
1. Neurology 65:1782-1787

36. **Taniguchi M, Kurahashi H, Noguchi S, Sese J, Okinaga T, Tsukahara T, Guicheney P, Ozono K, Nishino I, Morishita S, Toda T** 2006 Expression profiling of muscles from Fukuyama-type congenital muscular dystrophy and laminin-alpha 2 deficient congenital muscular dystrophy; is congenital muscular dystrophy a primary fibrotic disease?
5. Biochem Biophys Res Commun 342:489-502

37. **Taniguchi M, Kurahashi H, Noguchi S, Fukudome T, Okinaga T, Tsukahara T, Tajima Y, Ozono K, Nishino I, Nonaka I, Toda T** 2006 Aberrant neuromuscular junctions and delayed terminal muscle fiber maturation in alpha-dystroglycanopathies
1. Hum Mol Genet 15:1279-1289

38. **Noguchi S, Fujita M, Murayama K, Kurokawa R, Nishino I** 2005 Gene expression analyses in X-linked myotubular myopathy. Neurology 65:732-737

39. **Sanoudou D, Haslett JN, Kho AT, Guo S, Gazda HT, Greenberg SA, Lidov HG, Kohane IS, Kunkel LM, Beggs AH** 2003 Expression profiling reveals altered satellite cell numbers and glycolytic enzyme transcription in nemaline myopathy muscle
1. Proc Natl Acad Sci U S A 100:4666-4671

40. **Tezak Z, Hoffman EP, Lutz JL, Fedczyna TO, Stephan D, Bremer EG, Krasnoselska-Riz I, Kumar A, Pachman LM** 2002 Gene expression profiling in DQA1*0501+ children with untreated dermatomyositis: a novel model of pathogenesis
1. J Immunol 168:4154-4163

41. **Bakay M, Wang Z, Melcon G, Schiltz L, Xuan J, Zhao P, Sartorelli V, Seo J, Pegoraro E, Angelini C, Shneiderman B, Escolar D, Chen YW, Winokur ST, Pachman LM, Fan C, Mandler R, Nevo Y, Gordon E, Zhu Y, Dong Y, Wang Y, Hoffman EP** 2006 Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration
1. Brain 129:996-1013

42. **Sanoudou D, Frieden LA, Haslett JN, Kho AT, Greenberg SA, Kohane IS, Kunkel LM, Beggs AH** 2004 Molecular classification of nemaline myopathies: "nontyping" specimens exhibit unique patterns of gene expression
1. Neurobiol Dis 15:590-600

43. **Di Giovanni S., Molon A, Broccolini A, Melcon G, Mirabella M, Hoffman EP, Servidei S** 2004 Constitutive activation of MAPK cascade in acute

quadriplegic myopathy
1. Ann Neurol 55:195-206

44. **Greenberg SA, Pinkus JL, Pinkus GS, Burleson T, Sanoudou D, Tawil R, Barohn RJ, Saperstein DS, Briemberg HR, Ericsson M, Park P, Amato AA** 2005 Interferon-alpha/beta-mediated innate immune mechanisms in dermatomyositis
1. Ann Neurol 57:664-678

45. **Molon A, Di GS, Chen YW, Clarkson PM, Angelini C, Pegoraro E, Hoffman EP** 2004 Large-scale disruption of microtubule pathways in morphologically normal human spastin muscle
1. Neurology 62:1097-1104

46. **Bakay M, Chen YW, Borup R, Zhao P, Nagaraju K, Hoffman EP** 2002 Sources of variability and effect of experimental approach on expression profiling data interpretation
1. BMC Bioinformatics 3:4

47. **Crimi M, Bordoni A, Menozzi G, Riva L, Fortunato F, Galbiati S, Del BR, Pozzoli U, Bresolin N, Comi GP** 2005 Skeletal muscle gene expression profiling in mitochondrial disorders
1. FASEB J 19:866-868

48. **Rome S, Clement K, Rabasa-Lhoret R, Loizon E, Poitou C, Barsh GS, Riou JP, Laville M, Vidal H** 2003 Microarray profiling of human skeletal muscle reveals that insulin regulates approximately 800 genes during a hyperinsulinemic clamp
1. J Biol Chem 278:18063-18068

49. **Viguerie N, Clement K, Barbe P, Courtine M, Benis A, Larrouy D, Hanczar B, Pelloux V, Poitou C, Khalfallah Y, Barsh GS, Thalamas C, Zucker JD, Langin D** 2004 In vivo epinephrine-mediated regulation of gene expression in human skeletal muscle
1. J Clin Endocrinol Metab 89:2000-2014

50. **Balagopal P, Olney R, Darmaun D, Mougey E, Dokler M, Sieck G, Hammond D** 2006 Oxandrolone enhances skeletal muscle myosin synthesis and alters global gene expression profile in Duchenne muscular dystrophy
1. Am J Physiol Endocrinol Metab 290:E530-E539

51. **Cohn RD, Campbell KP** 2000 Molecular basis of muscular dystrophies. Muscle Nerve 23:1456-1471

52. **Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM** 2002 Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer
1. Cancer Res 62:4427-4433

53. **Rhodes DR, Chinnaiyan AM** 2005 Integrative analysis of the cancer transcriptome
4. Nat Genet 37 Suppl:S31-S37

54. **Segal E, Friedman N, Kaminski N, Regev A, Koller D** 2005 From signatures to models: understanding cancer using microarrays
1. Nat Genet 37 Suppl:S38-S45

55. **Kho AT, Kang PB, Kohane IS, Kunkel LM** 2006 Transcriptome-scale similarities between mouse and human skeletal muscles with normal and myopathic phenotypes
1. BMC Musculoskelet Disord 7:23

56. **Shaffer AL, Rosenwald A, Hurt EM, Giltnane JM, Lam LT, Pickeral OK, Staudt LM** 2001 Signatures of the immune response
1. Immunity 15:375-385

57. **Pietu G, Alibert O, Guichard V, Lamy B, Bois F, Leroy E, Mariage-Sampson R, Houlgatte R, Soularue P, Auffray C** 1996 Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 6:492-503

58. **Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK** 2003 Computational discovery of gene modules and regulatory networks
1. Nat Biotechnol 21:1337-1342

59. **Beer MA, Tavazoie S** 2004 Predicting gene expression from sequence
1. Cell 117:185-198

60. **Segal E, Yelensky R, Koller D** 2003 Genome-wide discovery of transcriptional modules from DNA sequence and gene expression
1. Bioinformatics 19 Suppl 1:i273-i282

61. **Bartel DP** 2004 MicroRNAs: genomics, biogenesis, mechanism, and function
1. Cell 116:281-297

62. **varez-Garcia I, Miska EA** 2005 MicroRNA functions in animal development and human disease
1. Development 132:4653-4662

63. **Kloosterman WP, Plasterk RH** 2006 The diverse functions of microRNAs in animal development and disease
4. Dev Cell 11:441-450

64. **Eisenberg I, Eran A, Nishino I, Moggio M, Lamperti C, Amato AA, Lidov HG, Kang PB, North KN, Mitrani-Rosenbaum S, Flanigan KM, Neely LA, Whitney D, Beggs AH, Kohane IS, Kunkel LM** 2007 Distinctive patterns of microRNA expression in primary muscular disorders
1. Proc Natl Acad Sci U S A 104:17016-17021

65. **Cheung VG, Spielman RS** 2002 The genetics of variation in gene expression
1. Nat Genet 32 Suppl:522-525

66. **Kruglyak L, Nickerson DA** 2001 Variation is the spice of life
2. Nat Genet 27:234-236

67. **Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA** 2003 Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans
1. Nat Genet 33:518-521

68. **Fan JB, Chen X, Halushka MK, Berno A, Huang X, Ryder T, Lipshutz RJ, Lockhart DJ, Chakravarti A** 2000 Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays
1. Genome Res 10:853-860

