UNIVERSITE MOHAMMED V
Faculté De Médecine et de Pharmacie
Centre d'Etudes Doctorales des Sciences de la Vie et de la santé
CEDOC-SVS

Année : 2017 THÈSE N° : 20/16CSVS

THESE DE DOCTORAT

Formation Doctorale : Biologie médicale, pathologie humaine et expérimentale et environnement

Présentée Par
Mme. Imane SABAOUNI

Identification des réseaux de gènes qui interviennent dans le phénotype myocardiale lie au CD36

" Analyse d'expression des gènes de trois modèles des souris "

Soutenu publiquement le 19/07/2017 devant le jury

Pr. Abderrahmane SBIHI	**Ecole National des Sciences Appliquées de Tanger ENSAT, Université Abdelmalek Essaâdi – Tétouan**	**Président**
Pr. Azeddine IBRAHIMI	**Laboratoire de Biotechnologie Médicale - MedBiotech- Faculté de Médecine et de Pharmacie, Université Mohammed V de Rabat**	**Directeur de thèse**
Pr. Rachida AMRI	**Centre Hospital-Universitaire Ibn Sina de Rabat- Hôpital d'Enfants de Rabat**	**Examinateur**
Pr. Naima HAFIDI	**Centre Hospital-Universitaire Ibn Sina de Rabat- Hôpital d'Enfants service des maladies infectueuses.**	**Examinateur**
Pr. Ahmed MOUSSA	**Ecole National des Sciences Appliquées de Tanger ENSAT, Université Abdelmalek Essaâdi – Tétouan**	**Rapporteur**
Pr. Rachid EL JAOUDI	**Faculté de Médecine et de Pharmacie, Université Mohammed V de Rabat**	**Rapporteur**

سبحانك لا علم لنا إلا ما علمتنا إنك أنت العليم الحكيم

سورة البقرة: الآية: 31

اللهم إنا نسألك علما نافعا وقلبا خاشعا
وشفاءا من كل داء وسقم

Résumé

Le gène CD36 code pour une protéine membranaire multiligand qui facilite le transport des acides gras à longues chaînes (AGLC) dans les tissus musculaires et fixe la thrombospondine sur les membranes des cellules endothéliales.

Les études au laboratoire utilisant le souris CD36-Knockout (KO) ont démontré que l'absence totale d'expression de la protéine CD36 impact la fonction cardiaque ainsi que la sensibilité à l'insuline dans les muscles squelettiques. Les cœurs de ces souris CD36-KO montrent, entre autres, des signes de cardiomyopathie avec une utilisation accrue de glucose au lieu des AGLC.

Dans l'une des études décrite dans ce manuscrit de thèse, nous avons pu corrigés ces dysfonctionnements en générant un modèle animal ; CD36-GR, où l'expression du gène CD36 a été rétablie dans les cœurs et des muscles des souris KO.

Dans une deuxième partie et afin d'identifier les réseaux de gènes impliqués dans l'hypertrophie cardiaque observée chez les souris KO en absence du gène CD36, une analyse globale d'expression des gènes en utilisant la technologie des puces à ADN a été réalisée. Nos résultats ont permis d'établir que l'absence d'expression du gène CD36 conduit à une régulation d'expression de trois groupes de gènes : ceux impliqués dans le métabolisme des AGLC, ceux inhibant l'angiogenèse et enfin des gènes de la restructuration cellulaire. Ces résultats semblent être cohérents avec les rôles joués par la protéine CD36 dans les cellules cardiaque et endothéliales.

Dans un troisième article et basée sur l'analyse des ARN et des réseaux des gènes et des protéines, cette étude a permit d'établir un répertoire de gènes exprimés au cours du développement de la myocardiopathie et d'identifier les gènes potentiellement impliqués.

En conclusion, ces travaux de thèse ont permit de caractériser les mécanismes de développement de la myocardiopathie liée à la protéine CD36 et de concevoir des stratégies thérapeutiques aux anomalies cardiaques liées à l'absence d'expression du gène CD36.

Mots clés : CD36, acides gras à longue chaine, Microarray, angiogenése, myocardiopathie hypertrophique, réseaux des gènes, réseaux des protéines.

Summary

CD36 gene encodes for a multiligand membrane protein that facilitates the transport of long chain fatty acids (LCFA) in muscle tissues and is involved in the binding of the thrombospondin to endothelial cell membrane.

Different studies using CD36-Knockout (KO) mouse have demonstrated that the complete lack of CD36 protein expression impacts cardiac function and insulin sensitivity in skeletal muscles. In fact, the hearts of these CD36-KO mice showed, among others, various signs of cardiomyopathy with a significant use of glucose instead of LCFA.

In one of the studies described in this thesis, we were able to solve these dysfunctions by generating a specific animal model, CD36-GR, where the expression of CD36 gene was restored in the hearts and the muscles of the KO mice.

In the second part of the thesis, a global analysis of genes expression using DNA chip technology was performed. The objective of this analysis is to identify the genes network involved in the cardiac hypertrophy observed in the KO mice in the absence of the CD36 gene. The results have demonstrated that the absence of CD36 gene expression leads to the expression regulation of three groups of genes; one group is involved in the metabolism of LCFAs, another one is responsible of the inhibition of angiogenesis and the last one is implicated in the cellular restructuring process. These outcomes are consistent with the roles of CD36 protein in cardiac and endothelial cells.

In a third study, the analysis of RNAs and genes/proteins networks has permitted to establish a repertory of genes that are expressed during the development of myocardiopathy. Moreover, it has helped identifying the genes potentially involved in this process.

In conclusion, this work has allowed to characterize various mechanisms of cardiomyopathy development related to the CD36 protein and to design therapeutic strategies to solve cardiac abnormalities that are specifically related to the absence of CD36 gene expression.

Key words: CD36, long chain fatty acids, Microarray, angiogenesis, hypertrophic cardiomyopathy, genes network, proteins network.

الملخص

الجينات CD36 تشفر بروتين غشائي الذي يسهل نقل الأحماض الدهنية طويلة السلسلة و ترومبوسبوندين في الأنسجة العضلية ثابتة على أغشية الخلايا البطانية.

وقد أظهرت الدراسات المخبرية باستخدام الفئران ذات الغياب التام للتعبير عن البروتين CD36 (KO) وظيفة تأثير القلب والحساسية للانسولين في العضلات والهيكل العظمي. قلوب هذه الفئران (KO) تظهر، من بين أمور أخرى، وعلامات اعتلال عضلة القلب مع زيادة استخدام الجلوكوز بدلا من الأحماض الدهنية.

في واحدة من الدراسات الموضحة في هذه المخطوطة أطروحة، كنا قادرين على تصحيح هذه الخلل عن طريق توليد نموذج حيواني. Gr-CD36، حيث تم استعادة التعبير عن الجين CD36 في قلوب وعضلات الفئران KO .

في الجزء الثاني و لأجل تحديد شبكات الجينات المعنية في تضخم القلب الملحوظة عند الفئران KO في غياب الجينات CD36، أجري تحليل التعبير الجيني العالمي باستخدام تكنولوجيا ميكروأري. وقد أظهرت نتائجنا أن غياب التعبير عن CD36 الجين يؤدي في تنظيم التعبير من ثلاث مجموعات من الجينات: المتورطين في عملية التمثيل الغذائي ، وتلك التي تمنع تكوين الأوعية الدموية، وأخيرا إعادة هيكلة الجينات الخلية. ويبدو أن هذه النتائج تتفق مع الأدوار التي يلعبها البروتين CD36 في خلايا القلب والخلايا البطانية.

في المادة الثالثة، واستنادا إلى تحليل الحمض النووي وشبكات الجينات والبروتينات، سمحت هذه الدراسة إلى إنشاء مرجع الجينات التي أعرب عنها خلال تطوير عضلة القلب وتحديد الجينات المحتمل أن تشارك.

في الختام، لقد سمح العمل في هذه أطروحة لوصف آليات تطوير عضلة القلب المتعلقة بCD36 وتصميم الاستراتيجيات العلاجية لتشوهات القلب المتعلقة بنقص التعبير عن الجينات CD36.

الكلمات المفتاحية: CD36، الأحماض الدهنية الطويلة السلسلة، ميكروأري، الأوعية الدموية، اعتلال عضلة القلب الضخامي، شبكات الجينات، شبكات البروتين.

Remerciement

J’aimerais exprimer ma profonde gratitude au professeur ***AZEDDINE IBRAHIMI*** Pour m’avoir accueillie dans son laboratoire, pour m’avoir généreusement offert la liberté de défricher des voies de recherche peu usuelles. Sa "manie" d’exiger la précision et la clarté de nos travaux, sa passion d’effectuer des croisements fertilisants entre les disciplines scientifiques ont été pour moi des leçons précieuses de la créativité du chercheur. Je garderai le souvenir de sa "méthode tordue" de m’assaillir inlassablement de directions et de pistes insolites, mais qui aboutissent sur des idées plus simples et plus élégantes, et je le remercie pour m’accorder ce sujet.

Un très grand Merci au responsable du Centre d’Etudes Doctorales des Sciences de la Vie et de la santé CEDOC-SVS ***Pr. Jamal TAOUFIK***, qui s’est montrée aimable et tellement compréhensive et m’a facilité l’intégration au sein du centre et n’a pas hésité une seconde à me présenter une aide précieuse afin d’avoir accès à l’information.

Au terme de ce travail, il m’est particulièrement agréable d’exprimer mes vifs remerciements et ma profonde gratitude Monsieur le ***Pr. Ahmed MOUSSA*** professeur à l’ENSA de Tanger, pour m’avoir soutenue et pour le temps et la patience qu’il m’a accordé le long du mémoire, je voudrais aussi le remercier pour ces précieux conseils, son exigence, son commentaire et ses critiques ont été très utiles pour structurer mon travail aussi bien que son suivi continu.

Je remercie ***Pr. Hassan BADIR*** avec qui j’ai eu la chance de pouvoir travailler. Sa rigueur, sa capacité d’analyse des problèmes et ses très nombreuses connaissances m’ont permis de progresser et ont répondu à plusieurs de mes préoccupations.

Mes plus sincères remerciements vont également à Madame la ***Pr. BRIGITTE VANNIER***, professeur à la Faculté de Sciences Fondamentales et Appliquées, Université de Poitiers - Campus, pour l’intérêt particulier qu’elle m’a accordée, son expérience et de sa compétence, ces conseils et ses commentaires auront été fort utiles.

Un énorme merci à tout le personnel du laboratoire de la biotechnogie MedBiotech de la faculté de médecine et de pharmacie de Rabat pour m'avoir permit une intégration facile au sein de leur établissement et une facilité à l'information irréprochable.

Je tiens à remercier ***Pr. Abderrahmane SBIHI*** d'avoir accepté d'être président du jury. Je remercie également tous les membres du jury d'avoir accepté d'assister à la présentation de ce travail, particulièrement le ***Pr. Rachid EL JAOUDI, Pr. Rachida AMRI et Pr Naima HAFIDI*** pour avoir généreusement accepté de rapporter et d'examiner mon travail.

Je tenais à remercier tous mes professeurs qui m'ont transmis, ne serait-ce qu'une partie de leurs impressionnantes connaissances et savoir.

Quoi que je puisse dire ou faire, ce ne sera jamais assez pour exprimer ma gratitude vis-à-vis de ces gens qui nous ont tellement donné et qui n'attendent pas vraiment grand-chose en retour, Merci, Mille Mercis à tous.

Dédicaces

A cœur vaillant rien d'impossible

A conscience tranquille tout est accessible

Quand il y a la soif d'apprendre
Tout vient à point à qui sait attendre

Quand il y a le souci de réaliser un dessein
Tout devient facile pour arriver à nos fins

Malgré les obstacles qui s'opposent
En dépit des difficultés qui s'interposent

Les études sont avant tout
Notre unique et seul atout

Ils représentent la lumière de notre existence
L'étoile brillante de notre réjouissance

Comme un vol de gerfauts hors du charnier natal
Nous partons ivres d'un rêve héroïque et brutal

Espérant des lendemains épiques
Un avenir glorieux et magique

Souhaitant que le fruit de nos efforts fournis
Jour et nuit, nous mènera vers le bonheur fleuri

Aujourd'hui, ici rassemblés auprès des jurys,
Nous prions dieu que cette soutenance
Fera signe de persévérance
Et que nous serions enchantés
Par notre travail honoré

Je dédie cette thèse à …?

A ma très chère mère lalla Touria

Affable, honorable, aimable : Tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi.

Ta prière et ta bénédiction m'ont été d'un grand secours pour mener à bien mes études.

Aucune dédicace ne saurait être assez éloquente pour exprimer ce que tu mérites pour tous les sacrifices que tu n'as cessé de me donner depuis ma naissance, durant mon enfance et même à l'âge adulte.

Tu as fait plus qu'une mère puisse faire pour que ses enfants suivent le bon chemin dans leur vie et leurs études.

Je te dédie ce travail en témoignage de mon profond amour. Puisse Dieu, le tout puissant, te préserver et t'accorder santé, longue vie et bonheur.

A mon très cher père sidi abdelmajid:

Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours pour vous.

Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être.

Tu es l'exemple que j'admire pour toutes les peines et les sacrifices que tu as consentis pour mon éducation et ma formation.

Ce travail est le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation.

Puisse dieu vous protèges et vous donner santé et long vie.

A mon très cher mari sidi said

Quand je t'ai connu, j'ai trouvé l'homme de ma vie, mon âme sœur et la lumière de mon chemin.

Ma vie à tes cotés est remplie de belles surprises.

Tes sacrifices, ton soutien moral et matériel, ta gentillesse sans égal, ton profond attachement m'ont permis de réussir mes études.

Sans ton aide, tes conseils et tes encouragements ce travail n'aurait vu le jour.

Que dieu réunisse nos chemins pour un long commun serein et que ce travail soit témoignage de ma reconnaissance et de mon amour sincère et fidèle.

A mes très chers frères et sœurs :

Mohamed, Amine, Nada, et Aya.

Veuillez trouver dans ce travail, le témoignage de mes sentiments les plus sincères et les plus affectueux. Que Dieu le tout-puissant, vous accorde longue vie, prospérité et bonheur.

A toutes mes familles, SABAOUNI, ELKASSIMI et ELBASRI, dont je ne saurai passer sous silence.

A tous mes collègues et amis et amies.

J'espère n'avoir oublié personne et si c'est le cas je m'en excuse.

Je vous aime

LISTE DES ABRÉVIATIONS

AG **:** Acide Gras

ADN : Acide DésoxyriboNucléique ;

ADNc : Acide DésoxyriboNucléique complémentaire ;

ADP : Adénosine DiPhosphate;

ATP : Adénosine TriPhosphate;

CD36 : Cluster of Differentiation 36;

GO : Gene Ontology;

Gr : GeneRescuse;

HGNC : *HUGO Gene Nomenclature Committee;*

KEGG : *Kyoto Encyclopedia of Genes and Genomes*;

Ko : knockout;

SHR : Rat Spontanément H ypertendu ;

Wt : Wild type ;

TABLE DES ILLUSTRATIONS

Sommaire

Article 2

Article 3

Chapitre I : Introduction Générale

1. Le récepteur éboueur de type B: le CD36

Le CD36 est un récepteur membranaire intégral qui fait partie de la grande famille des récepteurs éboueurs ou scavengers. Cette famille de récepteurs transmembranaires comprend plusieurs sous-familles de récepteurs qui présentent une grande variété structurelle incluant, entre autres, les récepteurs scavengers de classes A (SR-A), dont SR-AI, SR-AII et MARCO (*macrophage receptor with collagenous structure*) et les récepteurs scavengers de classe B (SR-B), qui incluent, outre notre récepteur d'intérêt CD36, CLA-1 (ou SR-BI chez la souris), LIMP-2 (*lysosomal integral membrane protein-2)* aussi appelé SR-BII, Emp et Croquemort. D'autres récepteurs scavengers sont aussi connus dont LOX-1 (*lectin-like oxidized low-density lipoprtotein receptor*) et CD68 (macrosialine). De façon générale, les récepteurs scavengers sont impliqués dans l'internalisation de certains ligands, dans l'adhésion cellulaire ainsi que dans le transfert d'acides gras de part et d'autre de la membrane plasmique. Le CD36 a d'abord été identifié sur la membrane des plaquettes sanguines comme étant la glycoprotéine IV (gpIV) (Tandon *et al.*, 1989a; Tandon *et al.*, 1989b). Quelques années plus tard, des études ont montré que le CD36 était aussi un récepteur pour les lipoprotéines modifiées exprimé par les monocytes/macrophages (Endemann *et al.*, 1993) et un transporteur pour les acides gras au niveau des adipocytes (Abumrad *et al.*, 1993). Le CD36 est aussi connu sous l'appellation FAT/CD36 (translocase d'acide gras), SCARB3, GP88 et gpIIIb. Il est exprimé par une grande variété de cellules, ce qui en fait un récepteur multifonctionnel jouant un rôle au niveau du système cardiovasculaire, le métabolisme des acides gras, l'angiogenèse et la biologie des plaquettes sanguines, ainsi que dans différentes pathologies telles l'athérosclérose, la maladie d'Alzheimer et le diabète (Febbraio & Silverstein, 2007).