69. **Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ** 2000 Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays
1. Genome Res 10:1126-1137

70. **Sellick GS, Longman C, Brockington M, Mahjneh I, Sagi L, Bushby K, Topaloglu H, Muntoni F, Houlston RS** 2005 Localisation of merosin-positive congenital muscular dystrophy to chromosome 4p16.3
1. Hum Genet 117:207-212

71. **Weiler T, Bashir R, Anderson LV, Davison K, Moss JA, Britton S, Nylen E, Keers S, Vafiadaki E, Greenberg CR, Bushby CR, Wrogemann K** 1999 Identical mutation in patients with limb girdle muscular dystrophy type 2B or Miyoshi myopathy suggests a role for modifier gene(s)
1. Hum Mol Genet 8:871-877

72. **Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO** 2001 Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF
1. Nature 409:533-538

73. **Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA** 2000 Genome-wide location and function of DNA binding proteins
2. Science 290:2306-2309

74. **Blais A, Tsikitis M, costa-Alvear D, Sharan R, Kluger Y, Dynlacht BD** 2005 An initial blueprint for myogenic differentiation. Genes Dev 19:553-569

75. **Cao Y, Kumar RM, Penn BH, Berkes CA, Kooperberg C, Boyer LA, Young RA, Tapscott SJ** 2006 Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters
5. EMBO J 25:502-511

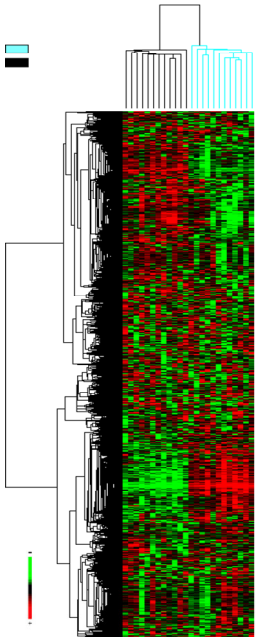
76. **Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA** 2006 Control of developmental regulators by Polycomb in human embryonic stem cells
1. Cell 125:301-313

77. **Dufour CR, Wilson BJ, Huss JM, Kelly DP, Alaynick WA, Downes M, Evans RM, Blanchette M, Giguere V** 2007 Genome-wide orchestration of cardiac functions by the orphan nuclear receptors ERRalpha and gamma
3. Cell Metab 5:345-356

78. **Ginestier C, Charafe-Jauffret E, Bertucci F, Eisinger F, Geneix J, Bechlian D, Conte N, Adelaide J, Toiron Y, Nguyen C, Viens P, Mozziconacci MJ, Houlgatte R, Birnbaum D, Jacquemier J** 2002 Distinct and complementary information provided by use of tissue and DNA microarrays in the study of breast tumor markers
2. Am J Pathol 161:1223-1233

79. **Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP** 1998 Tissue microarrays for high-throughput molecular profiling of tumor specimens
1. Nat Med 4:844-847

A



B

Fonctions	Gènes (symboles officiels)
structural molecule activity	ACTB; ACTG1; ACTN1; ADD3; ANK3; ARPC1B; ARPC3; ARPC5; C14ORF132; CAV1; CHI3L1; CLDN5; CLTC; COL14A1; COL15A1; COL16A1; COL1A2; COL3A1; COL4A1; COL4A2; COL5A2; COL6A1; COL6A2; COL6A3; EFEMP1; EMP1; EMP3; ENO1; EPB41L2; EPB41L3; FBLN1; FN1; GADD45A; GPM6B; HSPG2; JAG1; LAMA4; LAMB1; LUM; MAP1A; MGP; MYBPH; MYH8; MYL1; MYL4; MYL5; MYL6; NF2; PLP1; PNN; PTMA; RPL11; RPL19; RPL27A; RPL3; RPL5; RPL7; RPS12; RPS17; RPS27; RPS6KA1; SGCE; SPTAN1; SPTBN1; STAB1; TEK; THBS2; THBS3; THBS4; TMEM4; TNNT2; TUBB; VIM
extracellular matrix structural constituent	CHI3L1; COL14A1; COL15A1; COL16A1; COL1A2; COL3A1; COL4A1; COL4A2; COL5A2; COL6A1; COL6A2; COL6A3; FBLN1; FN1; LAMA4; LAMB1; LUM; MGP
cell adhesion molecule activity	ACTN1; AEBP1; CD44; CD99; CDH11; CLSTN3; COL15A1; COL6A1; COL6A2; COL6A3; DAG1; DPT; EFS; FAT; FN1; IGSF4; ITGA5; ITGA6; ITGA7; ITGB1; ITGB2; L1CAM; LAMA4; LAMB1; MAGED2; MYBPH; NCAM1; NID2; NRCAM; PTPRF; SELP; SH3GL1; SPP1; SRPX; STAB1; TGFBI; THBS2; THBS3; THBS4; TNC; TRO; TYRO3; VCAM1
MHC class II receptor activity	CD74; HLA-DMB; HLA-DPA1; HLA-DPB1; HLA-DQB1; HLA-DRA; HLA-DRB1
collagen binding	CD44; FN1; NID2; PCOLCE; SERPINH1; SPARC
actin binding	ACTN1; CALD1; CAPZA1; CFL1; CNN3; DBN1; DSTN; EPB41L2; EPB41L3; FLNA; FSCN1; GMFB; GMFG; LSP1; MARCKS; MYH3; MYH8; MYO10; MYO1B; PLS3; PTK9; SDCBP; SPTAN1; SPTBN1; TAGLN2; TMSB10; TMSB4X
calcium ion binding	ACTN1; AIF1; ANXA1; ANXA2; ANXA4; ANXA5; C1R; C1S; CANX; CAPN2; CASQ2; CBLB; CDH11; CETN2; CLSTN3; CSPG2; EFEMP1; F13A1; FAT; FBLN1; GRN; JAG1; MATN2; MGP; MMP2; MXRA5; MYL1; MYL4; MYL5; MYL6; NID2; NOTCH3; PLA2G2A; PLCG2; PLS3; PLSCR1; PRNP; PROS1; RCN1; S100A10; S100A11; S100A2; S100A4; SPARC; SPARCL1; SPTAN1; THBS2; THBS3; THBS4
glycosaminoglycan binding	APOE; CD44; CSPG2; FGFR1; FN1; MDK; NCAM1; PTN; SERPINE2; STAB1; TGFBR3; THBS2; THBS3; THBS4

Figure 1: Cluster de gènes co-exprimés (**A**) et fonction biologique sous-jacente (**B**). (**A**) classification hiérarchique des profils d'expression dans le muscle **DMD** (Dystrophie Musculaire de Duchenne) et dans le muscle **sain**. Les données brutes [Haslett et al., PNAS, 2002] ont subi une re-normalisation non-linéaire (LOWESS), une transformation logarithmique (Log2), un centrage médian au niveau des gènes et une classification hiérarchique sur la base des coefficients de corrélations. Chaque case de la matrice colorée représente le niveau d'expression d'un gène dans un échantillon avec le code couleur suivant : rouge pour une sur-expression, vert pour une sous-expression et noir pour un niveau d'expression proche de celui de la médiane (voir l'échelle en bas à gauche). Les résultats de l'analyse montrent un groupe de gènes corrélés (cluster) sur-exprimés dans le muscle **DMD** (entouré en blanc sur la figure). (**B**) Parmi les gènes sur-exprimés dans le muscle DMD, beaucoup codent pour des protéines de la matrice extracellulaire en relation avec la fibrose musculaire.