1.1. Historique de la découverte du CD36

Le récepteur CD36 fut initialement identifié à partir de membranes de plaquettes sanguines à la fin des années 1970. Un premier groupe décrivit le CD36, alors nommé glycoprotéine IV (GPIV), comme une protéine de 87 kDa résistante à la protéolyse lorsque des plaquettes humaines étaient digérées par la trypsine et la chymotrypsine. Le récepteur CD36 a par la suite été identifié par un autre groupe comme étant l'antigène réagissant avec l'anticorps OKM5 et fut alors nommé antigène OKM5. L'immunoprécipitation de l'antigène OKM5 des monocytes a mené à la caractérisation d'une protéine d'environ 88 kDa, soit un poids moléculaire similaire à la GPIV. Les termes GPIV et antigène OKM5 furent remplacés par CD36. Cependant le récepteur porta d'autres noms, notamment FAT (fatty acid translocase), GPIIIb (glycoprotéine IIIb et PAS IV (periodic acid/Schiff-positive band IV).

1.2. Structure du récepteur CD36

Le CD36 est localisé sur le chromosome 7q11.2 chez l'humain et sur les chromosomes 4 et 5 chez le rat et la souris, respectivement. La séquence des nucléotides prédit une protéine transmembranaire de 471 acides aminés ayant un poids moléculaire d'environ 53 kDa. Le gène CD36 est constitué de 15 exons, mais seulement une partie de l'exon 3, l'exon 4 à 13 et une partie de l'exon 14 codent pour la protéine CD36 (Figure 1) (Collot-Teixeira et al., 2007; Rac et al., 2007). Ce récepteur est constitué d'un large domaine extracellulaire qui possède six cystéines reliées par trois ponts disulfure ainsi que plusieurs sites de N-glycolysation qui lui confère une protection contre la protéolyse. Ces sites glycosylés entraînent une augmentation du poids moléculaire de la protéine à environ 88 plutôt que 53 kDa (Gruarin et al., 2000; Hoosdally et al., 2009). De plus, le domaine extracellulaire présente plusieurs domaines fonctionnels dont une poche hydrophobique située entre les acides aminés 184 et 204 qui lui permet d'interagir avec le feuillet externe de la membrane plasmique (Figure 2) (Greenwalt et al., 1992). Le domaine extracellulaire est responsable de la liaison des ligands et contient des sites de liaison distincts pour les protéines contenant un domaine homologue à la thrombospondine-1 (TSP-1) (Pearce et al., 1995) et les lipoprotéines oxydées, entre autres. En particulier, la séquence située entre les acides aminés 155 et 183 a été identifiée comme étant le site responsable de la liaison avec les lipoprotéines de faible densité oxydées (LDLox) (Puente Navazo et al., 1996), les produits de glycation avancée (AGE) (Ohgami et al., 2001) et les sécrétines de l'hormone de croissance (GHRPs). La méthionine 169 a été identifiée comme le point de contact de l'hexaréline sur le CD36 (Demers et al., 2004).

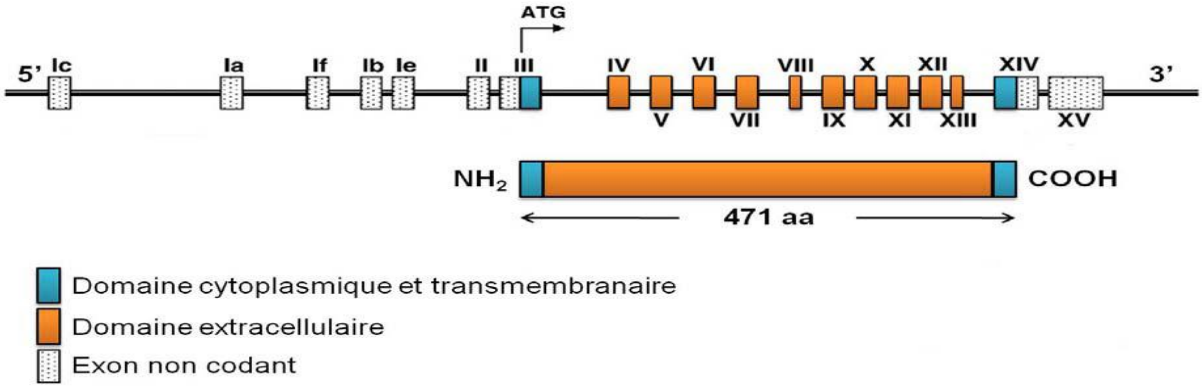


Figure 1. Représentation schématique du gène CD36

Le gène CD36 est constitué de 15 exons. La région non-transcrite en 5' est constituée de l'exon 1a, 1b, 1c, 1e, 1f, 2 et d'une partie de l'exon 3. Le restant de l'exon 3, les exons 4 à 13 et une partie de l'exon 14 codent pour la protéine CD36, tandis que le reste de l'exon 14 et l'exon 15 forment la région non-transcrite en 3' (Figure modifiée de (Collot-Teixeira *et al.*, 2007)).

D’autres domaines de liaison pour les LDLox ont été rapportés, notamment entre les acides aminés 28-93 et 120-155 (Pearce *et al.*, 1998). Il existe aussi deux domaines nommés CLESH (*CD36 LIMP-II Emp sequence homology*) qui sont impliqués dans la liaison et l’internalisation des cellules apoptotiques, et qui agissent de concert avec le complexe intégrine (αvβ3)/TSP-1 (Savill *et al.*, 1992; Navazo *et al.*, 1996; Dawson *et al.*, 1997; Simantov et al., 2001). La protéine CD36 comporte aussi deux domaines transmembranaires et deux petits domaines cytoplasmiques de sept à treize acides aminés, chacun caractérisé par la présence de cystéines palmitoylées aux positions 3, 7, 464 et 466 (Figure 2) (Tao et al., 1996). La palmitoylation joue un rôle essentiel dans le positionnement du CD36 au niveau des cavéoles et des radeaux lipidiques (Febbraio & Silverstein, 2007). Les deux domaines cytoplasmiques, N- et C-terminal, sont nécessaires pour que le CD36 soit exprimé à la surface membranaire. Il a été montré que le domaine C-terminal joue un rôle important dans la liaison, l’internalisation et la dégradation des LDLox puisque la délétion du dernier acide aminé ou bien une modification des six derniers acides aminés du domaine cytoplasmique en C-terminal engendre une diminution de l’interaction entre le CD36 et le LDLox (Malaud et al., 2002). De plus, il a été montré que l’activation du nuclear factor kappa B (NFκB) en réponse à l’activation du récepteur CD36 par les LDLox nécessite que le domaine cytoplasmique en C-terminal soit intact (Lipsky et al., 1997). Plus récemment, l’importance des résidus Tyr463 et Cys464 au niveau de l’extrémité C-terminale a aussi été montrée dans la liaison du ligand au récepteur ainsi que dans la signalisation cellulaire (Stuart et al., 2005a). McDermott-Roe et al., ont, quant à eux, montré que contrairement au domaine cytoplasmique en C-terminal, le domaine cytoplasmique en N-terminal n’est pas essentiel à l’internalisation des LDLox par le CD36 (McDermott-Roe et al., 2008).

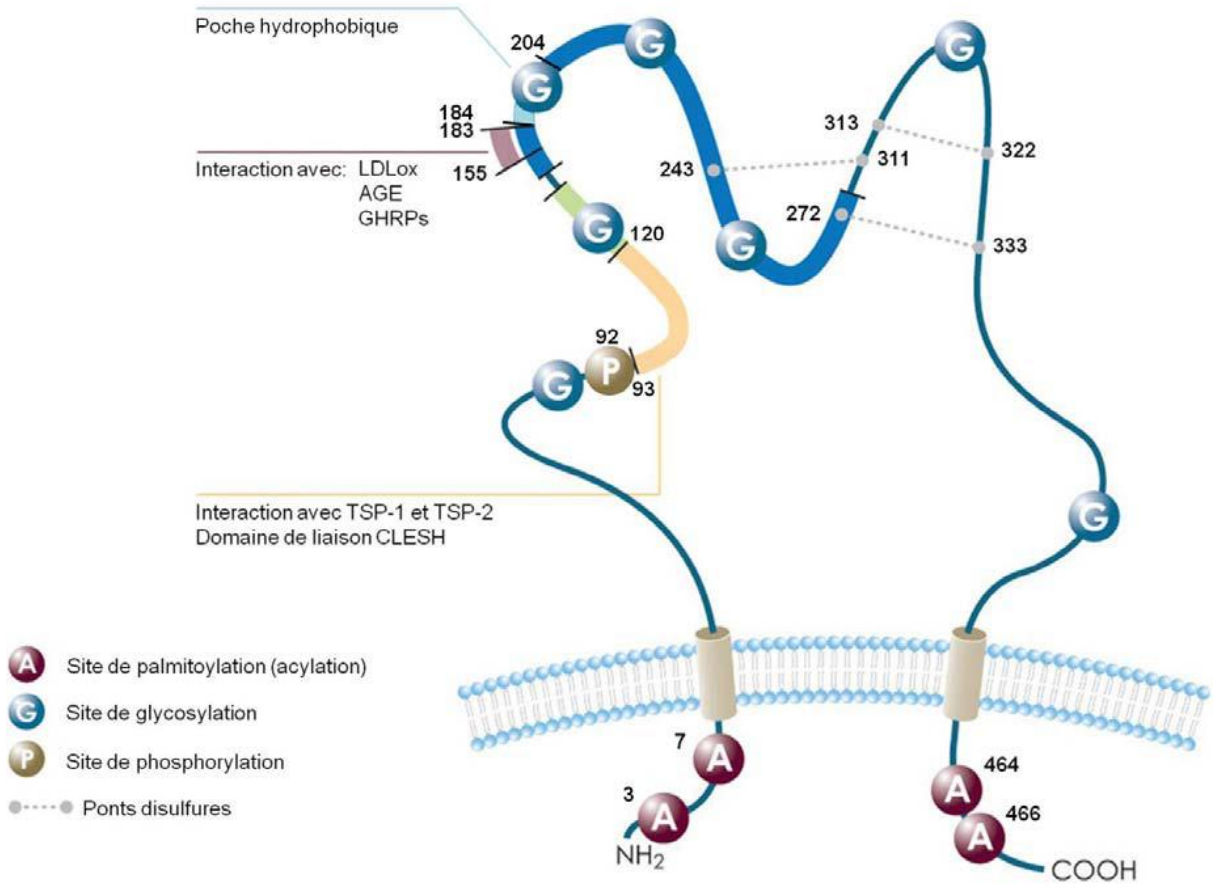


Figure 2. Structure du récepteur CD36

Le CD36 est un récepteur transmembranaire localisé dans les radeaux lipidiques de la membrane plasmique. Son large domaine extracellulaire contient trois ponts disulfures (Cys243-Cys311, Cys272-Cys333, et Cys313-Cys322), de multiples sites de N-glycosylation et différents domaines de liaison. Les deux petites queues cytoplasmiques contiennent des cystéines palmitoylées. La Tyr463 et Cys464 en C-terminal sont importantes pour la liaison des ligands et la cascade de signalisation intracellulaire (Figure inspirée de (Silverstein & Febbraio, 2009; Collot-Teixeira *et al.*, 2007).

1.3. Distribution et régulation de l'expression du récepteur CD36

Le CD36 a d'abord été identifié par Bolin et al. sur la membrane des plaquettes (Bolin *et al.*, 1981). Par la suite, Swerlick et al. ont montré que le CD36 est présent au niveau des cellules endothéliales de la microvasculature, contrairement aux cellules endothéliales des gros vaisseaux sanguins (Swerlick *et al.*, 1992). Les monocytes, et plus particulièrement les macrophages expriment aussi le CD36, et cette expression est régulée à la hausse après la différentiation des monocytes en macrophages (Endemann *et al.*, 1993; Huh *et al.*, 1996). Le récepteur CD36, de par son rôle facilitateur de la translocation des acides gras, est aussi exprimé dans les adipocytes, le muscle cardiaque et les muscles squelettiques ainsi que sur la muqueuse intestinale où l'on retrouve différents niveaux d'expression selon la région

intestinale. Le CD36 est principalement exprimé dans la partie proximale du petit intestin, incluant le duodénum et le jéjunum (Nassir *et al.*, 2007; Chen *et al.*, 2001a). Le récepteur est aussi exprimé au niveau des kératinocytes, des érythrocytes, des réticulocytes, ainsi qu'au niveau des cellules épithéliales mammaires et rénales, des cellules dendritiques, des cellules de la microglie et de l'épithélium pigmentaire rétinien (Febbraio *et al.*, 2001; Febbraio & Silverstein, 2007). Il a été montré que le CD36 peut aussi être exprimé au niveau des hépatocytes chez un modèle animal soumis à une diète riche en lipides (Koonen et al., 2007; Inoue et al., 2005), ainsi que dans les cellules musculaires lisses vasculaires (CMLV) (Matsumoto et al., 2000). Ainsi, le CD36 est exprimé dans une panoplie de La régulation de l'expression du récepteur CD36 (Tableau 1) a largement été étudiée dans les monocytes et les macrophages et quelques autres types cellulaires. L'expression du CD36 peut être régulée autant au niveau transcriptionnel, post-transcriptionnel que traductionnel et peut être modulée par l'adhérence et la différenciation cellulaire, ainsi que par divers médiateurs solubles ou par des récepteurs nucléaires, dont le principal est le peroxisome proliferator-activated receptor γ (PPARγ) (Chen et al., 2001b; Huh et al., 1995; Huh et al., 1996; Nicholson et al., 2000; Prieto et al., 1994; Tontonoz & Nagy, 1999; Tontonoz et al., 1998).

Tableau 1. Régulation de l'expression du récepteur CD36

Inducteurs	Types cellulaires	Mécanisme d'action	Références
Adhésion	monocytes/macrophages	Agit au niveau transcriptionnel	(Prieto *et al.*, 1994; Huh *et al.*, 1995)
Différentiation	monocytes/macrophages	Agit au niveau transcriptionnel	(Huh *et al.*, 1996)
M-CSF et GM-CSF	monocytes/macrophages	différentiation des monocytes en macrophages	(Yesner *et al.*, 1996)
PMA	monocytes/macrophages	différentiation des monocytes en macrophages	(Yesner *et al.*, 1996)
IL-4	monocytes/macrophages	Activation de PKC, activation de la lipoxygénase et production de 15d-PGJ2	(Feng *et al.*, 2000; Yesner *et al.*, 1996; Huang *et al.*, 1999)
cholestérol cellulaire	monocytes/macrophages	Agit au niveau transcriptionnel	(Han *et al.*, 1999)
LDL native	monocytes/macrophages	Agit au niveau transcriptionnel	(Han *et al.*, 1997)
LDLox (9-HODE et 13-HODE)	monocytes/macrophages	Activation du PPARγ	(Nagy *et al.*, 1998; Han *et al.*, 1997; Feng*et al.*, 2000)
15-HETE	monocytes/macrophages	Augmentation de	(Huang *et al.*, 1999

		l'expression d'IL-4 et PPARγ	
15d-PGJ2	monocytes/macrophages	Ligand naturel du PPARγ et activation du PPARγ	(Zuckerman et al., 2000; Han & Sidell, 2002)
4-Hydroxynonénal (4-HNE)	monocytes/macrophages	Activation de la kinase p38 et de la 5-lipoxygénase	(Yun et al., 2008; Yunet al., 2009)
Insuline/glucose	adipocytes coeur/intestin monocytes	Agit au niveau traductionnel Activation de la voie PI3-K Activation de PPARγ	(Griffin et al., 2001; Chabowski et al., 2004; Chen et al., 2006; Sampson et al., 2003; Yang et al., 2007)
9-cis Acide rétinoïque all-trans Acide rétinoïque	monocytes/macrophages	Activation des récepteurs de l'acide rétinoïque RAR/RXR	(Langmann et al., 2005; Wuttge et al., 2001)
Thiazolidinediones (TZD)	monocytes/macrophages adipocytes muscles	Ligands synthétiques qui activent le PPARγ	{4835;5685686; ;55687;5689}
Statines	monocytes/macrophages	Activation de COX-2 et production de 15d-PGJ2 Inhibe Rho GTPase	(Yano et al., 2007; Yeet al., 2007; Ruiz-Velasco et al., 2004)
Inhibiteurs de protéases du VIH	monocytes/macrophages	Activation PPARγ par PKC Inhibition du système ubiquitine-protéasome	(Dressman et al., 2003; Hui, 2003; Munteanu et al., 2004; Munteanu et al., 2005)
Ligands synthétiques de RAR et RXR	monocytes/macrophages	Activation de RAR et RXR, indépendemment du PPARγ	(Han & Sidell, 2002)
Ligands du FXR	Adipocytes	Activation PPARγ	(Abdelkarim et al., 2010)
Aspirine	monocytes/macrophages	diminution de PGE2	(Vinals et al., 2005)
Celecoxib	monocytes/macrophages	Inhibition COX-2, cPLA2 menant à une diminution PGE2	(Anwar et al., 2011)
Antagoniste récepteur AT1	Foie	Augmentation du récepteur PPARα	(Rong et al., 2010)
Agonistes PXR	Foie	Cible direct et activation du PPARγ	(Zhou et al., 2006; Zhou et al., 2008; Zhou et al., 2009)
Répresseurs	Types cellulaires	Mécanisme d'action	Références
LPS	monocytes/macrophages	NA	(Yesner et al., 1996)
Dexaméthasone	monocytes/macrophages	NA	(Yesner et al., 1996)
IL-10	monocytes/macrophages	Diminution expression PPARγ	(Rubic & Lorenz, 2006)
TGF β1 et β2	monocytes/macrophages	Activation de MAPK qui phosphoryle et inactive PPARγ	(Han et al., 2000)
Statines	monocytes/macrophages plaquettes	Diminution de NFκB	(Pietsch et al., 1996; Fuhrman et al., 2002a; Han et al., 2004; Puccetti et al., 2005; Bruni et al.,