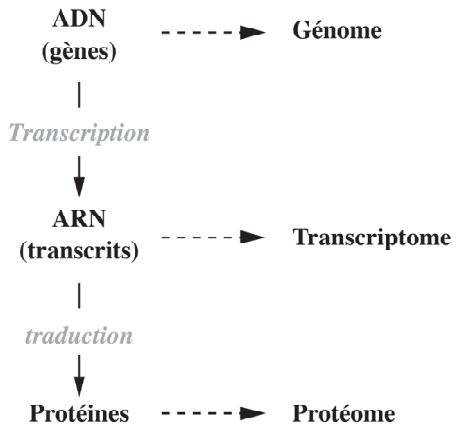


Figure 2: les différents niveaux d'approches de la génomique

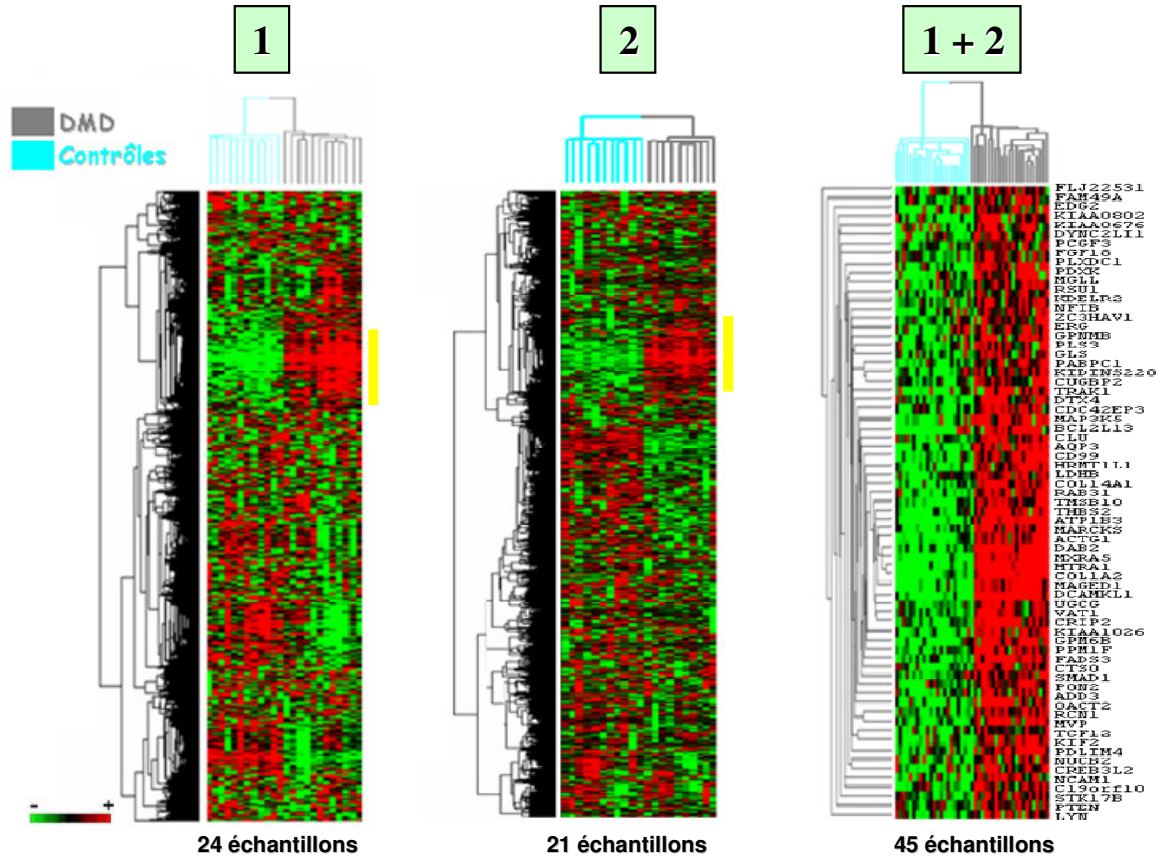


Figure 3: Comparaison de deux jeux de données de transcriptome de muscle **DMD** et de muscle **sain**. Les deux jeux de données originaux [1 : Haslett et al., PNAS, 2002 ; 2 : Haslett et al., Neurogenetics, 2003] correspondent à deux études indépendantes. Les données brutes ont subi une renormalisation non-linéaire (LOWESS), une transformation logarithmique (Log2), un centrage médian au niveau des gènes et une classification hiérarchique sur la base des coefficients de corrélations. Chaque case de la matrice colorée représente le niveau d'expression d'un gène dans un échantillon avec le code couleur suivant : rouge pour une sur-expression, vert pour une sous-expression et noir pour un niveau d'expression proche de celui de la médiane (voir l'échelle en bas à gauche). Chacun des deux jeux de données [1 et 2] montre un cluster de gènes sur-exprimés dans le muscle **DMD** (noté par une barre jaune). La méta-analyse de ces deux clusters [1+2] met en évidence une liste robuste de 69 gènes communs co-exprimés au travers des 45 échantillons de muscle.

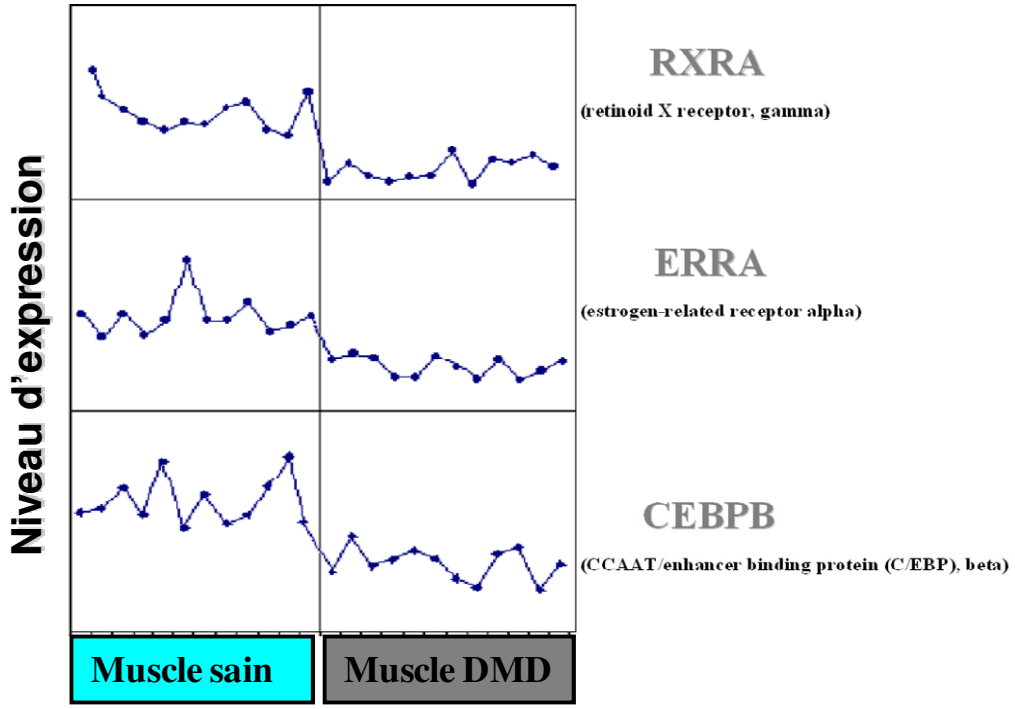


Figure 4: Exemples de trois facteurs de transcription sous-exprimés dans le muscle **DMD**. Les profils d'expression ont été générées à partir des données brutes [Haslett et al., PNAS, 2002] re-normalisées par LOWESS. RXRA: retinoid X receptor, gamma ; ERRA : estrogen-related receptor alpha ; CEBPB : CCAAT/enhancer binding protein (C/EBP), beta.