			2005; Mandosi et al., 2010)
α-tocophérol et autres anti-oxydants	monocyes/macrophages tCMLV	Agit au niveau transcriptionnel Diminution production enzymatique 9 et 13-HODE Inhibition de Tyk2 Liaison au CD36 engendrant son internalisation	(Devaraj et al., 2001; Fuhrman et al., 2002b; Ricciarelli et al., 2000; Venugopal et al., 2004; Ozer et al., 2006; Zingg et al., 2010; Gieseg et al., 2010)
H2S (Sulfide d'hydrogène)	monocytes/macrophages	Inhibition de ERK1/2	(Zhao et al., 2011)
PGE2	monocytes/macrophages	Indéterminé	(Chuang et al., 2010)
Dérivés opioïdes	monocytes/macrophages	Indéterminé	(Chiurchiu et al., 2011)
IL-33	monocytes/macrophages	répression transcriptionnelle ou activation de AP-1 et NFκB	(McLaren et al., 2010)
Vitamine D	monocytes/macrophages	Diminution phosporylation JNK, diminution PPARγ	(Riek et al., 2010; Oh et al., 2009)
HDL, HDLox	monocytes/macrophages	Activation PPARγ (p38-dépendant)	(Han et al., 2002; Renet al., 2010; Carvalho et al., 2010)
Ligands du FXR	monocytes/macrophages	Indéterminé	(Mencarelli et al., 2009; Mencarelli et al., 2010)
Inhibiteurs de protéases du VIH (chez femelles)	monocytes/macrophages	Oestrogène prévient activation de PKC	(Allred et al., 2006)

La régulation de l'expression du CD36 passe aussi par la redistribution d'un pool intracellulaire du récepteur vers la membrane plasmique (Huh *et al.*, 1996; Luiken *et al.*, 2002; Muller *et al.*, 2002). Dans le muscle, le CD36, qui est emmagasiné au niveau de vésicules endosomales, est acheminé vers la membrane plasmique en réponse à une contraction ou après une stimulation par l'insuline (Luiken *et al.*, 2003; Jeppesen *et al.*, 2011; Luiken *et al.*, 2002; Van Oort *et al.*, 2008). Il est à noter que chez les sujets souffrant d'obésité et de résistance à l'insuline, le CD36 est retenu à la surface cellulaire en conséquence d'un dérèglement du recyclage du CD36 entre le compartiment intracellulaire et la membrane plasmique. Cette situation entraîne une accumulation de triacylglycérol dans le muscle conduisant à une utilisation préférentielle des acides gras comme substrat énergétique au détriment du glucose (Bonen et al., 2004; Coort et al., 2004).

1.4. Les ligands du récepteur CD36

1.4.1. Ligands endogènes

Le CD36 est un récepteur qui se lie à une multitude de ligands endogènes. Parmi ces ligands, on retrouve la TSP-1 (Asch *et al.*, 1987), les acides gras à longue chaîne (AGLC) (Abumrad

et al., 1993), les LDLox (Endemann *et al.,* 1993), les lipoprotéines de haute densité (HDL), LDL et lipoprotéines de très faible densité (VLDL) natifs (Calvo *et al.,* 1998), le collagène I et IV (Tandon *et al.*, 1989a), les segments externes des photorécepteurs (Ryeom *et al.*, 1996), les phospholipides anioniques (Rigotti *et al.,* 1995), les cellules apoptotiques (Ren *et al.,* 1995) et les érythrocytes infectés avec le *Plasmodium falciparum* (Oquendo *et al.,* 1989).

1.4.2. Ligands exogènes: les sécrétines de l'hormone de croissance comme ligands exogènes du CD36

Les sécrétines de l'hormone de croissance (HC) ou GHRPs ont été synthétisées en 1977 alors que Bowers et al., développaient une série de petits peptides synthétiques dérivés de la méthionine-enképhaline à la recherche de composés ayant la propriété de stimuler la sécrétion de l'HC. Bowers et al. ont montré *in vitro* que la D-Trp2-Met-enképhaline stimule faiblement la sécrétion de l'HC par les cellules hypophysaires, tout en étant dépourvue d'activité opiacée (Bowers *et al.*, 1984). Des études ultérieures ont permis la synthèse d'hexapeptides tel que le GHRP-6 (His-D-Trp-Ala-Trp-D-Phe-Lys-NH2), un puissant sécrétagogue de l'HC tant chez les modèles expérimentaux que chez l'homme. Le GHRP-6 a servi de prototype pour la synthèse d'autres analogues variant de trois à sept résidus, dont l'heptapeptide GHRP-1 (Ala-His-D-beta-NaI-Ala-Trp-D-Phe-Lys-NH2) et l'hexapeptide GHRP-2 (D-Ala-D-beta-NaI-Ala-Trp-D-Phe-Lys-NH2). Les effets de ces peptides synthétiques sont médiés par le récepteur de la ghréline, un récepteur à sept passages transmembranaires couplé à une protéine G, appelé récepteur des sécrétagogues de l'HC de type 1a (GHS-R1a) (Bowers, 1998).

En 1994, Deghenghi et al. ont synthétisé un analogue métaboliquement plus stable des GHRPs en substituant, dans la structure du GHRP-6 (*Figure 3*), la Trp par le 2-methyl-Trp, donnant ainsi naissance à l'hexaréline (His-D-2-methyl-Trp-Ala-Trp-D-Phe-Lys-NH2). L'hexaréline est un peptide synthétique qui lie, comme ces prédécesseurs, le récepteur GHS-R1a, mais qui a aussi la capacité de lier le récepteur CD36 (Bodart *et al.,* 1999; Bodart *et al.,* 2002). Les études initiales chez un modèle de coeur isolé ont montré que l'hexaréline induit une vasoconstriction coronaire de façon dépendante de la dose et que cet effet n'est pas observé chez les souris déficientes en CD36 (Bodart *et al.,* 2002). De plus, des études ont montré que l'hexaréline se lie à un domaine de liaison qui chevauche le domaine de liaison des LDLox sur le récepteur CD36 (Demers *et al.*, 2004) et que l'hexaréline induit des effets anti-athérosclérotiques (Avallone *et al.,* 2005). Le EP 80317 (Haic-D-2-methyl-Trp-D-Lys-Trp-D-Phe-Lys-NH2), un analogue de l'hexaréline, est un ligand du récepteur CD36 qui, contrairement à l'hexaréline, ne se lie pas au GHRS-1a (Demers *et al.,* 2004).

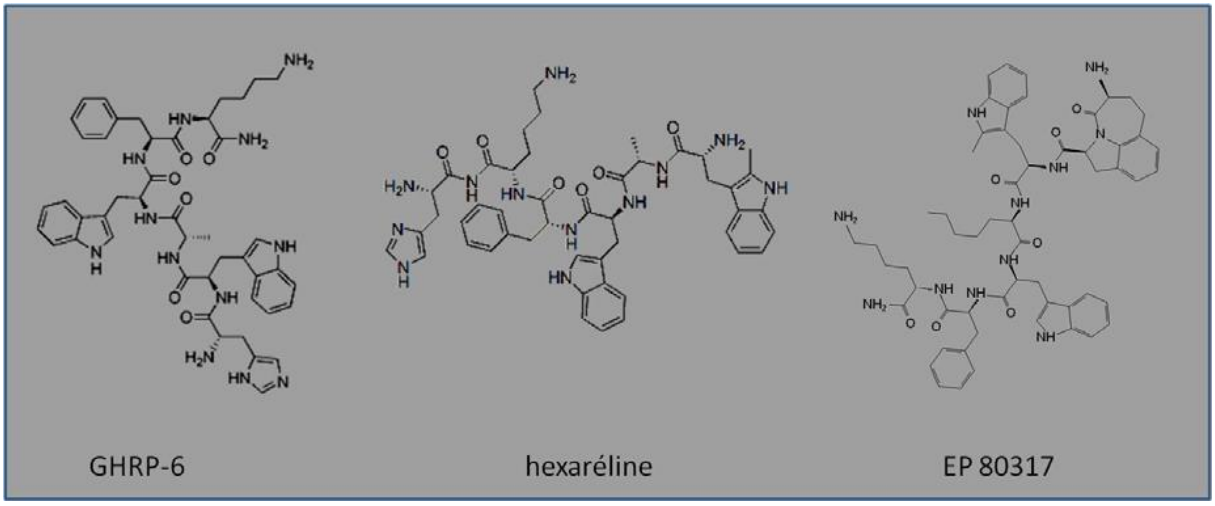


Figure 3. Structure des sécrétines de l'hormone de croissance

1.5. Rôle du récepteur CD36 dans le métabolisme lipidique

1.5.1. Le CD36: un récepteur-clé dans le transport des acides gras

Le transport des acides gras peut se faire par diffusion passive à travers la membrane plasmique, cependant il y plusieurs évidences qui montrent que les acides gras traversent la membrane par transport facilité grâce à des protéines membranaires comme le FAT/CD36, la protéine de liaison des acides gras (FABP) et la protéine de transfert des acides gras (FATP) (Ehehalt *et al.*, 2006; Su & Abumrad, 2009). Le transporteur FAT/CD36, l'homologue murin du CD36 humain, joue un rôle important dans le transport et le métabolisme des acides gras à longues chaînes dans les adipocytes et les muscles cardiaque et squelettique. Le mécanisme précis du transport des acides gras à travers la membrane plasmique reste à déterminer; cependant différents modèles, d

ans lesquels le FAT/CD36 joue un rôle-clé, ont été proposés: 1) le FAT/CD36, seul ou de concert avec la FABP membranaire (FABPpm), pourrait concentrer les acides gras à la membrane plasmique favorisant leur diffusion passive ou 2) transporter directement les acides gras à travers la membrane dans le cytosol, où les acides gras vont se lier à une FABP cytosolique (FABPc); 3) le FAT/CD36 pourrait servir de récepteur pour les acides gras et ainsi les acheminer à proximité des transporteurs FATP1 et FATP6 pour être internalisés (Schwenk *et al.*, 2010; Schwenk *et al.*, 2008) (*Figure 4*). Dans ces tissus riches en FAT/CD36, le récepteur est emmagasiné dans des vésicules intracellulaires formées à partir des endosomes de recyclage, et est transloqué à la membrane plasmique en réponse à l'insuline ou à une

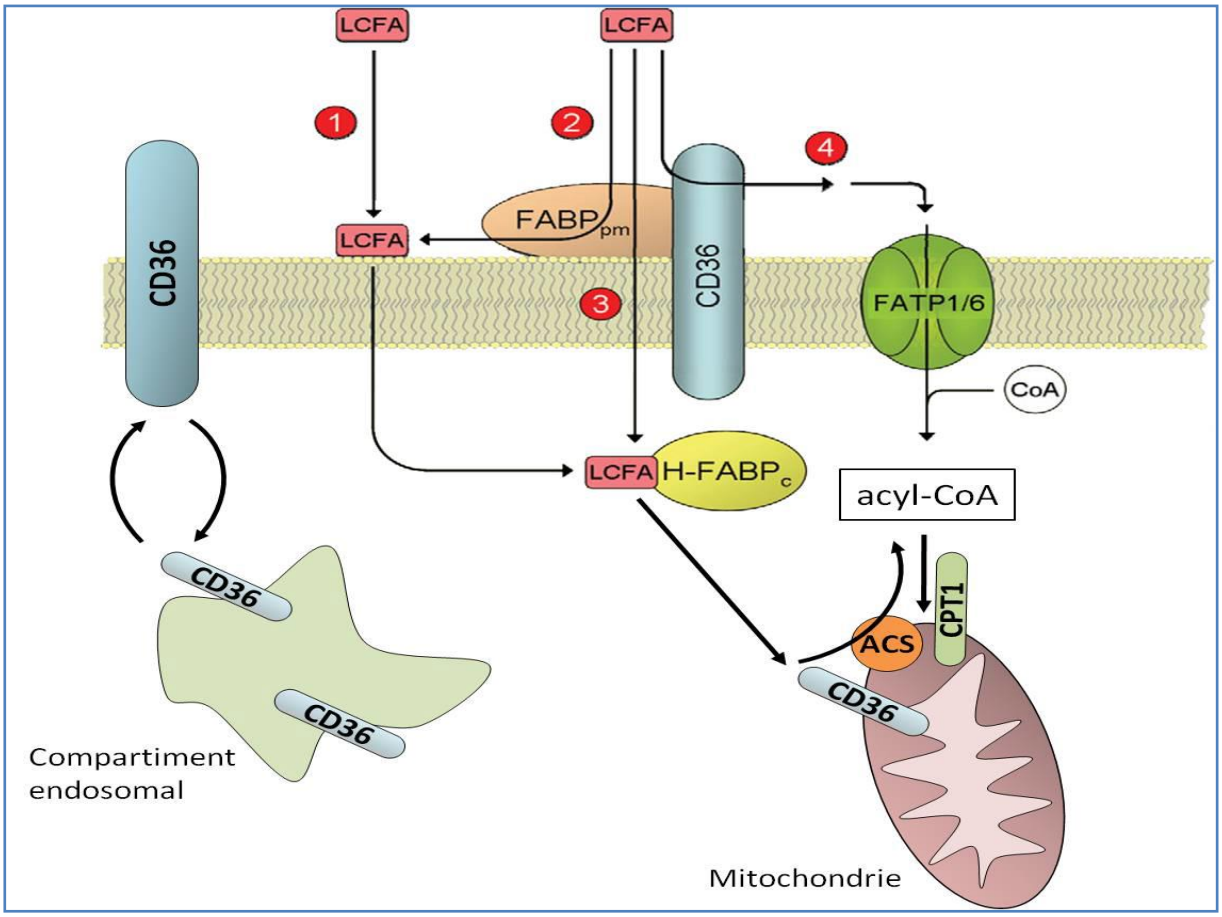


Figure 4. Transport des acides gras par le récepteur CD36
Comme le mécanisme exact du transport transmembranaire des acides gras à longues chaînes (AGLC ou LCFA, en anglais) est encore inconnu, différents modèles ont été proposés. (1) Compte tenu de leur caractère hydrophobe, les LCFA pourraient traverser la membrane cytoplasmique par diffusion passive. Cependant, l'internalisation des LCFA dépend en grande partie du récepteur CD36, mais il n'est pas spécifié si (2) le CD36, seul ou conjointement avec la FABPpm, concentre les LCFA à la surface cellulaire pour favoriser la diffusion passive, ou si (3) le CD36 transporte activement les LCFA à travers la membrane plasmique. (4) Le CD36 pourrait également servir de récepteur pour les LCFA et les acheminer à proximité des transporteurs FATP1 et FATP6 ce qui pourrait faciliter le transport des LCFA. Le CD36 est aussi exprimé au niveau de la mitochondrie où il joue un rôle dans l'oxydation des acides gras, de concert avec la carnitine palmitoyltransférase (CPT-1) (Figure modifiée de (Schwenk *et al.*, 2008).

contraction musculaire, où il intervient alors dans l'internalisation des acides gras (Bonen *et al.*, 2007; Schwenk *et al.*, 2008). Campbell et al. ont montré que le FAT/CD36 est aussi exprimé au niveau des mitochondries et joue un rôle dans l'oxydation des acides gras, de concert avec la carnitine palmitoyltransférase I (CPT I) (Campbell *et al.*, 2004). L'apparition des modèles de souris qui surexpriment ou qui sont déficientes en FAT/CD36 a permis de confirmer le rôle crucial de ce transporteur dans le transport des acides gras.

Une surexpression du FAT/CD36 dans le muscle entraîne une augmentation de l'oxydation des acides gras après une contraction et diminue les niveaux de lipides plasmatiques chez la souris (Ibrahimi *et al.*, 1999). À l'opposé, une déficience en FAT/CD36 diminue

significativement le captage et le métabolisme des acides gras dans les muscles cardiaque et squelettique ainsi que dans les adipocytes et engendre une élévation des taux d'acides gras libres non estérifiés (AGNE) et des triglycérides plasmatiques, ainsi qu'une résistance à l'insuline au niveau hépatique (Hajri & Abumrad, 2002; Goudriaan *et al.,* 2003). Le CD36 est aussi impliqué dans la perception sensorielle. Chez les rongeurs, le CD36 est exprimé à la membrane apicale des cellules sensorielles situées dans les papilles linguales entraînant une préférence pour les aliments riches en lipides. La prise de nourriture riche en acides gras, comme l'acide linoléique, transmet un signal aux fibres nerveuses conduisant à la perception du goût et à la sécrétion d'acides biliaires, préparant ainsi le système digestif pour l'absorption des acides gras.

De plus, le FAT/CD36 est fortement exprimé dans le petit intestin (Nassir *et al.,* 2007; Chen *et al.,* 2001a). Le FAT/CD36 est principalement localisé au niveau de la membrane apicale des entérocytes, sur les 2/3 supérieurs des villosités. L'expression du FAT/CD36 et sa localisation au niveau intestinal suggère un rôle dans l'absorption des lipides mais cette fonction a longtemps été controversée. Chez les souris déficientes en FAT/CD36, il n'y a pas de modification significative de l'absorption intestinale des acides gras. Cependant, une déficience en FAT/CD36 diminue l'absorption intestinale des acides gras à très longues chaînes (AGTLC) chez les souris sous diète à haute teneur en gras (Drover *et al.,* 2008). De plus, l'ablation du FAT/CD36 engendre une diminution de la sécrétion des chylomicrons au niveau de l'intestin proximal et la production de chylomicrons de plus petites tailles avec moins d'apoB-48, en plus d'entraîner une diminution de la clairance des lipoprotéines riches en triglycérides due à l'inhibition de la lipase endothéliale (Drover *et al.*, 2005; Goudriaan *et al.*, 2005; Nauli *et al.,* 2006). Ces données indiquent donc que le FAT/CD36 joue un rôle dans la formation et la sécrétion des chylomicrons mais ne montrent pas que le FAT/CD36 peut contribuer à l'absorption des acides gras au niveau intestinal. Récemment, Nassir et al. ont montré que le FAT/CD36 est impliqué dans le captage du cholestérol et des acides gras dans l'intestin proximal (voir section 2.2.3.2.1.). Ils ont observé une perturbation de la sécrétion des lipoprotéines par les entérocytes ainsi qu'une diminution de la production des apoB48 et apoA-IV. De plus, une augmentation de la protéine *Niemann Pick C1 like 1* (NPC1L1) est observée chez les souris déficientes en FAT/CD36 suggérant une coopération entre les deux protéines dans l'absorption du cholestérol (Nassir *et al.,* 2007).

1.5.2. Le CD36 : un récepteur-clé dans l'internalisation des lipoprotéines oxydées et le développement de l'athérosclérose

Le CD36 exprimé à la surface des macrophages joue un rôle dans l'internalisation des lipoprotéines modifiées et dans le développement des lésions athérosclérotiques. Des études

ont montré que la délétion du gène CD36 chez les souris déficientes en apoE nourries avec une diète enrichie en cholestérol et en lipides entraînent une réduction de 80% des lésions athérosclérotiques associée à une diminution de l'internalisation des LDLox dans les macrophages (Febbraio *et al.*, 1999; Febbraio *et al.*, 2000). Pour bien comprendre les effets anti-athérosclérotiques des ligands du récepteur CD36, il est essentiel de bien comprendre les différentes étapes du développement de l'athérosclérose.

Le CD36 exprimé à la surface des macrophages joue un rôle important dans l'internalisation des LDLox. C'est la fraction lipidique des LDLox qui est à l'origine de leur interaction avec le récepteur CD36, contrairement à leur interaction avec les autres récepteurs *scavengers* qui se fait plutôt via la fraction protéique (Nicholson *et al.*, 1995). On retrouve sur le récepteur CD36 un domaine composé d'un groupement de trois lysines chargées positivement, lysine 164-166 (Kar *et al.*, 2008), qui permet l'interaction avec les LDLox qui sont chargées négativement (Doi *et al.*, 1994). Podrez et al. ont identifié une famille de phospholipides oxydés formés lors de l'oxydation des particules de LDL ou des phospholipides cellulaires, particulièrement la phosphatidylcholine (PC), contenant des acides gras polyinsaturés (PCoxCD36). Ces phospholipides oxydés, contenant un acide gras avec un groupement carbonyl insaturé présentant un groupe γ-hydroxy(ou oxo)-α,β, estérifié en position sn-2, sont responsables de la liaison de haute affinité avec le récepteur CD36 (*Figure 5*) (Podrez *et al.*, 2002a; Podrez *et al.*, 2002b). En plus du PCoxCD36, 36

d'autres phospholipides oxydés ont récemment été identifiés comme ligands potentiels des récepteurs scavengers de type B (*Figure 5*) (Gao *et al.*, 2010).

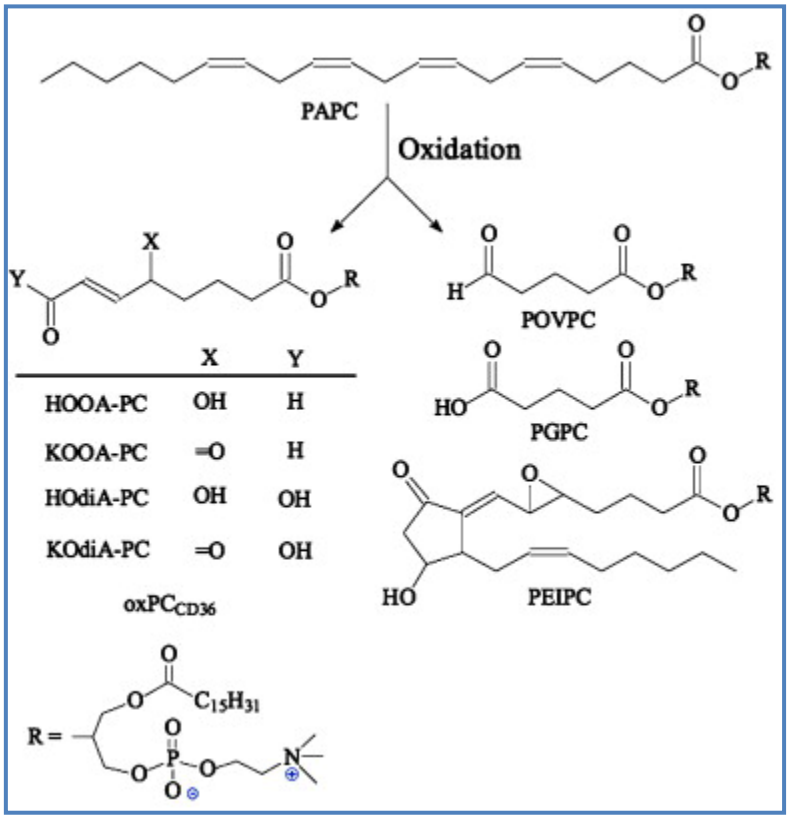


Figure 5. Schéma des différents phospholipides oxydés formés lors de l'oxydation du 1-palmitoyl-2-arachidonoyl-sn-glycero-3-phosphorylcholine (PAPC) (Figure tirée de (Ashraf *et al.*, 2009)).

L'activation du récepteur CD36 peut engendrer d'une part, l'internalisation et la dégradation des LDLox qui entraînent la libération d'oxystérols et de dérivés d'acides gras oxydés comme 9 et 13-HODE, des ligands des récepteurs nucléaires LXRα et PPARγ, respectivement (Brown *et al.*, 1996; Brown *et al.*, 2000; Janowski *et al.*, 1999; Nagy *et al.*, 1998). L'activation du PPARγ par les 9- et 13-HODE s'accompagne d'une augmentation de la transcription du récepteur CD36, régulant ainsi à la hausse l'internalisation des LDLox (Tontonoz *et al.*, 1998). Cependant, l'activation de ce récepteur nucléaire entraîne aussi la régulation des gènes impliqués dans le métabolisme lipidique (Tontonoz *et al.*, 1998; Nagy *et al.*, 1998; Moore *et al.*, 2001; Chinetti *et al.*, 2001; Li *et al.*, 2004). La signalisation intracellulaire résultant de l'activation du PPARγ par le LDLox via le récepteur CD36 peut induire l'activation d'une PKC conventionnelle (Feng *et al.*, 2000). De plus, le CD36, en dépit de ses courts domaines intracytoplasmiques, engendre une cascade de signalisation lorsque le récepteur CD36 est activé par ses divers ligands. Cette cascade implique le recrutement, au domaine C-terminal intracytoplasmique du CD36, de tyrosine kinase de la famille des src (fyn, lyn et yes) (Moore *et al.*, 2002; Medeiros *et al.*, 2004; Jimenez *et al.*, 2000; Janabi *et al.*, 2000; Bamberger *et al.*, 2003). Leur activation conduit à l'activation de MAPK spécifiques, c-Jun N-terminal kinase (JNK) 1 et 2 (Rahaman *et al.*, 2006). Cette voie de signalisation est impliquée dans la formation de cellules spumeuses puisque l'inhibition pharmacologique de JNK conduit à une réduction de l'athérosclérose, tout comme l'athéroprotection observée chez les souris déficientes en JNK2 et en src kinase lyn (Sumara *et al.*, 2005; Ricci *et al.*, 2004; Miki *et al.*, 2001). De plus, des études plus récentes ont montré que la liaison du CD36 avec les LDLox peut entraîner l'activation de la kinase p44/42 et de COX-2, et ainsi augmenter la production de 15d-PGJ2 menant à l'activation du PPARγ (Taketa *et al.*, 2008).

Bien que ces mécanismes montrent que le récepteur CD36 joue un rôle-clé dans la formation des cellules spumeuses, l'internalisation des LDLox par les macrophages est aussi associée à la sécrétion de facteurs pro-inflammatoires (section 1.5).

1.6. Rôle du CD36 dans la régulation de l'angiogenèse

Tel que mentionné précédemment, le CD36 est un récepteur pour la TSP-1 ainsi que pour les molécules qui présentent un motif répétitif de type 1 (TSRs) et leur domaine de liaison 43 sur le CD36 est nommé le domaine CLESH-1 (Armstrong & Bornstein, 2003; Crombie & Silverstein, 1998). La TSP-1 est un inhibiteur endogène de l'angiogenèse via l'activation du CD36 (Cf. *Figure 6*). Outre l'inflammation, le CD36 joue ainsi un rôle potentiellement important dans la

croissance tumorale et toutes autres pathologies dans lesquelles on retrouve un processus de néovascularisation (Silverstein & Febbraio, 2009). Le CD36 agit sur l'angiogenèse tumorale en mobilisant les facteurs pro-angiogéniques responsables de la prolifération, la migration et la survie des cellules endothéliales, et en générant des facteurs anti-angiogéniques (Dawson *et al.*, 1997; Jimenez *et al.*, 2000). La TSP-1 inhibe le développement des vaisseaux sanguins mais pas celui des vaisseaux lymphatiques, en raison de l'absence du CD36 au niveau de l'endothélium lymphatique. L'inhibition de l'angiogenèse par le complexe TSP-1/CD36 est initiée par le recrutement de fyn, une protéine tyrosine kinase de la famille des *src*, qui d'une part active une cascade de signalisation impliquant les protéine kinases p38 et JNK ainsi que le ligand Fas et la caspase-3 (Jimenez *et al.*, 2000; Volpert *et al.*, 2002) et d'autre part inhibe la phosphorylation d'akt induite par le VEGF (Sun *et al.*, 2009). De plus, une fois liée au CD36, la TSP-1 active le TGFβ latent qui inhibe l'activation des métalloprotéinases (en particulier la MMP-9) et entraîne la séquestration du facteur de croissance endothéliale vasculaire (VEGF) dans la matrice extracellulaire résultant en une inhibition de la migration cellulaire tout en induisant l'apoptose (Yehualaeshet *et al.*, 1999; Jimenez *et al.*, 2000; Rodriguez-Manzaneque *et al.*, 2001; Gabison *et al.*, 2003). Ces éléments, en plus du CD36, sont essentiels pour inhiber l'angiogenèse et induire l'apoptose des cellules endothéliales activées. La production de certains facteurs pro-apoptotiques comme le TNF-44 α et le ligand Fas seraient aussi impliqués dans le processus (Volpert *et al.*, 2002; Rege *et al.*, 2009).

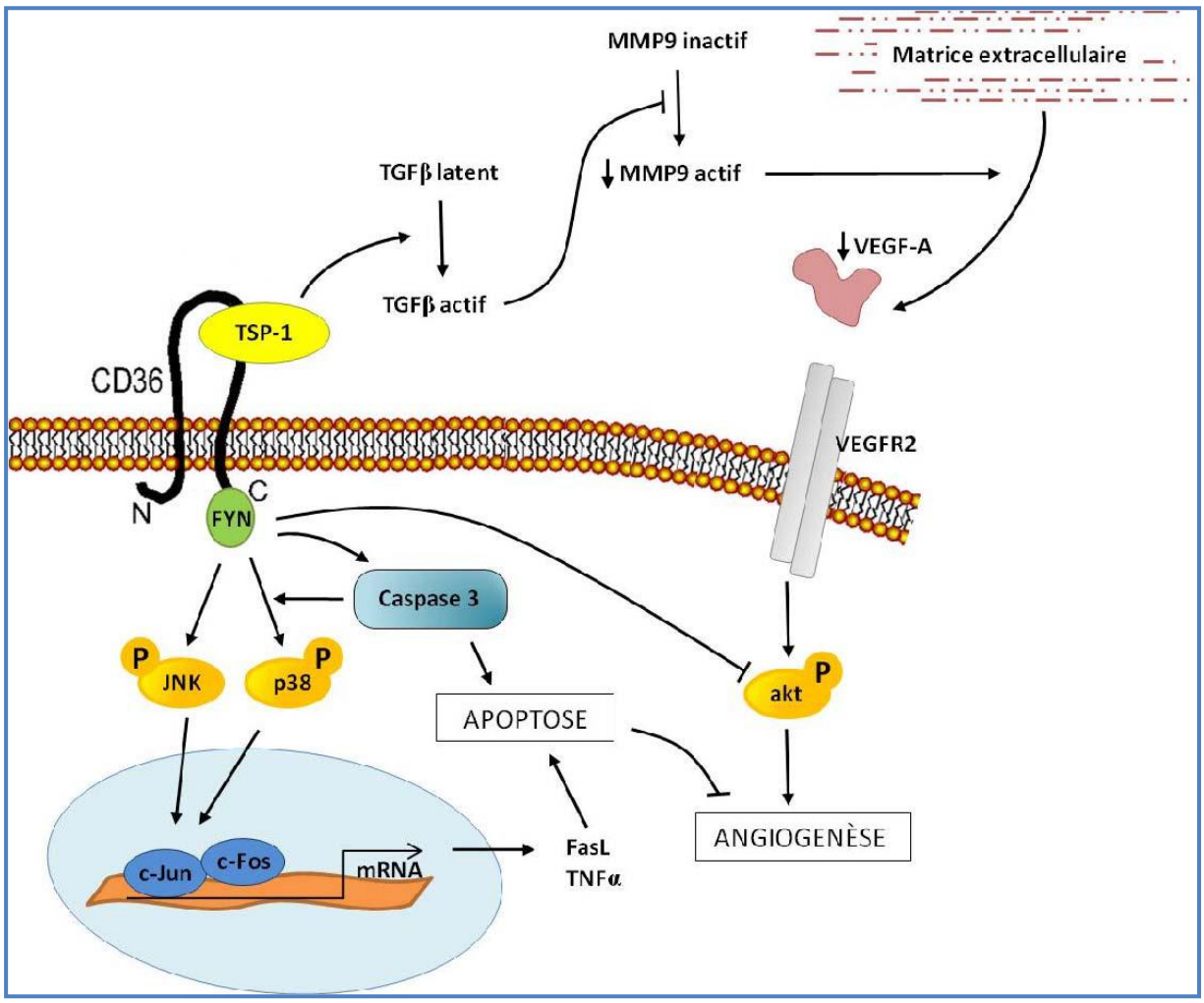


Figure 6. Rôle du CD36 dans la régulation de l'angiogenèse

La liaison de la TSP-1 avec le récepteur CD36 à la surface des cellules endothéliales entraîne une cascade de signalisation qui passe par fyn, la caspase-3 et les MAPK p38 et JNK et engendre une augmentation de l'expression des gènes impliqués dans l'apoptose. La TSP-1 peut aussi inhiber l'angiogenèse de façon indirecte puisqu'elle a la capacité d'activer le TGFβ et ainsi d'inhiber la MMP-9. L'inhibition de l'activation de la MMP9 entraîne une diminution de la libération du VEGF-A qui est séquestré dans la matrice extracellulaire, ce qui entraîne une diminution de la quantité de VEGF-A qui se lie au récepteur VEGFR2 et donc engendre un effet anti-angiogénique. De plus, l'activation du CD36 et de la kinase fyn par la TSP-1 semble inhiber la phosphorylation d'akt induite par le VEGF-A (Bujold K., 2011).

Récemment, un parallèle a été fait entre l'athérosclérose et la dégénérescence maculaire liée à l'âge. La dégénérescence maculaire est caractérisée, d'une part, par une néovascularisation cornéenne (Mwaikambo *et al.*, 2008b; Mwaikambo *et al.*, 2008a) et d'autre part, par l'accumulation de lipoprotéines contenant l'apoB100 (LDL et VLDL) dans la membrane de Bruch, suivie par l'apoptose des cellules de l'épithélium pigmentaire rétinien (RPE) (Dunaief *et al.*, 2002; Li *et al.*, 2005; Malek *et al.*, 2003). Le VEGF joue un rôle-clé dans la

néovascularisation cornéenne et un lien entre le CD36 et le VEGF est appuyé par l'observation qu'une délétion du récepteur CD36 induit une augmentation de l'expression de VEGF-A, JNK-1 et c-Jun ainsi qu'une augmentation de la néovascularisation. Il en résulte la formation d'un voile au niveau de la cornée qui est amplifiée chez les souris plus âgées (Mwaikambo *et al.*, 2008b; Mwaikambo *et al.*, 2008a). Au contraire, l'activation du récepteur CD36 entraîne plutôt une diminution de l'expression du VEGF-A et une diminution de la formation de nouveaux vaisseaux, voir même la régression de la néovascularisation (Mwaikambo *et al.*, 2008b; Mwaikambo *et al.*, 2008a).

1.7. La maladie cardiomyopathie

La cardiomyopathie hypertrophique est caractérisée par des altérations structurales telles que l'hypertrophie, la nécrose et la fibrose myocardique qui s'accentuent avec l'évolution de la maladie. En effet, un certain nombre d'études chez l'homme comme chez l'animal ont montré des changements structuraux en parallèle de changements fonctionnels chez le cœur cardiomyopathique tel que la fibrose et l'hypertrophie myocardiques.