Puces à ADN

ITBM-RBM 28 (2007) 210–215

Daniel Baron [1,2], Mahatsangy Raharijaona [1,2], Rémi Houlgatte [1,2,3,4]
1 : INSERM, U533, Nantes, F-44000 France.
2 : Université de Nantes, Faculté de Médecine, l'institut du thorax, Nantes, F-44000 France.
3 : CHU Nantes, l'institut du thorax, F-44000 France.
4 : remi.houlgatte@nantes.inserm.fr

Résumé
Les puces à ADN sont un réseau de sondes d'ADN très dense déposé sur une surface solide. Elles permettent de mesurer la quantité des ADN complémentaires à ces sondes en une seule hybridation. Les puces à ADN permettent de mesurer les variations du génome ou du transcriptome, de reséquencer un gène ou une région génomique, de déterminer les sites de fixation d'un facteur de transcription sur la chromatine ou bien de doser la présence d'un organisme pathogène dans des fluides biologiques. Les puces à ADN sont l'outil le plus abouti de la génomique qui vise à étudier la séquence et l'expression des génomes. L'application la plus courante des puces à ADN est la mesure du niveau d'expression de tous les gènes d'un génome : le transcriptome.

Abstract
DNA microarrays are a dense network of DNA probes spotted on a solid surface. They allow the measurement of the complementary DNA of each probe in a single hybridization. DNA microarrays allow to measure variations of genome sequence or expression, to resequence a gene or a genomic region, to determine the genomic binding sites of a transcription factor on the chromatine, or to detect the presence of a pathogen in biological fluids. DNA microarrays are the most achieved tool of functional genomics which aim to study sequence and expression of genomes.
The most common application of DNA microarrays is the measurement of gene expression of all the genes of a genome: the transcriptome.

Mots clés : Génomique fonctionnelle ; Profil d'expression ; Puce à ADN ; Transcriptome
Keywords: DNA microarray; Expression profiling; Functional genomics; Transcriptome

Le séquencage des génomes a ouvert la voie à des études fonctionnelles à grande échelle. Ces études de génomique fonctionnelle visent à comprendre le fonctionnement de l'organisme par l'analyse de l'expression de ces gènes. Parmi les outils de la génomique fonctionnelle, les puces à ADN sont probablement 'outil le plus abouti. Initialement développées comme outil de criblage de banque [1], les puces à ADN se sont imposées comme l'outil de la mesure d'expression des gènes [2–6]. Contrairement à d'autres techniques, elles permettent de mesurer l'expression des gènes de l'ensemble d'un génome même complexe, en une seule hybridation.L'existence de puces commerciales de plus en plus fiables

et de moins en moins chères et de protocoles d'hybridation et d'analyses standardisées rendent cet outil indispensable à tous les domaines de la biologie. Enfin, le champ d'application des puces à ADN n'a cessé d'augmenter. À l'heure actuelle, les puces à ADN permettent de mesurer les variations du génome, du transcriptome, de reséquencer un gène ou bien de caractériser les sites de fixation d'un acteur de transcription sur la chromatine.

1. Aspects technologiques

Les puces à ADN sont un réseau de sondes très dense déposé sur une surface solide. Ces sondes sont hybridées avec l'ADN (ou l'ADN complémentaire) que l'on cherche à mesurer. Cet ADN à doser est marqué par de la radioactivité, une enzyme, de la fluorescence ou de la biotine... Le signal obtenu sur chaque sonde de la puce reflète l'abondance de la séquence complémentaire dans l'ADN à doser (Fig. 1).

Les sondes utilisées pour fabriquer ces puces à ADN peuvent être des clones d'ADNc [2], des plasmides [3], des produits de PCR [4,5] ou des oligonucléotides de synthèse [6]. Ces sondes sont déposées sur le support de nylon, de verre, de silicium... au moyen de robots de dépôts utilisant des pointes pleines, des plumes avec réservoir, des imprimantes à jet d'encre ou bien sont synthétisées in situ. Ces différentes techniques présentent des avantages respectifs, mais il est maintenant admis que l'utilisation d'oligonucléotides de synthèse constitue le meilleur choix pour des sondes.

Ces puces à oligonucléotides, malgré un coût plus élevé, offrent une plus grande fiabilité, une meilleure spécificité, une flexibilité, une simplicité et une sécurité d'utilisation. Cela est dû au fait qu'il est beaucoup plus facile d'égaliser les tailles, les concentrations et les *melting temperature* (Tm) d'oligonucléotides de synthèse que de clones d'ADNc. De même, il est toujours possible de concevoir des oligonucléotides permettant de distinguer des gènes orthologues ayantmême de forte similarité, en concevant des oligonucléotides dans les zones les plus divergentes [7]. Cela est beaucoup plus difficile pour des clones d'ADNc de grande taille. Enfin, les protocoles de synthèse des oligonucléotides étant totalement robotisés, les seules manipulations à réaliser sont le dépôt. Il n'en va pas de même des clones d'ADNc qu'il faut faire pousser sous agitation, extraire (partiellement) le plasmide, puis amplifier par PCR. Ces nombreuses manipulations et amplifications, en culture, puis par PCR, offrent autant de sources de contamination et d'artefacts. Enfin, il est plus facile de concevoir de nouveaux oligonucléotides en remplacement d'oligonucléotides ne donnant pas de signal (par exemple à cause d'une mauvaise conception), que de remplacer un clone dans les différentes plaques (plaques clones, plasmides et PCR) de puces à ADNc.

Malgré ces très nombreux avantages en faveur des puces à oligonucléotides, les différentes techniques de puces coexistent encore. En effet, la conception de puces à oligonucléotides repose sur une connaissance approfondie du génome. En effet, pour concevoir un oligonucléotide spécifique d'un gène, il faut connaître la séquence de ce gène, de ses éventuels orthologues et du génome dans sa totalité, afin d'éviter les hybridations croisées indésirables. Dans le cas d'un génome non séquencé ou mal séquencé, les puces à produits de PCR issus de clones d'ADNc sont une solution simple et peu chère pour commencer des études de transcriptome. L'existence de banques d'ADNc utilisant un même vecteur (pour standardiser l'amplification PCR) et partiellement séquencées sous forme d'étiquettes de séquences exprimées (EST) [8–10] constituent un outil appréciable pour

concevoir facilement des puces à ADNc. La disponibilité de ces clones dans des centres de ressources, comme, par exemple, les clones IMAGE [11,12] s'avèrent un outil important pour la communauté scientifique travaillant sur un génome non encore séquencé.

Les puces commerciales actuelles utilisent toutes des oligonucléotides déposés ou synthétisés sur une surface solide. Le nombre de sondes disponibles est de plusieurs milliers à plusieurs centaines de milliers, ce qui est suffisant pour mesurer l'ensemble des gènes des génomes les plus grands. L'amélioration des protocoles, des procédures, et des appareillages font des puces à ADN des outils de plus en plus sensibles et reproductibles. Par exemple, il est maintenant couramment admis qu'il est possible de détecter une copie par cellule d'un ARN donné [6,13]. De même, l'amélioration et la standardisation des méthodes d'amplification linéaire à la ARN polymérase T7 [14,15] permettent de mesurer l'expression des gènes à partir de quelques milliers de cellules [16]. Ces performances font des puces à ADN un outil incontournable de la biologie.

2. L'étude du transcriptome

L'application la plus habituelle des puces à ADN est la mesure du transcriptome, c'est-à-dire la mesure du niveau d'expression de l'ensemble des gènes d'un génome. Cela en fait un outil majeur de la génomique qui vise à étudier les génomes en terme de séquence (et de variation de séquence) et de fonctionnement (d'expression). L'étude du transcriptome d'un échantillon biologique permet de caractériser des signatures aractéristiques de cet échantillon. C'est dans le domaine du cancer où l'apport des puces à ADN est le plus évident. Cela est dû au fait que les cancers, du moins les tumeurs solides, s'accompagnent de remaniements géniques majeurs (remaniements chromosomiques, amplifications, insertions, délétions, mutations, acétylation des histones, méthylation des CpG. . .) qui ne peuvent être appréhendés dans leur globalité qu'avec des techniques à grande échelle. L'analyse du transcriptome des cancers a permis de montrer qu'il est possible de classer les tumeurs d'une façon au moins aussi efficace que l'anatomopathologie classique, au moyen de signatures moléculaires caractéristiques [17,18]. De plus, l'analyse du transcriptome permet de découvrir de nouvelles classes dans des groupes que l'on pensait homogènes ou dans des groupes d'échantillons inclassables. C'est, par exemple, la découverte de deux classes parmi les lymphomes B diffus à grandes cellules [19], l'existence de cinq sous-groupes caractéristiques de cancer du sein [20,21], ou l'existence de trois nouvelles classes parmi les lymphomes T inclassables [22]. De même, l'analyse du transcriptome a permis de mettre en évidence des signatures associées au pronostic [23] ou à la sensibilité au traitement [24–26] des cancers.