L'hypertrophie ventriculaire constitue une des caractéristiques essentielles de la cardiomyopathie. Cette hypertrophie ventriculaire est retrouvée dans les études épidémiologiques de Rubler et coll. (1972) ou encore de Hamby et coll. (1974) chez des patients diabétiques présentant des complications sévères et une insuffisance cardiaque. En effet, ces patients présentent une hypertrophie myocardique associée à une augmentation du poids du cœur. Les études expérimentales réalisées chez l'animal aboutissent à la même conclusion que les études épidémiologiques. En effet, le ratio poids du cœur / poids du corps, calculé comme index de l'hypertrophie cardiaque, est significativement augmenté dans de nombreuses études chez des rats diabétiques (Golfman et coll. 1999, Bidasee et coll. 2001, Kim et coll. 2001, Netticadan et coll. 2001). L'augmentation du ratio poids du cœur / poids du corps résulte d'une hypertrophie cellulaire à laquelle s'associent des processus complexes de remodelage, telle que la fibrose (Swynghedauw 1999). L'utilisation du ratio poids du ventricule gauche (VG) / poids du coeur par Golfman et coll. (1996) suggère plus précisément une hypertrophie ventriculaire. En effet, la masse, la taille et l'épaisseur du ventricule gauche vont progressivement augmenter à mesure que la cardiomyopathie diabétique devient de plus en plus sévère (Fang et coll. 2004). Plus récemment, le recours à l'imagerie par résonnance magnétique a permis à l'équipe de Loganathan et coll. (2006) de mettre en évidence une augmentation du ratio volume du VG / poids du corps, suggérant une hypertrophie du VG chez des rats 8 semaines après l'induction du diabète. L'ensemble de ces données épidémiologiques, cliniques et expérimentales confirment l'existence d'une mort cellulaire, d'une fibrose myocardique et d'une hypertrophie ventriculaire, caractéristiques de la

cardiomyopathie. Il est important de préciser que ces changements structuraux sont à l'origine de la diminution de la compliance myocardique observée au cours du diabète. De plus, il apparaît que la perte de fonction systolique dépendrait du degré de nécrose/apoptose des myocytes cardiaques tandis que la perte de fonction diastolique serait directement liée au degré de fibrose (Fang et coll. 2004). Cependant, le développement de la cardiomyopathie diabétique est aussi fréquemment associé à des désordres métaboliques.

1.8. Les fonctions de CD36 qui peut impact la maladie cardiomyopathie

Le cœur est l'organe le plus consommateur d'énergie. Le cœur humain sain est une pompe très efficace qui propulse ~5 L/min de sang, totalisant >7000 L/jour et >2600000 L/année (Taegtmeyer, 2004) (Shen et al 2010). Étant donné que dans les conditions normales, ses réserves cardiaques sont assez faibles (5 μmol/g de poids humide) (Opie, 1998) et que le turnover du pool d'ATP myocardique est d'environ 10 secondes, le cœur doit ainsi produire une quantité suffisante d'ATP pour maintenir sa contractilité et s'adapter rapidement aux changements physiologiques et à la disponibilité des différents substrats. Le renouvèlement structurel extracellulaire et intracellulaire se produit rapidement à partir d'un approvisionnement régulier en acides aminés, des lipides et carbohydrates. D'un point de vue métabolique, le myocarde est alors considéré comme un tissu «omnivore», car il peut oxyder simultanément les AG et le glucose et dans des proportions variables (Randle, 1998). D'un point de vue énergétique et comparativement au glucose, les AG sont des substrats plus efficaces, puisque, pour une même quantité de substrats oxydés, les AG produisent une plus grande quantité d'ATP. Par contre, si l'on considère la dépense en oxygène pour leur oxydation, le glucose est un substrat plus efficace que les AG qui nécessitent 10% plus d'oxygène pour être oxydés (Lopaschuk et al., 2003). Sur 80% de la consommation en O2 couplée à la synthèse de l'ATP, environ 30% sont utilisés pour la synthèse protéique, 25% par la Na+ -K+ -ATPase, environ 8% par les Ca2+ -ATP-ases et 10% pour la gluconéogenèse. Le produit final; l'acétyl-CoA (molécule de 2 carbones) de l'oxydation des différents substrats, via différentes voies métaboliques, va ensuite entrer dans le cycle de Krebs (CK) pour y être oxydé. Ce sont les équivalents réduits produits par le CK, la β-oxydation (β-OX) des AG, la glycolyse et l'oxydation du pyruvate et du lactate (le nicotinamide adénine dinucléotide (NADH) et la flavine adénine dinucléotide (FADH2) qui vont par la suite transporter des électrons à la chaîne respiratoire (CR) où est produit l'ATP à partir de l'adénosine diphosphate (ADP) par phosphorylation oxydative (Cf. figure 7).

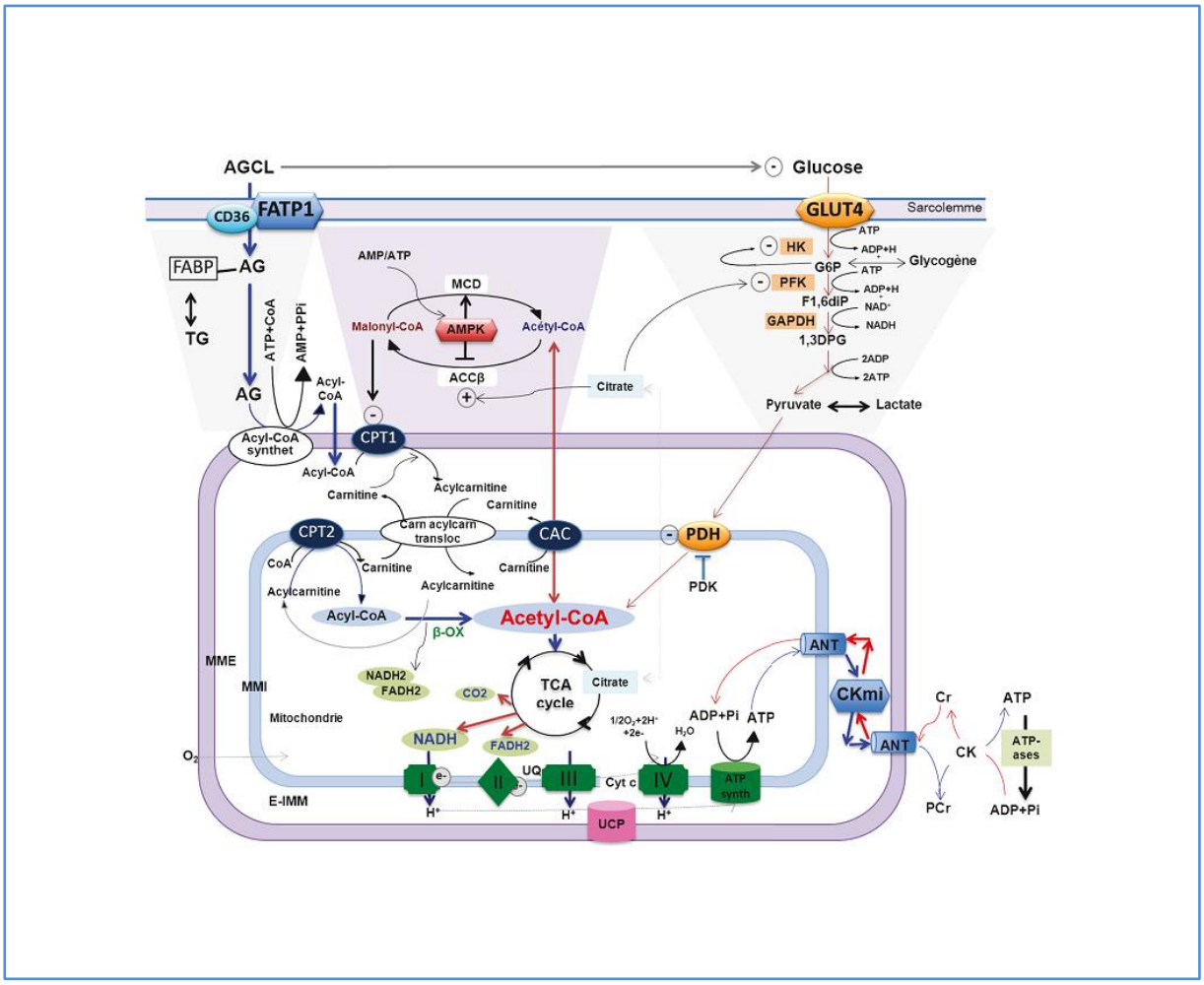


Figure7 : Métabolisme des acides gras et le glucose au niveau du tissu cardiaque.

Le métabolisme des AG est sous un contrôle plus compliqué que celui du glucose et dépend d'un certain nombre de facteurs comme 1) l'apport d'AG; 2) la présence d'autres substrats énergétiques; 3) la demande énergétique; 4) l'approvisionnement en oxygène; 5) le contrôle allostérique de la captation, l'estérification, et le transport mitochondrial d'AG; et 6) le contrôle de la fonction mitochondriale, y compris le contrôle direct de -oxydation, l'activité du cycle de Krebs et de la chaîne de transport d'électrons (CTE). Le contrôle transcriptionnel des enzymes impliquées dans le métabolisme des AG et de la biogenèse mitochondriale sont aussi des déterminants importants du taux de -oxydation. La figure 8 illustre l'essentiel du métabolisme lipidique cardiaque et les différentes étapes décrites ci-dessous sont tirées de plusieurs articles de revue (Glatz et al.,2010) , (Lopaschuk et al., 2010) , (Eaton, 2002) , (Wanders et al., 1999)

.

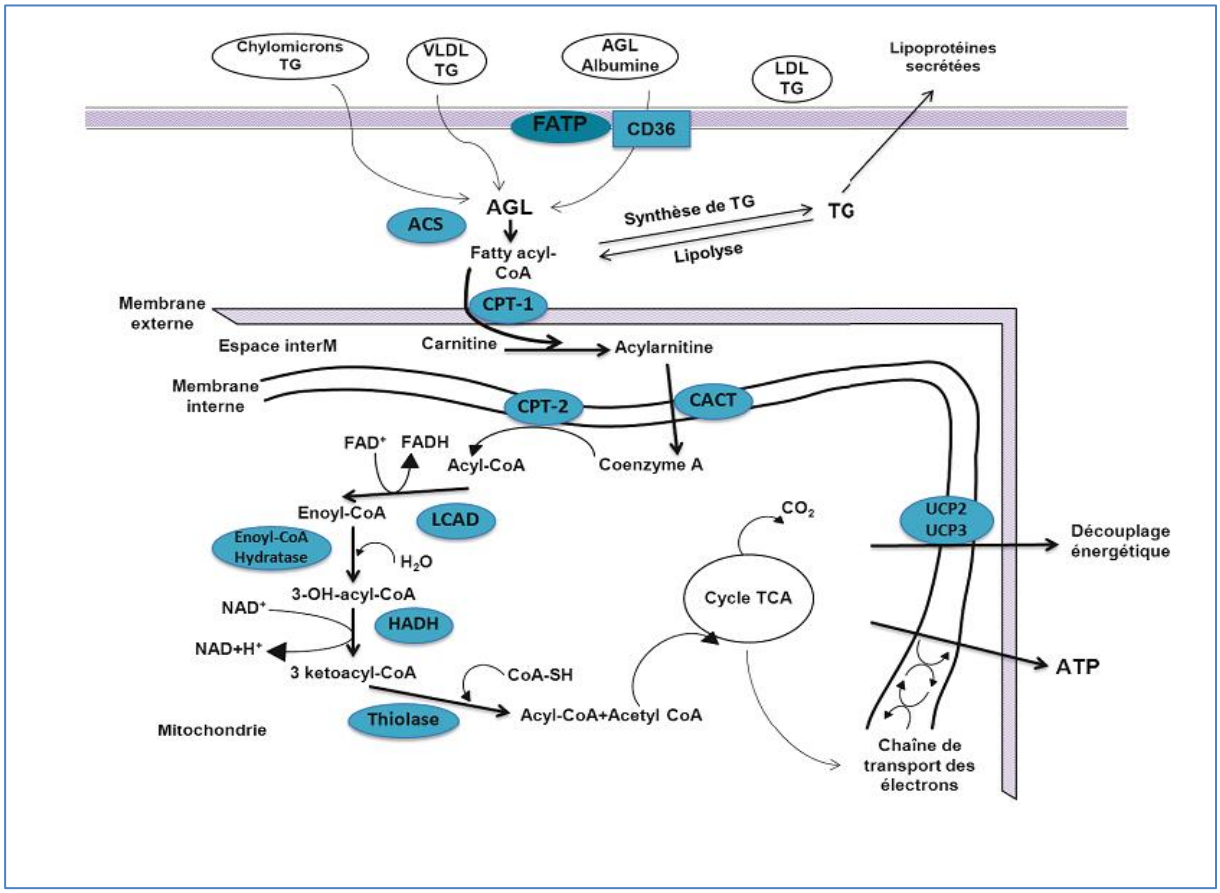


Figure 8 : Le contrôle transcriptionnel des enzymes impliquées dans le métabolisme des AG et de la biogenèse mitochondriale sont aussi des déterminants importants du taux de -oxydation

Les AGL peuvent entrer dans le cardiomyocyte soit par diffusion passive par un mécanisme de flip-flop à travers la bicouche phospholipidique (Schaffer, 2002), ou facilité par une protéine de transport membranaire (Luiken et al., 1999) (Cf. Figure 8). Jusqu'à ce jour, trois protéines impliquées dans la captation des AGCL ont été identifiées: 1) l'isoforme membranaire de la protéine de liaison des AG (fatty acid binding protein ; FABPpm), 2) les protéines de transport des AG (fatty acid transport proteins ; FATP) et 3) les translocases qui, dans le cœur, sont connues sous la forme du CD36 (fatty acid translocase/cluster of differentiation ; FAT/CD36) Parmi les trois transporteurs d'AG identifiés, FAT/CD36 est l'un des plus étudié qui joue un rôle essentiel dans le transport des AG à travers la membrane plasmique. La distribution tissulaire de FAT/CD36 montre qu'il est fortement exprimé dans les tissus ayant une capacité métabolique pour les AG élevée, comme le cœur

(Glatz et al., 2010). Des études utilisant des souris déficientes en FAT/CD36 ont montré que cette translocase est responsable de l'absorption de 50 à 80% des AG par le

cœur. Des travaux antérieurs ont trouvé que l'inhibition de l'absorption d'AG médiée par CD36 par délétion génétique conduit à une réduction significative de 40 à 60% dans l'oxydation des AG dans le cœur, ainsi qu'une diminution des niveaux d'estérification de TG, ce qui suggère que le degré d'absorption des AG par FAT/CD36 a un impact considérable (Kuang et al., 2004) , (Brinkmann et al., 2002).

FAT/CD36 est une glycoprotéine membranaire de 88 kDa. Sa fonction dans l'absorption des AGCL est largement régulée par sa translocation intracellulaire et le contrôle transcriptionnel. Dans le cœur environ 50% des FAT/CD36 sont stockés dans des compartiments intracellulaires, et doivent être acheminés vers la membrane plasmique afin de participer activement au transport des AG (Luiken et al., 2002). Il a été démontré que des stimuli physiologiques, y compris la contraction, l'exercice et l'insuline, peuvent induire sa translocation vers la membrane pour accélérer le transport et l'oxydation des AG (Luiken et al., 2002), (Luiken et al., 2003). Les mécanismes de la contraction et de la translocation de FAT/CD36 induite par l'insuline peut impliquer l'activation des voies de PDK/Akt et de l'AMPK, respectivement (Glatz et al., 2010). Comme pour d'autres protéines impliquées dans le métabolisme des lipides, l'expression de FAT/CD36 est sous le contrôle transcriptionnel des PPARs (Madrazo et Kelly, 2008).

2. Etat des lieux des puces à ADN et de leurs méthodes d'analyse

Pour comprendre l'explosion des méthodes statistiques d'analyse des puces et les étudier, il faut savoir : Quel a été l'apport des puces dans la compréhension des mécanismes biologiques. A quoi correspondent biologiquement et techniquement les puces. Comment les données sont traitées et analysées.

2.1. Apport des puces à ADN pour mesurer le niveau d'expression des gènes

Les puces à ADN permettent de visualiser simultanément le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier. 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. Le concept de puce à ADN date du début des années 1990. Toutefois, le principe fondateur remonte à 1975. En effet, la technologie des puces à ADN se base sur la technique d'hybridation entre des séquences

complémentaires d'ADN, conformément aux observations de E. Southern en 1975. De ces observations sont nées les techniques de Southernet Northern blot qui sont à l'origine des premières puces à ADN (Lander, 1999).

Les puces à ADN ont d'abord été conçues sur de grandes membranes poreuses en nylon ou macroarrays (Gresset al., 1992; Nguyenet al., 1995). La miniaturisation, rendue possible par les progrès de la robotique, a ensuite permis le développement des microarrays. Comme leur nom l'indique, ces puces à ADN sont de plus petites surfaces telles une lame de microscope (Schenaet al., 1995) ou une petite membrane nylon (Jordan, 1998). Elles présentent également l'avantage de pouvoir être de très haute densité et par conséquent sont susceptibles de recouvrir l'intégralité du génome d'un organisme.