3. Applications des puces à ADN

Il existe de très nombreuses applications des puces à ADN (Tableau 1A). Les puces à ADN permettent de cribler des banques génomiques ou d'ADNc ou bien de caractériser leurs clones par empreintes d'oligonucléotides. Elles permettent de mesurer l'expression des gènes, mais aussi de caractériser la taille des ARN messagers en hybridant sur différentes puces les différentes fractions d'un gel d'électrophorèse. Elles permettent aussi de détecter et de valider les gènes dans un génome ou
une région génomique. Pour cela, une puce à ADN comportant des sondes régulièrement réparties sur la séquence à tester est hybridée avec des ARN provenant de différents tissus.

Ces hybridations mettent en évidence les régions transcrites et la cohérence des signaux obtenus avec les différents tissus permet de déterminer si plusieurs régions voisines transcrites proviennent d'un seul ou de plusieurs gènes.

Au niveau de l'ADN, les puces à ADN permettent de déterminer des délétions ou des amplifications de régions chromosomiques par hybridation génomique comparative (CGH) sur des puces couvrant la région à tester au moyen de sondes régulièrement réparties. On peut même reséquencer cette région en constituant une puce couvrant cette région base par base (Fig. 2). C'est le même principe qui permet d'analyser les points de polymorphisme (SNP) des génomes : pour chaque SNP, des oligonucléotides correspondant aux différents variants connus sont déposés sur une puce qui sera hybridée avec l'ADN d'un individu à tester. Ces stratégies permettent de caractériser des variants (différentes souches par exemple) d'un organisme donné en effectuant un phénotypage moléculaire.

Concernant les protéines, les puces à ADN permettent de déterminer les sites potentiellement reconnus par un facteur de transcription, en hybridant ce facteur sur des oligonucléotides doubles brins correspondant à toutes les séquences de taille N possibles (Fig. 3).

Une autre application concernant les facteurs de transcription est le *chromatine immunoprecipitation* (ChIP)–chip [27,28]. Il s'agit d'immunoprécipiter un facteur de transcription en même temps que les segments d'ADN sur lequel il est fixé. Ces segments d'ADN sont ensuite caractérisés par hybridation sur une puce à ADN (chip) constituée de sondes des régions promotrices, des îlots CpG ou couvrant l'ensemble du génome. Cette technique est particulièrement intéressante car c'est une méthode directe applicable à tout type d'échantillons biologiques (cellules, tissu, organe) et ne modifiant pas le système étudié. Les méthodes concurrentes sont, soit indirectes (recherche de motif dans les promoteurs de gènes, retard sur gel), soit applicables uniquement à des modèles cellulaires (transfection, SiRNA) ou bien perturbant le système étudié (*knock-out*, *knock-in*, *transfection*). Il a, par exemple, pu être montré que la surexpression de Myc se traduit par la fixation du facteur sur de nombreux sites non utilisés habituellement [29]. Cette méthode, de plus, permet de déterminer les mécanismes mis en jeu lors des activations de gènes observées au niveau du transcriptome, en analysant le rôle de différents facteurs de transcription dans le système étudié.

Il existe probablement de nombreuses applications des puces à ADN non décrites ici. Les puces à ADN remplacent progressivement beaucoup de techniques de la biologie moléculaire. Les applications de suivi de réplication de plasmides ou de mesure de la taille des ARN en sont deux exemples flagrants. De même, l'analyse du transcriptome ne se limite pas à la mesure du niveau d'expression des ARN messagers. D'autres paramètres peuvent être mesurés (Tableau 1B). Il s'agit, par exemple, de la mesure du taux de traduction des protéines en mesurant la liaison aux polysomes de leur ARN messager ou bien de leur localisation subcellulaire par l'analyse de la localisation subcellulaire de leur ARN. Chez la levure, l'analyse cinétique du transcriptome d'un mutant thermosensible de la RNA polymérase II, a permis de mesurer le taux de dégradation des ARN. Il est probablement envisageable de mesurer les taux de synthèse et de dégradation de tous les ARN par des expériences de marquages courts avec chasse isotopique (*Pulse-chase*). Cette application souvent citée dans les congrès n'a toutefois pas encore été publiée.

4. Approches génomiques intégrées

À l'heure actuelle de plus en plus d'études de l'expression des génomes combinent plusieurs mesures génomiques utilisant des puces à ADN. Il peut s'agir de combiner une approche de génotypage CGH avec l'analyse du transcriptome [30] afin de déterminer l'impact des amplifications ou des délétions géniques sur le transcriptome. Il peut aussi s'agir de la comparaison du transcriptome de différentes espèces afin de mettre en évidence des modules transcriptionnels fortement conservés dans l'évolution [31]. Deux autres approches intégrées ouvrent des pistes particulièrement intéressantes.

La première vise à intégrer les données du transcriptome avec celles du ChIP–chip. Les données du transcriptome permettent de mettre en évidence des expressions particulières à une situation biologique donnée. Il peut, par exemple, s'agir des gènes sur- ou sous-exprimés dans des tumeurs de mauvais pronostic. L'analyse des sites de fixation des facteurs de transcription sur le promoteur de ces gènes permet de comprendre les mécanismes de régulation conduisant à la malignité particulière de ces tumeurs. L'analyse à grande échelle du transcriptome et du rôle des facteurs de transcription dans un système biologique donné ouvre même la voie à la modélisation statique ou dynamique de ce système, la biologie des systèmes. Cette modélisation permettrait même de modéliser l'effet d'un médicament ou d'une mutation sur un système biologique donné.

La deuxième approche vise à coupler le génotypage SNP et l'analyse du transcriptome. L'apport du transcriptome à la génétique n'est pas nouveau. Par exemple, dans les approches de clonage positionnel, on détermine une région contenant plusieurs gènes candidats. L'analyse du transcriptome permet de déterminer les gènes exprimés dans le tissu atteint et d'expression différentielle chez les individus normaux et atteints. Ces gènes constituent de bons gènes candidats à analyser en premier. Récemment, une approche originale appelée génomique génétique a été développée [32]. Elle vise à faciliter la découverte de gènes de susceptibilité dans les pathologies multifactorielles en couplant le génotypage SNP à l'analyse du transcriptome. L'idée principale est d'utiliser les expressions géniques comme des traits quantitatifs *expressed quantitative trait loci* (eQTL). La recherche de polymorphismes génétiques affectant l'expression des gènes permet de relier les polymorphismes directement à l'expression des gènes. L'ensemble des gènes affectés par un même polymorphisme participe à une même fonction biologique, c'est-à-dire à un même trait quantitatif.

5. Offres académiques et commerciales

Àl'heure actuelle de nombreuses plates-formes académiques fournissent un soutien aux projets académiques d'analyse du transcriptome ou pour d'autres applications : hybridation génomique comparative, pucesSNPou ChIP-chip. Ce soutien s'avère très important pour permettre d'utiliser des protocoles et des méthodes d'analyse et d'interprétation appropriées. Ces expertises remplacent progressivement l'activité de fabrication des supports qui, compte tenu des prix sans cesse en baisse des puces commerciales, ne se justifient plus sauf pour les espèces non séquencées.

Au niveau commercial, plusieurs sociétés utilisant différentes technologies toutes basées sur des sondes oligonucléotidiques, coexistent. On peut distinguer les sociétés utilisant des oligonucléotides longs présynthétisés puis déposés sur un support solide (Applied Biosystem). Cette approche s'avère particulièrement peu flexible. Le coût de synthèse à grande échelle de dizaines de milliers d'oligonucléotides ne permet pas

d'évolution rapide du jeu de sondes. La fabrication de puces à fac¸on est impossible car elle nécessiterait des réorganisations fréquentes du jeu de sondes. Cela fait que cette société a limité ses applications aux plus fréquentes : la mesure du transcriptome chez l'homme, le
rat et la souris.

Une autre technique repose sur la synthèse d'oligonucléotides longs sur des billes (Illumina) qui sont ensuite déposées dans de minuscules cavités sur un support solide. Cette technologie souffre du même manque de flexibilité que la précédente, et cette société a une offre limitée à la mesure du transcriptome chez l'homme, le rat et la souris, et au génotypageSNPchez l'homme. Elle possède toutefois l'avantage d'un partenariat privilégié avec le consortium HapMap impliqué dans de génotypage SNP.

La technologie dominante à l'heure actuelle utilise la synthèse in situ d'oligonucléotides courts (Affymetrix). Cette synthèse in situ offre une très grande flexibilité puisqu'il est possible de modifier à volonté les séquences des sondes. Cette technologie utilise toutefois des masques pour la déprotection des bases lors de la synthèse, comme ceux utilisés dans l'industrie des microprocesseurs. Pour chaque ajout de nucléotide, il faut un masque de déprotection pour chacune des bases A, T, C et G, ce sont donc 100 masques (25 mères×quatre bases) qu'il faut synthétiser pour concevoir une nouvelle puce. Cela limite la flexibilité du système. Une seconde limitation est que cette méthode de synthèse est limitée à des oligonucléotides courts (25 mères) dont la sensibilité est réduite. Cela est très largement compensé par l'avance technologique de cette société, qui est présente dans ce domaine depuis les années 1990 (elle possède d'ailleurs de très nombreux brevets), et par une capacité d'intégration de sondes inégalée (plusieurs millions de sondes par puce). Elle dispose de puces à ADN pour étudier le transcriptome de nombreuses espèces même partiellement séquencées sous forme d'EST. Elle dispose de plus de puces couvrant l'ensemble du génome, le génome mitochondrial, certains chromosomes particuliers, les exons, les SNP, les régions promotrices pour l'homme, la souris et d'autres espèces. Ces puces couvrent la plupart des applications des puces à ADN: expression, génotypage, reséquenc¸age, ChIP-chip... La société offre même un service de puces à fac¸on, un peu limité par le prix des masques.

Le principal concurrent d'Affymetrix, utilise une synthèse in situ d'oligonucléotides longs sans masque au moyen d'imprimantes à jet d'encre (Agilent). L'imprimante sert à déposer la base appropriée à ajouter à chaque oligonucléotide en cours de synthèse. Cette technologie offre donc une flexibilité maximale. L'utilisateur peut choisir les séquences de ses oligonucléotides, lesquelles sont envoyées aux ordinateurs pilotant les imprimantes, et obtenir ainsi sa propre puce à ADN. Cette flexibilité permet toutes les applications pour toutes les espèces séquencées même partiellement.

Une technologie très voisine de celle d'Affymetrix utilise la synthèse in situ d'oligonucléotides longs avec des masques programmables (Nimblegen). Les masques préusinés sont ici remplacés par micromiroirs programmables. Cette technologie offre aussi une flexibilité maximum et permet de synthétiser à volonté des oligonucléotides de taille variable et de n'importe quelle séquence. Cette technologie est limitée par les nombreux brevets possédés par Affymetrix dans le domaine des puces à ADN. Ces obstacles ont été partiellement levés par l'installation de la société en Islande, pays dans lequel Affymetrix n'avait pas étendu ces brevets, puis plus récemment par un partenariat avec Affymetrix. Malgré cela cette technologie reste encore confidentielle. Le rachat récent

de cette société par Roche (le leader mondial dans le domaine de la biologie) pourrait peut-être changer les choses dans l'avenir.
Une dernière technologie très similaire à la précédente utilise des micromiroirs programmables pour synthétiser in situ des oligonucléotides longs (Febit). Cette société possède l'originalité de vendre une machine très compacte permettant à n'importe qui de réaliser la synthèse, l'hybridation et la lecture de puces à ADN chez soi. Cette machine permet d'analyser de huit à 16 puces à ADN par jour, avec une flexibilité inégalée. Cette technologie encore peu répandue, préjuge peut-être de l'avenir des puces à ADN où chaque utilisateur pourra analyser la séquence ou l'expression de n'importe quel génome sur sa paillasse.

Références

[1] Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach H. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. Mamm Genome 1992;3(11):609–19.
[2] Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, et al. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. Genomics 1995;29(1):207–16.
[3] Zhao N, Hashida H, Takahashi N, MisumiY, SakakiY. High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156(2):207–13.
[4] Pietu G, Alibert O, Guichard V, Lamy B, Bois F, Leroy E, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6(6):492–503.
[5] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270(5235):467–70.
[6] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. Expression monitoring by hybridization to high-density
oligonucleotide arrays. Nat Biotechnol 1996;14(13):1675–80. [7] Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res 2000;28(22):4552–7.
[8] Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, et al. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet 1992;2(3):173–9.
[9] Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, et al. Sequence identification of 2,375 human brain genes. Nature 1992;355(6361):632–4.
[10] Houlgatte R, Mariage-Samson R, Duprat S, Tessier A, Bentolila S, Lamy B, et al. The Genexpress Index: a resource for gene discovery and the genic map of the human genome. Genome Res 1995;5(3):272–304.
[11] Auffray C, Behar G, Bois F, Bouchier C, Da Silva C, Devignes MD, et al. IMAGE: molecular integration of the analysis of the human genome and its expression. C R Acad Sci III 1995;318(2):263–72.
[12] Lennon G, Auffray C, Polymeropoulos M, Soares MB. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics 1996;33(1):151–2.
[13] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000;405(6788):827–36.
[14] Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci U S A 1990;87(5):1663–7.
[15] Phillips J, Eberwine JH. Antisense RNA amplification: A linear amplification method for analyzing the mRNA population from single living cells. Methods 1996;10(3):283–8.
[16] Wang E, Miller LD, Ohnmacht GA, Liu ET, Marincola FM. High fidelity mRNA amplification for gene profiling. Nat Biotechnol 2000;18(4): 457–9.
[17] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.
[18] Thieblemont C, Nasser V, Felman P, Leroy K, Gazzo S, Callet-Bauchu E, et al. Small lymphocytic lymphoma, marginal zone B-cell lymphoma, and mantle cell lymphoma exhibit distinct gene-expression profiles allowing molecular diagnosis. Blood 2004;103(7):2727–37.
[19] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403(6769):503–11.
[20] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98(19): 10869–74.
[21] Bertucci F, Finetti P, Rougemont J, Charafe-Jauffret E, Cervera N, Tarpin C, et al. Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. Cancer Res 2005;65(6):2170–8.
[22] Ballester B, Ramuz O, Gisselbrecht C, Doucet G, Loi L, Loriod B, et al. Gene expression profiling identifies molecular subgroups among nodal peripheral T-cell lymphomas. Oncogene 2006;25(10):1560–70.
[23] Bertucci F, Nasser V, Granjeaud S, Eisinger F, Adelaide J, Tagett R, et al. Gene expression profiles of poor-prognosis primary breast cancer correlate with survival. Hum Mol Genet 2002;11(8):863–72.
[24] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24(3):227–35.
[25] Scherf U, Ross DT,Waltham M, Smith LH, Lee JK, Tanabe L, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24(3):236–44.
[26] Zembutsu H, Ohnishi Y, Tsunoda T, Furukawa Y, Katagiri T, Ueyama Y, et al. Genome-wide cDNA microarray screening to correlate gene expression profiles with sensitivity of 85 human cancer xenografts to anticancer drugs. Cancer Res 2002;62(2):518–27.
[27] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, et al. Genome-wide location and function of DNA binding proteins. Science 2000;290(5500):2306–9.
[28] Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 2001;409(6819):533–8.