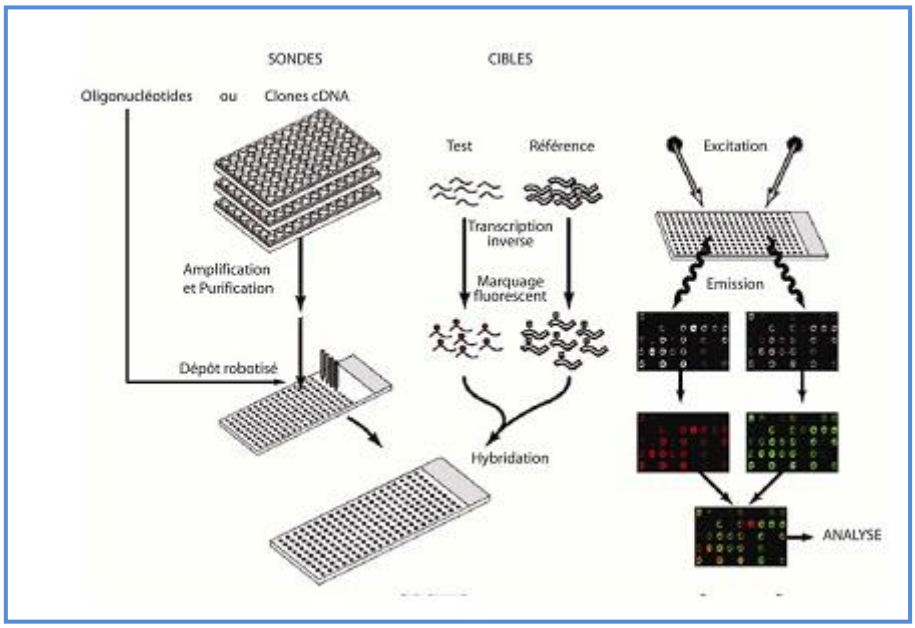


Figure 9. Schématisation de la technique d'analyse du transcriptome par la technologie des puces à DNA (d'après Duggan et al., 1999).

Les sondes (oligonucléotides ou clones d'ADNc purifiés et amplifiés) sont déposées mécaniquement sur une lame de verre. Parallèlement, les cibles sont couplées à des marqueurs fluorescents (parfois amplifiés) par transcription inverse. Par exemple, la cible test est marquée par une Cyanine 5 (Cy5) rouge et la cible de référence par une Cyanine 3 (Cy3) verte. Les cibles sont assemblées pour former un mélange complexe. Ce mélange pourra s'hybrider, dans des conditions de stringence particulières, avec les sondes présentes sur la puce. La lecture est réalisée par un scanner muni d'un microscope confocal, couplé à deux lasers. Ces lasers possèdent des longueurs d 'ondes d'excitation spécifiques, correspondant à celles des deux marqueurs fluorescents. L'excitation et l'émission (amplifiée par des photomultiplicateurs) des fluorochromes permettent l'obtention de deux images (une pour

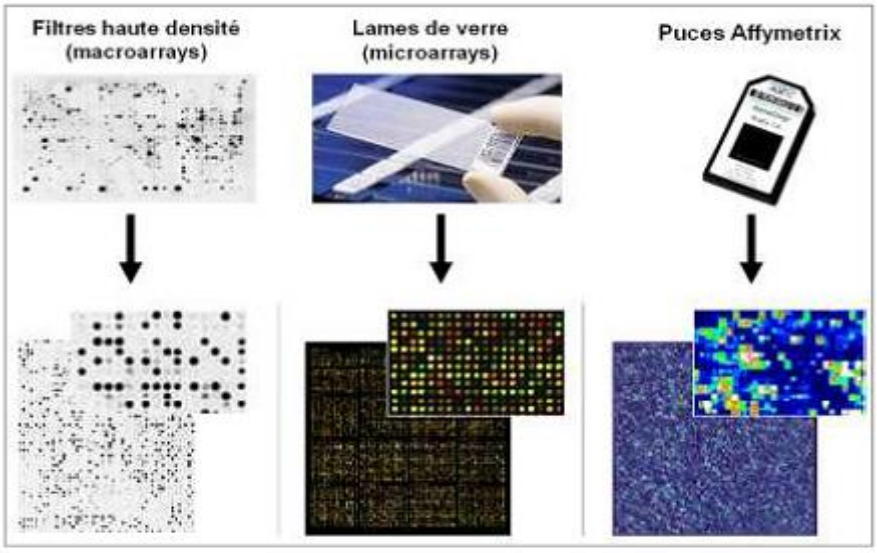
chaque marqueur) en niveau de gris. Ces images sont ensuite converties en pseudo-couleur et fusionnées pour être analysées par un logiciel d'analyse d'images.

2.2. Principe

Sur une puce à ADN, des dizaines de milliers d'hybridations peuvent être réalisées simultanément. Les hybridations se font entre des sondes nucléotidiques (probe ou reporters) ordonnées sur un support solide et des cibles (target) marquées, présentes dans un mélange complexe (Dugganet al., 1999) (Cf. Figure. 9). Les sondes et les cibles représentent respectivement les gènes du transcriptome à analyser. Le signal d'intensité, recueilli pour chaque hybridation spécifique « sonde-cible », permet d'apprécier le niveau d'expression de chaque gène étudié dans le tissu analysé. Un profil d'expression est obtenu pour chaque échantillon.

2.3. Technologies

La particularité des puces à ADN, par rapport aux macroarrays, réside dans la miniaturisation du procédé permettant l'utilisation d'une moindre quantité de matériel génétique pour une densité plus importante de sondes. Plusieurs types de puces à ADN existent selon le support, la nature des fragments fixés à la surface, le mode de fabrication, la densité, le mode de marquage des cibles et les méthodes d'hybridation (Cf. Tableau 2). Les supports sur lesquels sont fixées les sondes sont des supports solides, de surface plane généralement inférieure à $1cm^2$. Les matériaux qui les composent peuvent être du verre, des polymères, du silicium, de l'or ou encore du platine. Quel que soit le support choisi, il est traité pour former un réseau dense et régulier de micro-surfaces où seront greffées les sondes. Les sondes sont qualifiées de « gène reporté » car elles représentent des fragments de gènes et rapportent leur niveau d'expression. Ces gène reporters, ordonnés sur les lames, peuvent être des produits de PCR (puce à ADNc) (Schenaet al., 1995) ou des oligonucléotides plus ou moins longs (25 à 70 mers). Les produits de PCR et les oligonucléotides issus de synthèses chimiques (50-70 mers) sont greffés sur les puces à ADN par adressage mécanique ou électrochimique (Leung et Cavalieri, 2003). Les oligonucléotides peuvent également être synthétisés in situ. Breveté par la société Affymetrix®1, la synthèse in situpar photolithographie, ou adressage photochimique, rappelle une technique couramment utilisée pour la fabrication des puces électroniques (Lockhartet al., 1996).



Support	Membrane de nylon	Verre ou silice avec revêtement Chimique	Verre avec revêtement chimique
Densité	quelques centaines des spots/cm^2	1000-10000 spots/cm^2	~ 250 000 spots/cm^2
Sonde	ADNc	ADNc ou oligonucléotides	oligonucléotides (synthèsein situ)
Longueur de la sonde (nucléotides)	100 à 1000	ADNc ~100-1000 oligonucléotides ~ 30-70	25-60 Cibl
Cibles	ADNc	ADNc	ARNc
Marquage de l'échantillon	Radioactivité (33P)	Fluorescence (double : cyanine3 et 5)	Fluorescence (simple : biotine-streptavidine)
Quantité d'échantillon nécessaire (µg)	1-100	10-100	0.05-5
Sensibilité	+++	++	++
Spécificité	++	+++	+++
Principales applications	Analyse du niveau d'expression des Gènes	Analyse du niveau d'expression des gènes, ChIP-Chip, CGH-array …	Analyse du niveau d'expression des gènes, étude des polymorphismes…

Tableau 2 : Exemples de technologies de puces à ADN

Les cibles sont les échantillons à étudier. Elles peuvent avoir différentes origines (tissu, une culture cellulaire…) et de différentes natures (ARNm, ADNc…). Selon la technologie de puce utilisée, les cibles sont identifiées par un marquage radioactif ou fluorescent. Bien que moins sensibles que les marquages radioactifs (Tab. 1.), certains systèmes de marquages

fluorescents présentent l'avantage de pouvoir identifier plusieurs cibles sur la même puce. Par exemple, un tissu « anormal » peut être marqué par une cyanine verte (Cy3) et un tissu « sain » peut être identifié par une cyanine rouge (Cy5) (Figure. 9). Le rapport (ratio)des intensités obtenues pour chaque fluorochrome offre une comparaison directe des variations d'expression entre les deux échantillons.

La lecture des résultats d'hybridation se fait grâce à un scanner. Dans le cas des technologies à fluorescence, son principe est celui d'un microscope confocal couplé à un ou plusieurs lasers. Chaque laser excite spécifiquement un fluorochrome. L'émission est amplifiée par un photomultiplicateur et transformée en signal digital, i.e. en image. Chaque pixel de l'image scannée représente une mesure de fluorescence. Pour les puces à ADN deux couleurs, deux images en niveau de gris sont générées (une pour chaque fluorochrome). Ces images sont converties en fausses couleurs (allant généralement du vert au rouge) et superposées. Un logiciel d'analyse d'images extrait des informations qualitatives (diamètre, niveau de saturation) et semi quantitatives (intensité du signal et du bruit de fond) pour chaque complexe sonde-cible (spot) dans chacun des fluorochromes. Des méthodes et outils informatiques sont ensuite nécessaires pour analyser et extraire la connaissance des données.

Le choix de l'unité INSERM U533, au sein de l'IFR26 et de la plate-forme puce à ADN Ouest Genopole® de Nantes, s'est tout d'abord porté sur l'emploi des puces à ADNc puis à oligonucléotides longs (50 mers). Ces oligonucléotides sont issus d'une synthèse chimique et sont adressés mécaniquement sur les lames de verre. Le marquage des cibles se fait au moyen de deux fluorochromes, les cyanines 3 et 5. Par conséquent, dans la suite du présent manuscrit, seules les questions concernant cette technologie seront abordées.

2.4. Applications des puces pour la génomique

De nombreuses applications des puces se développent pour l'étude du génome des organismes en dehors de la simple expression des gènes. La description ci-dessous reste succincte et non exhaustive puisque le reste de cette thèse ne porte pas sur ses applications. On pourra se référer à la revue de Mockler *et al.* (Mockler, 2005) pour de plus amples détails.

2.4.1. Etude du polymorphisme

Au départ, les puces n'ont pas été mises au point afin d'étudier le niveau d'expression des gènes mais afin de séquencer les génomes. Grâce aux puces à ADN, différents aspects génomiques sont étudiés.

Ces expériences requièrent parfois l'utilisation de puces génomiques, c'est-à-dire des puces où sont représentées à la fois des séquences de gènes et des séquences intergéniques.

En 1988-89 Fodor *et al* et Southern *et al.* de la société Affymetrix ont développé une méthode de séquençage par hybridation (*sequencing by hybridization* SBH). Le gène étudié est considéré comme un ensemble de plusieurs séquences chevauchantes dont la détermination simultanée puis l'assemblage permettent de reconstituer la séquence (Hacia, 1999). Pease *et al.* en 1994 ont complété cette technologie. Afin d'avoir des détails sur les techniques existantes et sur leurs utilisations, on peut se référer à la revue (Hacia, 1999). Les puces à ADN sont actuellement utilisées pour le re-séquençage afin d'élucider les différences entre des séquences d'origines différentes. Plus particulièrement, on recherche des mutations entre différents individus ou populations comme le séquençage de différentes souches de streptocoques afin d'adapter les traitements aux infections (Davignon et al., 2005).

Des puces à oligonucléotides ont été fabriquées afin d'identifier les polymorphismes sur une base spécifique (*single nucleotide polymorphism* ou SNP). Les oligonucléotides sont organisés en tétrades au sein desquelles une séquence est strictement identique à celle de type sauvage alors que les trois autres sont caractérisées par une substitution de base localisée au milieu de la séquence. Ces puces (Lipshutz, 1999) permettent la détection de différents allèles ainsi que leurs positions (génotypage) mais sont inadaptées à l'étude de polymorphisme long comme des délétions ou des insertions. Plusieurs génotypages ou criblages de mutations ont été effectués grâce à cette technique (Ahrendt et al., 1999) (Wang et al 1998) (Winzeler et al., 1998) hrendt . La revue (Shi, 2001) récapitule l'ensemble des techniques utilisées actuellement pour la détection de SNP dont les puces à ADN Les puces peuvent également être utilisées afin d'étudier la séquence d'organismes proches de celui pour lequel elles ont été conçues. Jusqu'à présent, la séquence complète du génome n'est disponible que pour un petit nombre des systèmes modèles (Renn et al., 2004). Pour les organismes non modèles, l'approche la plus accessible est l'utilisation de puces à ADNc afin d'établir des différences de séquences (insertion/délétion) avec des organismes proches. Cette application est également étendue à la recherche de régions délétées ou insérées entre différentes souches ou populations. Lashkari *et al.* ont présenté des puces permettant la comparaison génomique de deux souches de levure. Après hybridation de l'ADN génomique marqué, les gènes pour lequel le signal est extrêmement faible sont d'éventuels gènes délétés ou extrêmement divergents entre différentes souches. Une autre technique, les puces CGH (*comparative genome hybridization*), puces génomiques, permettent également l'identification de grandes délétions/ insertions. Comme des cancers peuvent se caractériser par des délétions/ insertions ou amplifications de certaines séquences, les puces CGH ont été utilisées sur des lignées cellulaires tumorales (Kallioniemi et al., 1992) (Pollack et al., 1999) et ont permis de détecter des séquences amplifiées et délétées dans différents types de cancer.

2.4.2. Analyse des mécanismes de régulation d'expression (hybridation d'ADN)

Toujours grâce à l'hybridation de l'ADN aux puces, des études se sont focalisées sur une échelle plus réduite, le gène et ses environs. Les puces ont ouvert la voie à des cartographies des motifs de régulations de l'ADN à grande échelle. On cherche à identifier des régions d'ADN se liant avec les facteurs de transcription. Pour cela, la chromatine liée à des protéines est immunoprécipitée grâce à des anticorps spécifiques de la protéine étudiée. La séquence est ensuite identifiée grâce à des puces génomiques. Cette technique, ChIP-on-chip, et les découvertes réalisées grâce à elle sont décrites en détail dans les revues (Rodriguez and Huang, 2005) (Hanlon and Lieb, 2004). L'étude de la liaison de deux facteurs de transcription (Iyer et al., 2001) a montré leur implication dans l'activation de gènes qui font partie de deux voies métaboliques différentes: la synthèse de la membrane cellulaire et la réplication et la réparation de l'ADN. Cette spécialisation des facteurs de transcription permet d'expliquer la régulation de processus cellulaires indépendants.

Globalement, les puces génomiques permettent donc d'aborder le système de régulation d'expression de manière complètement nouvelle. Ainsi, une grande partie des sites de fixation de facteurs de régulation identifiés grâce à cette technique se situent en dehors des régions promotrices prédites jusqu'alors (Mockler et al., 2005) (Cawley et al., 2004).

Une autre application concerne la régulation de l'expression des gènes *via* la méthylation des cytosines de l'ADN. Plusieurs techniques sont utilisées afin d'identifier les sites de méthylation de l'ADN (cf. la revue (Mockler et al., 2005)). Globalement, une réaction préalable permet de différencier les cytosines où l'ADN est méthylé de celles où il ne l'est pas. Des enzymes de clivage pour les ADN non méthylés ou la conversion des cytosines en uracile grâce au bisulfate de soude sont deux types de réactions utilisées. Ces morceaux d'ADN méthylés sont ensuite identifiés grâce à l'hybridation sur des puces.

2.4.3. Analyse au niveau de l'ARN et des mécanismes d'expression

Les puces à ADN génomique permettent de découvrir de nouveaux gènes (Mockler et al., 2005) et de définir leur structure (intron/exons pour les eucaryotes). Des puces génomiques qui représentent soit l'ensemble du génome soit des portions du génome régulièrement espacées permettent la détection des portions d'ADN transcrites. Pour l'homme et le génome des plantes, les puces génomiques révèlent que certaines régions considérées comme non codantes sont effectivement exprimées (Kapranov et al., 2002) (Kampa et al., 2004) (Yamada et al., 2003). Parmi les séquences exprimées, environ 50% étaient annotées comme étant des gènes alors que 50% ne présentaient aucune annotation codante. La revue de Johnson *et al.*

(Johnson et al 2005) permet d'approfondir cette question et montre notamment qu'une partie des séquences transcrites sont en fait des ARN non traduits ou correspondent à des transcrits anti-sens.

Une autre application des puces à ADN est l'analyse des transcrits alternatifs difficilement prédits par les programmes actuels de la bioinformatique. Cela a été appliqué chez la levure (Castle et al., 2003). La technique n'est pas encore sans défaut. On peut se référer à la revue afin de connaître ses limites et les différentes applications existantes (Lee and Roy, 2004).

L'étude des séquences cibles des protéines liant l'ARN permet de comprendre les mécanismes de modifications post-transcriptionnelles grâce aux puces. Cette analyse peut être couplée avec l'épissage alternatif afin de comprendre comment les différents ARNm sont synthétisés (Castle et al 2003). Le principe est simple et ressemble à celui de ChIP-on-chip : les séquences d'ARN qui se lient avec une protéine sont purifiées par immunoprécipitation ou par colonne d'affinité. La puce à ADN permet ensuite d'identifier ces séquences. La revue (Mata et al. , 2005) référence l'ensemble des applications des puces quant à l'identification des séquences d'association entre l'ARN et des protéines.