[29] Fernandez PC, Frank SR, Wang L, Schroeder M, Liu S, Greene J, et al. Genomic targets of the human c-Myc protein. Genes Dev 2003;17:1115–29.
[30] Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 1999;23(1):41–6.
[31] Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, et al. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet 2005;37(1): 48–55.
[32] Jansen RC, Nap JP. Genetical genomics: the added value from segregation.
Trends Genet 2001;17(7):388–91.

Tables et figures

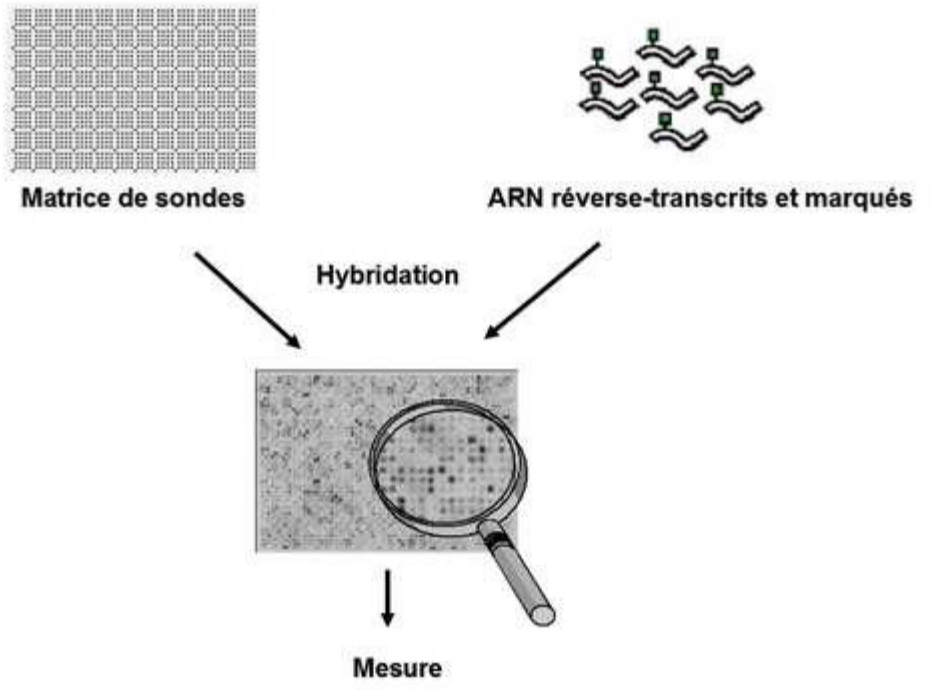


Figure 1 : Principe des puces à ADN.
Une puce à ADN est un réseau régulier et très dense de sondes d'ADN déposé sur un support solide. Chaque sonde est spécifique d'un gène donné. Elle est hybridée avec de l'ARN provenant d'un échantillon biologique, réverse-transcrits et marqués. Le signal observé sur chaque sonde reflète l'abondance dans l'échantillon biologique étudié, de l'ARN correspondant à cette sonde. L'application présentée ici est l'analyse du transcriptome.

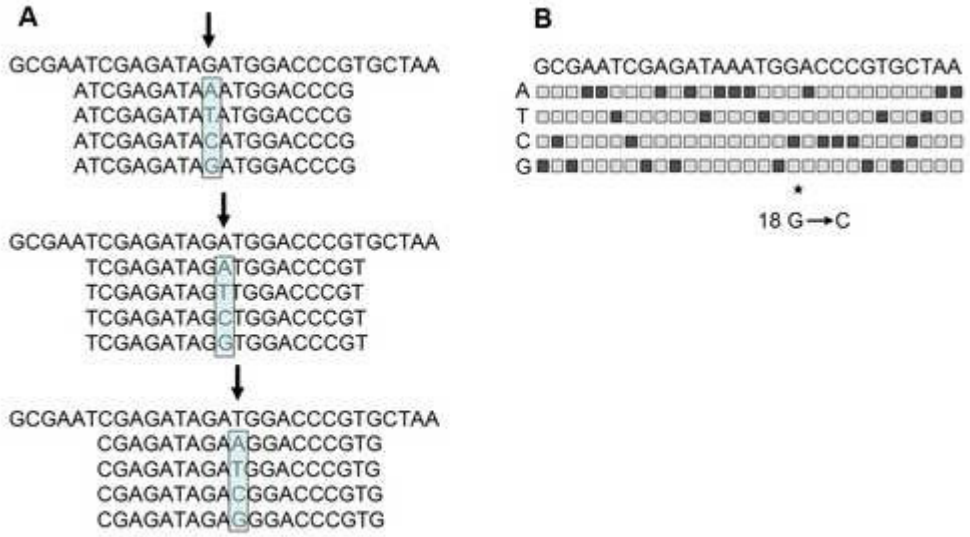


Figure 2 : Principe du reséquençage de gène
A : Conception des sondes. Pour chaque position (marquée d'une flèche) de la séquence que l'on veut reséquencer (la première séquence de chaque bloc), on conçoit 4 oligonucléotides courts (les 4 dernières séquences de chaque bloc) centrés sur cette position et possédant en position centrale les 4 bases (A, T, C ou G) possibles. Pour un gène de taille N, il faut donc 4N oligonucléotides.
B : Hybridation de la puce à ADN. L'ensemble de ces oligonucléotides sont ensuite déposés sur une puce à ADN qui est hybridée avec l'ADN que l'on veut reséquencer. Chaque carré représente un oligonucléotide centré sur les différentes positions de la séquence à reséquencer (la séquence en haut) et possédant en position centrale la base située à gauche. Les signaux d'hybridations (carrés noirs) permettent la lecture de la séquence à déterminer. Les signaux d'hybridation montrent un remplacement d'un G par un C en position 18 (position marquée d'une étoile noire).

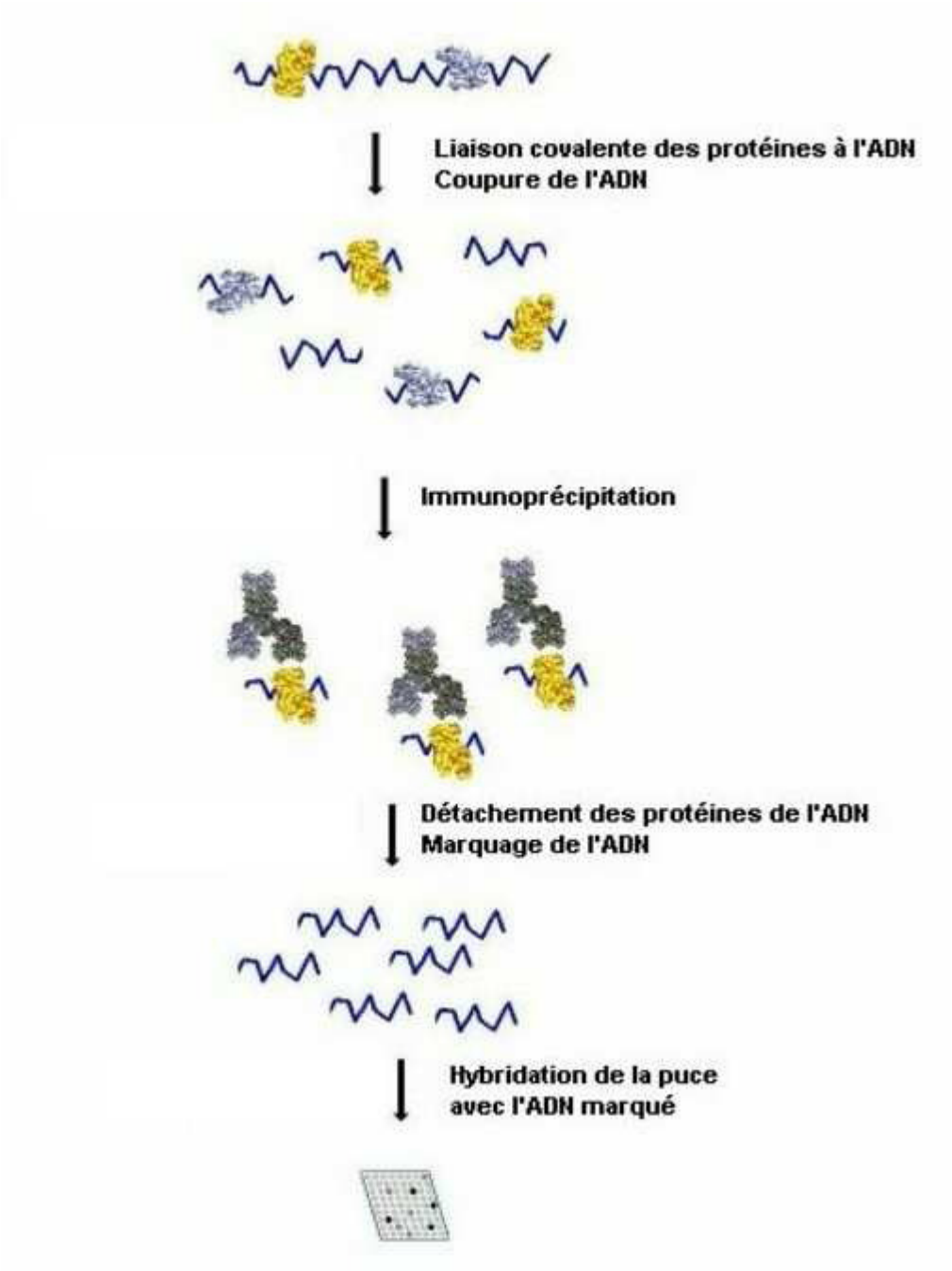


Figure 3 : Principe du ChIP-chip
L'objectif du ChIP-chip est de déterminer les sites de fixation d'un facteur de transcription sur la chromatine (ici, le facteur jaune). La première étape consiste à lier de façon covalente l'ADN aux protéines au moyen par exemple de formaldéhyde, et de couper l'ADN par sonication. Le facteur à étudier est ensuite immunoprécipité (ChIP : Chromatine ImmunoPrecipitation) au moyen d'un anticorps spécifique (la molécule noire). En immunoprécipitant le facteur étudié, on immunoprécipite simultanément les séquences génomiques sur lesquels il est fixé. Après une étape de réversion des liaisons covalentes, l'ADN est marqué et hybridé sur une puce à ADN (chip) constituée de régions promotrices d'îlots CpG ou couvrant l'ensemble du génome. Les signaux d'hybridation correspondent alors aux différentes régions sur lesquelles était fixé le facteur de transcription.

Tableau 1A
Applications des puces à ADN

Matériel hybridé	Applications	Référence
ADNc	Criblage de banque	[1]
Oligonucléotides	Caractérisation par empreinte d'oligonucléotides	Drmanac S. et al. (1996) Genomics 37: 29–40
ARNm	Mesure du transcriptome	[5]
ARNm	Mesure de la taille des ARNm	Hurowitz et al. (2003) GenBiol 5: R2
ARNm	Détection de gènes	Penn et al. (2000) Nature Genet 26: 315–318
ARNm	Validation de gènes	Shoemaker et al. (2001) Nature 409: 922–927
ADN	Reséquençage de gène	Kozal MJ et al. (1996) Nat Med 2: 753–759
ADN	Phénotypage	Gingeras et al. (1998) Genome Res 8: 435–448
ADN	Génotypage CGH	[30]
ADN	Génotypage SNP	Winzeler et al. (1998) Science 281: 1194–1197
Plasmide	Suivi de réplication	Khodursky et al. (2000) PNAS 97: 9419–9424
Protéine	Sites reconnus par un facteur de transcription	Bulyk et al. (1999) Nature Biotech 17: 573–577
ADN	Sites de fixation d'un facteur de transcription sur la chromatine	[28]

Tableau 1B
Applications des puces à ADN au niveau du transcriptome

Matériel hybridé	Propriété mesurée	Référence
ARNm	Transcriptome	[5]
ARNm	Taux de dégradation des ARNm	Wang et al. 2002 PNAS 99: 5860–5865
ARNm liés aux polysomes	Taux de traduction des ARNm	Johannes et al. 1999 PNAS 96: 13118–13123
ARNm liés aux membranes	Localisation subcellulaire des protéines	Diehn et al. 2000 Nature Genetics 25: 58–62
ARNm liés aux polysomes mitochondriaux	Localisation subcellulaire des protéines	Marc et al. 2002 EMBO reports 31 (21)

TITRE : De la génomique fonctionnelle vers la génomique intégrative de pathologies humaines.

L'achèvement du séquençage du génome humain ainsi que celui de nombreux organismes a contribué à l'explosion de la génomique. Cette discipline s'attache notamment à la façon dont un génome s'exprime, au mode d'action des gènes, à l'étude de l'ensemble des produits qu'il code dans un système biologique donné. Des technologies de mesure du statut du génome à grande échelle, comme les biopuces que nous avons utilisées dans cette thèse, ont été mises en oeuvre. Ce travail a porté sur l'étude des variations du transcriptome dans différentes conditions physiologiques ou pathologiques. Nous avons pu mesurer l'influence de facteurs génétiques et environnementaux, détecter des biomarqueurs ou des profils d'expression spécifiques de sous-classe pathologiques dans des lymphomes et dans des maladies thyroïdiennes. Afin de comprendre des mécanismes moléculaires sous-jacents de ces modifications d'expression, nous avons progressivement intégré d'autres données génomiques. Ceci incluait la détection de délétions ou d'amplifications de régions génomiques par puces CGH, ou l'identification de sites de liaison de facteurs de transcription sur l'ADN par analyse bioinformatique ou par ChIP-chip. Par cette approche combinée, une modélisation de réseaux biologiques est alors envisageable. Elle permettra de mieux comprendre le fonctionnement d'un système biologique, suscitant de nombreux espoirs dans la recherche de cibles thérapeutiques encore plus pertinentes.

Mots-clés: Génomique, génomique intégrative, transcriptome, bioinformatique, découverte de motifs, CGH, site de liaison de facteur de transcription, ChIP-chip

TITLE: From functional genomics to integrative genomics of human pathologies

The complete human genome sequence has contributed to the expansion of genomics. This field notably describes how a genome is expressed, how some sets of genes and their products work together in biological systems. High-throughput technologies for genome research, like microarrays which were used in this thesis, were set up. This work deals with transcriptomic variations observed in different physiological or pathological conditions. We appreciated the effect of genetic and environmental factors on expression profiles, to detect biomarkers specific to pathological subclasses of lymphoma and thyroid lesions. To understand molecular mechanisms underlying these gene expression modifications, we integrated gradually other genomic data. This included detections of genomic deletions or amplifications using CGH arrays, or the identification of transcription factor binding sites by sequence analysis or by ChIP-chip methodology. With this combined approach, modeling biological networks modeling is then conceivable. It will allow a better understanding of a biological system and to detect more reliable therapeutic targets.

Keywords: Genomics, integrative genomics, microarrays, transcriptome, bioinformatics, motif discovery, CGH, Transcription factor binding sites, ChIP-chip.

RAHARIJAONA Mahatsangy
IRT - UN
l'institut du thorax
INSERM UMR 915
8 quai Moncousu
BP 70721
44007 Nantes Cedex 1