Enfin, de nouvelles puces sont actuellement développées afin d'étudier l'effet des *ARN interference* (ARNi) sur les cellules (Wheeler et al., 2005). Des lots d'ARNi sont fixés sur la puce puis mis en présence de cellules à transfecter. Le but est d'observer les phénotypes obtenus sur les cellules transfectées par un ARNi particulier.

2.5. Applications des puces pour l'étude de l'expression des gènes

Actuellement, la principale utilisation des puces à ADN est de mesurer les niveaux d'expression relatifs des gènes dans différentes conditions. Si la génomique permet de connaître la structure du génome et les différents gènes existants, elle ne permet d'aborder la fonction des gènes que *via* les similitudes de séquences, c'est-à-dire de manière statique et dépendante des connaissances déjà acquises. Or, comme le précise Lipshutz *et al.* en 1999 (Lipshutz et al.,1999), pour comprendre la fonction d'un gène il est utile de savoir quand et où il est exprimé et dans quelles circonstances son niveau d'expression est affecté. Les puces permettent d'aborder ces problèmes. Néanmoins outre les études concernant un gène précis, elles permettent d'aborder l'étude des réseaux fonctionnels existants. Au départ, les puces engendraient de nombreux espoirs. En 1999, Brown et Botstein (Brown and Botstein, 1999) parlent de nouvelles cartes génomiques fondées sur les données d'expression, qui permettraient de comprendre les fonctions des gènes ou leur régulation. Pour eux, dans un avenir proche, c'est à dire à peu près de nos jours, les effets des mutations pour chaque gène de la levure seraient connus. Tel n'est pas le cas. Si les puces constituent un outil qui fait

avancer la compréhension du fonctionnement des gènes, il ne répond pas entièrement aux attentes formulées par Brown et Botstein du fait de ses limites. Malgré cela, de nombreux aspects de l'expression des gènes ont été abordés grâce aux puces comme nous allons le détailler ci-après. On peut classer les applications existantes des puces en différents types (Kutalik et al., 2004) : la comparaison des niveaux d'expression de gènes selon différentes conditions, l'identification de gènes fonctionnellement liés, la recherche de gènes discriminants grâce au clustering, l'approche des réseaux d'interaction géniques grâce aux séries temporelles.

2.5.1. Comparaison des niveaux d'expression des gènes selon différentes conditions

La première utilisation des puces pour les niveaux d'expression (Schena et al., 1995) a été l'analyse des différences d'expression de 45 gènes entre les feuilles et les racines d'*Arabidopsis thaliana*. Depuis, de nombreuses expériences de recherche de gènes différentiellement exprimés entre plusieurs conditions ont été réalisées. Le but est de comprendre certains phénomènes biologiques et d'identifier les gènes impliqués dans ces processus.

Plusieurs phénomènes ont été étudiés : la différenciation et la maturation de cellules dendritiques humaines et les gènes impliqués dans ces phénomènes (Le et al., 2001). Clark *et al.* ont identifié des gènes dont l'expression induit le passage d'une cellule tumorale en métastase chez la souris. DeRisi *et al.* se sont intéressés aux différences d'expression entre des lignées de cellules cancéreuses humaines et des lignées mutantes qui suppriment les tumeurs. Hedenfalk *et al.* ont étudié les différences d'expression entre des mutations de BRCA1 et BRCA2 dans des cas de cancer du sein. En plus de ces études du cycle cellulaire et du cancer, certaines applications visent à analyser le comportement des gènes face à des modifications environnementales notamment le stress.

Les puces ont également permis d'identifier les gènes dont l'expression est gouvernée par différents régulateurs, par les brassinostéroïdes et les gibbérellines pour le riz (Yang and Komatsu, 2004) ou encore d'identifier les mécanismes de régulation des défenses d'*Arabidopsis thaliana* (Eulgem,2005).

Différents métabolismes ont encore été étudiés comme le mécanisme de l'assimilation du soufre chez *Bacillus subtilis* (Sekowska et al., 2001).

2.5.2. Identification de gènes fonctionnellement liés

Une des applications les plus courantes des puces est l'étude de gènes dont les niveaux d'expression sont corrélés. Eisen *et al.* (Eisen et al., 1998) ont mis au point une méthode permettant d'identifier les gènes partageant le même profil d'expression. L'hypothèse suppose que des gènes co-exprimés sont impliqués dans un même processus cellulaire ou voie métabolique. Il serait ainsi possible d'identifier les fonctions de gènes jusqu'alors inconnues en analysant les fonctions des gènes ayant le même profil d'expression. La revue de Niehrs and Pollet, 1999) (Niehrs ans Polletrecense, chez les eucaryotes, différents groupes de gènes qui ont été identifiés comme co-exprimés en 1999 et qui partagent une fonction commune.
Une autre hypothèse est que les gènes identifiés comme co-exprimés seraient régulés par des facteurs de régulations communs et auraient donc, dans leur promoteur, des sites de fixation identiques. Plusieurs études combinent ces deux hypothèses comme celle de Chu *et al.* en 1998 sur les différentes étapes de la sporulation chez la levure. Les différents prélèvements ont permis de distinguer différentes phases de la sporulation et de caractériser les gènes associés. Lors de leur travail, ils ont tenté d'identifier des séquences nucléiques communes aux gènes identifiés afin de découvrir des sites de régulations communs. Des études similaires ont été réalisées afin d'étudier le cycle cellulaire de la levure (Cho et al., 1998) (Spellman et al. 1998), le cycle circadien d'*Arabidopsis thaliana* (Schaffer et al., 2001). Les cycles cellulaires de fibroblastes humains (Cho et al., 2001) ou de cellules cancéreuses humaines (Whitfield et al., 2000) ont également été étudiés afin d'identifier de nouveaux gènes impliqués dans différentes phases du cycle. En plus des séries temporelles, d'autres études portent sur la compréhension de la réponse d'organismes à différentes conditions comme l'étude des réponses de la levure à différents stress environnementaux (Gasch et al., 2000).

2.5.3. Recherche de gènes discriminants

Les puces peuvent également être un outil de diagnostic afin de discriminer entre différents types cellulaires. Dans le domaine clinique notamment, la recherche vise à identifier des gènes marqueurs qui permettent de réaliser un diagnostic fiable de maladies comme le cancer. Alizadeh *et al.* en 2000 ont démontré qu'il existait des signatures d'expression différentes pour des cellules de lymphomes qui proviennent soit de patients qui répondent à un traitement soit de patients qui ne répondent pas et succombent. Van't Veer *et al.* ont ainsi utilisé les puces afin de déterminer un ensemble de gènes dont les profils d'expression « prédisent » les possibilités de métastases pour des tumeurs du sein. Plusieurs études (Garber et al. ; 2001) (Beer et al., 2002) ont identifié des groupes de gènes pouvant prédire la possibilité de survie pour des patients atteints de cancer des poumons.

L'étude de l'expression des gènes permet également de distinguer des types cellulaires ou de cancer difficilement identifiables par les critères physiologiques classiques (Golub et al., 1999) (Bittner et al., 2000) (Su et al., 2001). Les puces permettent donc de faire des diagnostics de maladie et d'adapter les traitements aux différents types de maladies.

2.5.4. Approche des réseaux d'interaction géniques

Les puces à ADN mettent à disposition des jeux relativement larges de données d'expression d'une grande partie des gènes d'un organisme. Parallèlement à l'étude de groupes de gènes co-exprimés se sont développées des méthodes qui permettent d'inférer les réseaux d'interaction géniques (régulation) d'un organisme. Ces études mènent à une visualisation du réseau plus ou moins dense. Rung et al. ont analysé un important jeu de données chez la levure.

2.6. La partie technique et biologique des puces

2.6.1. Le principe des puces

Le principe de la puce à ADN est l'hybridation spécifique de deux molécules présentant la même séquence. Sur une surface de quelques centimètres carrés, des fragments synthétiques d'ADN (les sondes) sont greffés et espacés de quelques micromètres. Les sondes sont regroupées en spots représentatifs de chacun des gènes étudiés. Ce micro-dispositif est ensuite mis au contact des acides nucléiques à analyser, au cours de l'étape d'hybridation. Ces acides nucléiques, appelés cibles, correspondent aux ARNm ou aux ADNc préalablement couplés à un marqueur fluorescent ou radioactif. Ce contact entre cibles et sondes conduit à la formation d'hybrides qualifiés par leurs coordonnées et quantifiés grâce à la lecture des signaux radioactifs ou fluorescents. Les sondes sont toujours en excès par rapport aux cibles.

L'utilisation des puces à ADN se décompose en différentes étapes qui suscitent plus ou moins l'attention des biologistes : le choix d'un plan expérimental, le choix ou la fabrication de la puce utilisée, l'extraction des ARNm, la synthèse et le marquage des ADNc, l'hybridation, la lecture des données, les étapes de normalisation et d'analyse et enfin la confrontation avec des données externes. Dans cette partie, nous allons nous attacher à la description des différentes étapes de manipulation et d'analyse des résultats ainsi qu'à celle des biais éventuels qui en résultent. Dans le chapitre suivant, nous nous concentrerons sur les particularités des données et leur analyse. Nous rappelons que dans l'ensemble de ce manuscrit le terme de puces à ADN correspondra à l'ensemble des techniques permettant l'acquisition des données de transcriptome et comprendra donc les micro-réseaux, les filtres à membranes et les puces à oligonucléotides. Toutefois, dans la description des différents supports existants, nous ferons la distinction entre les différentes techniques.

2.6.2. Le choix du plan d'expérience

Le plan d'expérience est une des étapes les plus importantes (Churchill , 2002) (Yang and Speed, 2002) (Simon et al., 2002) (Kerr and Chutchill, 2001) (Barrett and Kawasaki, 2003) (Leung and Cavalieri, 2003) dans la mise en place de la mesure du transcriptome et ce, d'autant plus que les puces génèrent de grandes quantités de données qu'il faut ensuite analyser. Malheureusement, cette étape est fréquemment négligée (Kerr and Chutchill, 2001) lors des expériences et se résume souvent au choix d'un protocole expérimental.

Le choix du plan expérimental dépend des questions posées lors de l'expérience et des hypothèses qui en résultent (Kerr and Chutchill, 2001) (Hess et al., 2001). Il comprend :

- la détermination des conditions expérimentales étudiées et donc des facteurs pris en compte dans l'étude
- leur agencement.

a) Première étape : définir précisément le but de l'expérience et le(s) facteur(s) d'intérêt

La première question à se poser lors de la mise en place d'une expérience de transcriptome est de définir précisément son but, ce que l'on recherche. Si cette interrogation semble simple en théorie, elle n'est pas si évidente en pratique. Définir précisément quel est le facteur d'intérêt n'est pas une chose aisée. Parfois la question posée correspond à une combinaison de facteurs. Par ailleurs, si le but est de comprendre un mécanisme physiologique particulier, le choix des conditions étudiées influencera grandement les résultats. Ainsi, si des conditions relativement éloignées sont choisies (exemple un type de cancer *versus* en bonne santé), les niveaux d'expression de nombreux gènes seront affectés : les gènes directement impliqués dans le mécanisme à étudier (gènes impliqués dans le type de cancer) mais aussi l'ensemble des gènes dont les niveaux d'expression ont été affectés par ces fortes différences de conditions physiologiques. Bref, il sera relativement difficile d'identifier les gènes responsables ou impliqués dans un mécanisme précis.

Par ailleurs, quand il y a beaucoup de variations non contrôlées au cours de l'expérimentation, il devient difficile de distinguer des fluctuations aléatoires d'un effet spécifique (Hess et al. 2001). Si l'on cherche à identifier un nombre réduit de gènes responsables d'un mécanisme précis, il est avantageux d'obtenir des profils d'expression relativement proches pour la plupart des gènes. Il faut donc choisir des conditions qui ne font varier que très peu de facteurs et n'impliquent qu'un nombre réduit de gènes.

Dans leur étude sur le mécanisme d'entrée et de métabolisme de deux sources de soufre, Sekowska *et al.* ont choisi d'étudier le transcriptome dans des conditions très proches, à

savoir la croissance des bactéries dans un milieu contrôlé en présence de méthionine ou de methylthioribose comme source de soufre (au plus une dizaine de gènes impliqués théoriquement). Malgré les conditions contrôlées, de nombreux gènes étaient détectés car le changement de conditions déclenchait leur voie métabolique. Ici la voie de la synthèse de l'arginine a été activée indirectement lors du changement de source de soufre. Les gènes de synthèse de l'arginine ne sont cependant pas impliqués directement dans l'assimilation du soufre. Par ailleurs, l'expérience a révélé le déclenchement involontaire de transitions dans les conditions environnementales avec les cascades de régulations qu'elles provoquent. Ces différences d'expression seraient dues à des différences de températures de la pièce pour la période comprise entre la phase préculture et la phase de croissance. Ainsi, même en choisissant des conditions proches, on retrouve d'autres mécanismes qui se mettent en place indépendamment du phénomène étudié.

De petites différences dans des conditions expérimentales non contrôlées peuvent mener à des différences d'expression de gènes visibles et cohérentes, concernant par exemple les gènes de compétence ou de sporulation. Si l'on avait choisi des conditions lointaines, il aurait été encore plus difficile de distinguer les gènes réellement impliqués dans le mécanisme d'intérêt des gènes annexes.

b) Deuxième étape : choix des autres facteurs pris en compte

Généralement, les facteurs autres que le facteur d'intérêt ne sont pas identifiés en tant que tels. On parle souvent de répétitions ou de réplications. Les réplications ou répétitions des conditions expérimentales sont nécessaires afin de distinguer les variations d'expression présentes par hasard de celles reproductibles et réellement liées au facteur d'intérêt. Sans réplication ou répétition il est impossible d'estimer l'erreur ou le bruit, paramètre nécessaire pour les méthodes d'analyse ultérieures (Kerr and Churchill, 2001).

Il est possible de distinguer deux niveaux de réplications :

- les répétitions techniques qui correspondent, par exemple, à deux spots du même gène sur la puce ou le même ARNm hybridé sur plusieurs puces ou encore différents protocoles d'extraction ou de marquage. Ces répétitions permettent d'évaluer la variabilité liée au protocole expérimental.
- la réplication vraie ou biologique comme l'ARNm de plusieurs individus, échantillons, spécimens. Cette réplication permet de prendre en compte la variabilité biologique. Les conclusions de l'analyse portent plus sur la population étudiée que sur l'échantillon en particulier. Elle comprend également des réplications de l'expérience à deux jours/ temps

différents afin d'étudier la reproductibilité de l'expérience dans le temps. Les résultats obtenus sont donc plus fiables et reproductibles que si aucune réplication biologique n'avait été réalisée (Allison et al., 2006). Par rapport aux réplications et aux répétitions, on peut trouver deux types de comportements erronés dans la mise en place des puces. Le premier est de craindre que les différences génétiques n'interfèrent dans les effets du facteur d'intérêt. Ainsi, certains expérimentateurs cherchent à homogénéiser les individus ou échantillons utilisés ou encore à sélectionner une lignée consanguine la plus adaptée au problème étudié. Or l'article de Turk *et al.* montre que les variations d'expression entre deux lignées consanguines de souris sont relativement faibles. Ceci est rendu particulièrement évident par la comparaison du nombre de gènes différentiellement exprimés entre deux lignées ou entre un tissu atteint d'une maladie et un tissu sain. Ils en concluent que le profil génétique ne devrait interférer que marginalement dans l'analyse du transcriptome. Au contraire, Whitehead et Crawford constatent que dans l'étude d'un groupe de gènes essentiels dans le métabolisme cellulaire (192 gènes), seuls 31% des gènes différentiellement exprimés entre deux tissus sont conservés entre des populations de poissons différentes. Ils en concluent que lorsque l'on fait les mesures sur une seule population, l'observation de différences très significatives d'expression dans certains tissus ne correspond pas nécessairement à des gènes représentatifs des différences fonctionnelles ou morphologiques entre ces tissus. Les grandes variations d'expression entre individus (Whitehead and Crawford, 2005) montrent qu'il est important d'inclure des replicats biologiques à l'intérieur des groupes de traitement afin d'attribuer des différences d'expression au traitement plutôt qu'à des variations entre individus ou populations. Les gènes détectés indépendamment de l'individu ou de la population expliqueront de manière plus probable les différences fonctionnelles et morphologiques étudiées.

Le deuxième comportement erroné est d'accorder beaucoup plus d'importance aux répétitions techniques qu'aux réplications biologiques. Souvent, on observe des plans d'expérience où il y a de nombreuses répétitions techniques mais peu de réplications biologiques. S'il est naturel d'avoir des résultats différents entre deux individus, on peut trouver plus inquiétant de ne pas avoir le même résultat pour un même échantillon selon le marquage ou la position sur la puce. La question se pose alors de la fiabilité et de la précision des mesures. Actuellement les mesures de puces sont globalement fiables dans la mesure où elles sont reproductibles au sein d'un même laboratoire. Il faut garder à l'esprit que tout comme la biologie, l'instrumentation subit l'influence de facteurs environnementaux et les variations techniques ne sont pas plus étonnantes que les variations biologiques. Si l'accent est mis sur les répétitions techniques au dépend des répétitions biologiques, les résultats obtenus sont alors précis et indépendants de

la technique employée mais ne sont pas forcement généralisables à tout échantillon biologique similaire.

Pour conclure, lors de l'élaboration du plan expérimental, il est donc important de définir les différentes questions posées et de les hiérarchiser entre elles mais aussi de prendre en compte les contraintes matérielles comme le nombre d'hybridations possibles ainsi que les sources de variabilités techniques et biologiques (Churchill, 2002). On classe les facteurs de variations d'expression en trois catégories (Churchill, 2002) :

- la variabilité biologique qui est intrinsèque à tous les organismes et dépend des facteurs génétiques et environnementaux. Elle est évaluée grâce aux différents réplicats
- la deuxième est technique (puce/extraction/ marquage/hybridation). Elle est évaluée grâce aux répétitions techniques
- la dernière est l'erreur de mesure.

Dans une expérience sur l'étude du métabolisme du soufre chez *B. subtilis*, les sources de variations les plus fortes sont la quantité d'ARNm utilisée pour synthétiser de l'ADNc qui influence plus certains gènes que d'autres (variabilité technique) et la difficulté de reproduire exactement les conditions expérimentales d'une expérience à une autre (jour de l'expérience, variation biologique). Les variations d'expression selon le facteur d'intérêt étaient faibles : cela était souhaité car les différences de conditions ne devaient faire intervenir qu'une dizaine de gènes.

2.6.3. L'agencement des différents facteurs : le plan d'expérience proprement dit

Prenons le cas où l'impact du facteur expérimental d'intérêt (mutation, conditions de culture, …) est confondu avec les variations causées par d'autres facteurs, par exemple deux jours de culture différents nécessités par le plan expérimental. Il est impossible de savoir si les variations d'un gène donné sont dues au facteur d'intérêt ou à un autre facteur (Kerr andChurchill, 2001). Pour être plus clair, les résultats ne peuvent pas être interprétés correctement à cause du biais expérimental et l'ensemble de l'étude ne repond pas au but initial (Yang and Speed. 2002)

L'agencement des facteurs par rapport aux conditions d'expérience est donc primordial.

Revenons maintenant désormais à l'élaboration du plan expérimental. Il comprend différents points :

1. La définition de l'unité expérimentale : quel est le meilleur échantillon à utiliser pour réduire par exemple, la variance biologique : *pooler* ou non les échantillons. En théorie, rassembler les échantillons de plusieurs individus devrait augmenter la précision de la mesure en réduisant la variance des comparaisons d'intérêt (Whitehead and Crawford, 2005).

Cependant, cette option présente l'inconvénient de potentiellement laisser un seul échantillon avoir une influence trop forte sur les résultats. Par ailleurs, elle ne permet pas d'estimer la variance biologique entre individus (Whitehead and Crawford, 2005). Le pooling s'avère parfois nécessaire notamment lorsque la quantité d'ARNm prélevée est faible.

2. Le nombre de répétitions. Actuellement, il n'existe pas de consensus sur le nombre de répétitions nécessaire. La réponse différerait selon l'expérience (la thématique, l'organisme, les conditions et autres). La tendance générale est : « plus on fait de répétitions, mieux c'est ». Il existe cependant des méthodes pour définir *a posteriori* si le nombre d'échantillons biologiques est suffisant pour discriminer différents groupes cellulaires (Hwang et al., 2002) par exemple différents types de leucémies. Mais comme ces méthodes sont réalisées *a posteriori*, elles ne sont pas de grande utilité pour définir un plan expérimental.

3. La manière d'associer les échantillons. Les plans d'expérience complets et réguliers sont ceux qui permettent les analyses les plus puissantes. Sinon, pour un nombre donné de puces, les plans d'expérience équilibrés pour les facteurs d'intérêt sont les plus efficaces (Kerr andChurchill, 2001). Il n'est pas toujours facile de les mettre en place, car soient ils requièrent un grand nombre de mesures, ce qui pose des problèmes de coût, soit parce qu'un individu – chez l'Homme par exemple– ne peut présenter les différents modes du facteur d'intérêt. De manière caricaturale, un individu peut être difficilement à la fois malade et sain. Plusieurs plans expérimentaux sont décrits dans Kerr *et al.* dans le cadre d'une puce à ADN avec hybridation simultanée de deux échantillons, un marqué en rouge, l'autre en vert :

- le carré latin communément appelé dans le domaine *dye swap* : les échantillons d'ARNm sont marqués par deux marqueurs différents une fois rouges et une fois verts. Ce plan d'expérience permet de mesurer les biais du marquage sur les mesures de transcriptome.
- un autre plan communément utilisé est l'utilisation d'une mesure de référence. Les échantillons d'intérêt sont toujours marqués de la même couleur et l'échantillon de référence de l'autre. Ce plan d'expérience ne permet donc pas l'estimation de l'effet fluorochrome (une répétition technique). En outre, avec ce plan expérimental on obtient beaucoup plus de mesures de l'échantillon de référence que des échantillons d'intérêt. Il y a donc une perte d'argent et de temps en faisant des mesures « inutiles » (Kerr andChurchill, 2001) (Thygesen and Zwinderman, 2004).
- utiliser à la place un plan équilibré permet d'acquérir deux fois plus de données pour autant de moyens (Kerr andChurchill, 2001). Les mesures ainsi effectuées sont donc beaucoup plus précises. Dans un plan équilibré, chaque mode de facteur expérimental est répété le même nombre de fois. Le plan expérimental équilibré de l'article de Sekowska *et al.* peut être une base pour mettre au point un plan expérimental adapté à la problématique choisie.

2.6.4. Le choix de la puce ou la préparation de la puce

Si une des premières mesures d'expression (Schena et al., 1995) a utilisé une puce contenant 45 ADNc d'*Arabidopsis thaliana*, les supports actuels permettent la mesure simultanée de l'expression de plusieurs dizaines de milliers de gènes. Désormais, il existe une grande variété de puces en adéquation avec l'objectif de l'expérience. Certaines de ces puces sont génomiques et représentent l'ensemble du génome de l'organisme, d'autres sont dédiées à quelques gènes généralement impliqués dans un même processus cellulaire. En plus de la diversité des gènes étudiés, il existe différents types de supports, de sondes, de densités et de marquages de la cible utilisés (Vrana et al., 2003). Le choix de la puce est plus ou moins ample suivant qu'on utilise des puces commerciales ou que l'on produise la puce. La description suivante précise les différents types de puces existants ainsi que leurs qualités respectives. Historiquement, les *macroarray*, les *microarrays* et les « véritables » puces à ADN correspondent à trois techniques différentes (Vrana et al., 2003).

- Les *macroarrays* utilisent comme sonde des clones d'ADN complémentaire (ADNc) disposés sur des membranes de nylon avec un espacement de l'ordre du millimètre en association avec des cibles radioactives. Les sondes représentent entre 200 et 8 000 gènes. Leur densité est donc relativement faible. Cette technique ne demande pas d'équipement particulier à part le *phosphorimager*.
- Les *microarrays* plus miniaturisés, comportent quelques milliers de gènes représentés par des produits PCR déposés tous les 200 à 400 microns sur une lame de verre et des cibles marquées par fluorescence. Cette technique permet l'hybridation compétitive.
- Les « véritables » puces à ADN associaient à chacun des gènes d'un organisme un ensemble de sondes sous la forme d'oligonucléotides synthétisés *in situ* sur la surface de la matrice. Les oligonucléotides mesurent au plus vingt-cinq bases à cause de l'efficacité finie de chaque étape (Barrett and Kawasaki, 2003). Chaque puce peut comporter jusqu'à 40 000 à un million d'oligonucléotides. Afin de mesurer la possibilité d'hybridation croisée, certains oligonucléotides correspondent à la séquence exacte et d'autres comportent une mutation « mismatch » au milieu.

Aujourd'hui, ces trois distinctions n'ont plus vraiment lieu d'être, d'autant plus que ces techniques sont utilisées de façon croisée comme le montre l'exemple de puces à ADN utilisant des produits PCR et des cibles radioactives. Globalement, il est encore possible de distinguer les *microarrays* des *macroarrays* du fait de la densité des sondes sur la puce. Les terminologies « puces à ADN » et « *microarray* » sont donc employées de façon indifférente. Les termes « biopuces » ou « microréseau » sont également employés.

Comme il est difficile de définir chaque type de puce précisément, nous aborderons plutôt leurs différents composants : le support, le type de sonde, le type de marquage.

a) <u>Le type de sonde</u>

✓ ***Les sondes ADNc***

L'utilisation de sondes ADNc ne nécessite pas la séquence complète du génome mais l'utilisation d'ADNc issus des différents EST (Expressed Sequence Tag) disponibles. Les banquesvd'ADNc correspondent à des inserts dans des plasmides ou des chromosomes bactériens (Holloway et al., 2002). L'ADNc est ensuite récupéré grâce à l'extraction et la purification des inserts puis à l'amplification PCR avec des primer universels. Les produits sont ensuite séparés sur gel d'électrophorèse, quantifiés et déposés sur le support (Barrett and Kawasaki, 2003).

Ce type de sonde est donc recommandé pour les organismes dont le génome n'est pas séquencé. Il est cependant également largement utilisé pour l'ensemble des organismes (dont le génome est séquencé ou non) car il ne nécessite pas de matériel spécifique comme les oligonucléotides. Ces sondes sont relativement longues puisqu'elles comprennent, en moyenne, entre une centaine et cinq cents paires de bases. Cette particularité limite les possibilités d'hybridation croisée sauf dans le cas de séquences très proches comme certaines familles de gènes très similaires chez les plantes ou pour des isoformes d'une même famille (Kothapalli et al., 2002). La modification du design des puces est par contre une chose aisée puisqu'il suffit de rajouter un spot contenant l'ADNc qui correspond au gène à étudier. Une des difficultés d'utiliser des ADNc est le maintien des banques d'ADNc qui peut s'avérer coûteux et lourd pour un laboratoire isolé (Barrett and Kawasaki, 2003).. Ces banques contiennent entre 1 et 5% de séquences redondantes, mal annotées ou encore contaminées (Holloway et al., 2002). Certaines banques d'ADNc commerciales sont désormais disponibles pour les organismes d'intérêt tels que l'homme, la souris, le rat et le chien. Par ailleurs Lipshutz *et al.* pointaient le risque de mauvaise identification du spot ou de l'ADNc déposé et du coup une mauvaise attribution du niveau d'expression. Ce risque a été confirmé par Kothapalli *et al.* qui a re-séquencé 17 ADNc qui correspondent à des gènes différentiellement exprimés dans leur expérience. Parmi ces dix-sept ADNc, quatre (24%) présentent des séquences incorrectes qui ne correspondent pas au gène qu'ils devaient représenter.

✓ **Les oligonucléotides courts**

Les puces à oligonucléotides sont fondées en majorité sur une technique de fabrication de puces informatiques (Fodor et al., 1991) adaptée ensuite par la société Affymetrix à l'étude de l'expression des gènes (Lockhart et al., 1996) (Lipshutz et al., 1999). Les sondes sont des oligonucléotides courts composés de 25 paires de bases synthétisées *in situ* par des dépôts de

couches successives de quatre nucléotides. La technique débute par l'attache à la surface de la puce de liaisons synthétiques munie de groupes chimiques qui peuvent être enlevés sous l'effet de la lumière. On envoie ensuite de la lumière à certaines localisations définies *via* un masque photolithographique afin de déprotéger les liants. Le rajout de deoxynucleosides avec un groupe photolabile permet la liaison avec les groupes déprotégés. Un autre masque dont la configuration varie pour chaque couche déposée et qui assure ainsi une succession correcte des bases est ensuite mis en place et ainsi de suite jusqu'à la synthèse complète des oligonucléotides. La Figure 10 illustre cette technique.

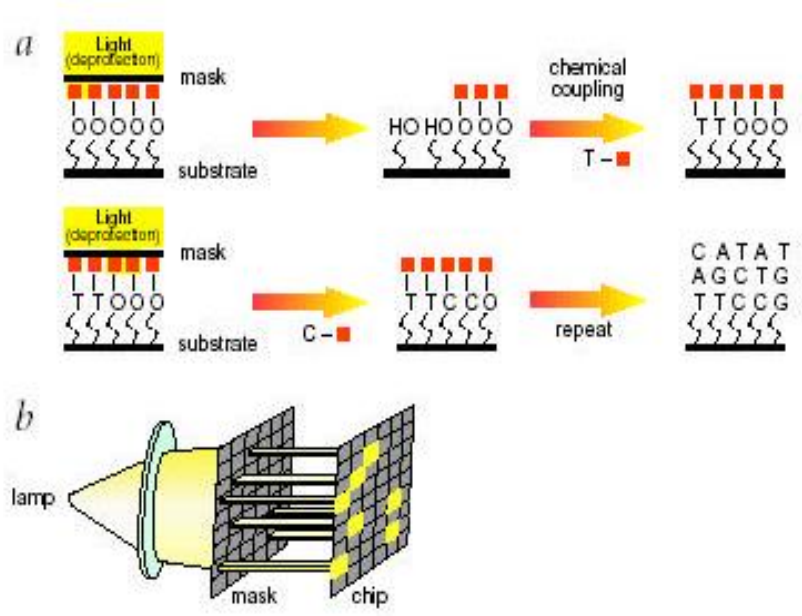


Figure 10 : Technique de synthèse des oligonucléotides courts (figure tirée de (Lipshutz et al., 1999)**)**
la Synthèse d'oligonucléotides par photo-activation. Des molécules terminées par un groupe protecteur photolabile sont fixées sur un support solide. De la lumière est dirigée à travers un masque afin de déprotéger et d'activer les sites sélectionnés. Des nucléotides protégés s'assemblent alors avec les sites activés. Le processus est répété : activation de différents sites puis assemblage des différentes bases. Ce procédé permet la construction de sondes ADN spécifiques sur chaque site. b : représentation schématique de la lampe, du masque et de la puce.

Toutefois, la technique de l'impression jet d'encre (Blanchard et al., 1996) est également utilisée afin de fabriquer les oligonucléotides *in situ*. Le principe est simple : une aiguille de l'imprimante permet d'appliquer une seule goutte de dix picolitres à une position identifiée. Les étapes sont les mêmes que la synthèse par photolithographie : protection, déprotection, synthèse. A la place de l'action de la lumière, des agents chimiques permettent la réalisation des réactions qui conduisent à la synthèse des oligonucléotides. Pour les puces Affymetrix, l'expression de chaque gène est mesurée par une vingtaine d'oligonucléotides différents (Barrett and Cavalieri, 2003). Afin d'identifier d'éventuelles hybridations croisées et éliminer le bruit résultant, chaque portion du gène choisie est représentée par des oligonucléotides qui

correspondent à la séquence exacte de la cible (PM ou Perfect Match) et par d'autres dont la séquence est identique sauf une mutation située à la position centrale (MM : Mismatch) (Figure 11).

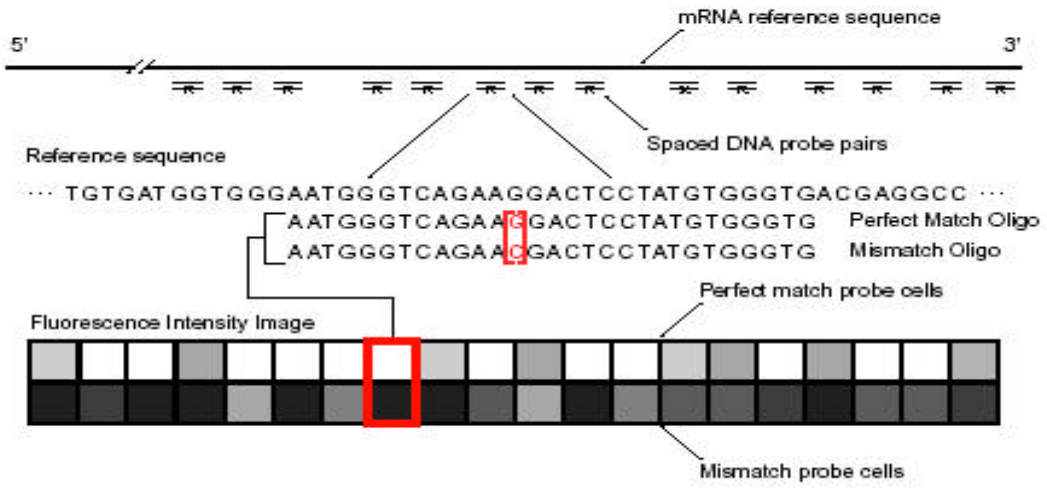


Figure 11: Exemple d'oligonucléotides PM et MM

Toutes les sondes sont de même longueur et présentent, autant que faire se peut, la même composition en G/C ce qui leur procure l'avantage d'avoir une température d'hybridation commune ou presque, contrairement aux ADNc. Par ailleurs, chaque spot contient la même quantité d'oligonucléotides contrairement, là encore, aux ADNc (Barrett and Cavalieri, 2003).

Les puces à oligonucléotides sont préférées pour l'analyse complète des génomes. La possibilité de représenter toute séquence présente dans un génome, la petite longueur de ces sondes et la possibilité de sondes chevauchantes multiples permet de détecter des caractéristiques génomiques comme de petits polymorphismes, des variants d'épissage, la distinction entre des membres d'une famille génique ou la distinction de régions répétitives.

Les oligonucléotides présentent l'avantage de ne pas à avoir à entretenir une banque d'ADNc et réduit le risque d'avoir des spots mal identifiés (Lipshutz et al., 1999). Chaque gène est représenté par plusieurs sondes, soit plusieurs mesures pour un même gène.

La principale difficulté selon Lipshutz *et al.* est la sélection des oligonucléotides dont la séquence doit être spécifique du gène et si possible non chevauchante avec les autres oligonucléotides chargés de détecter le même gène. Par ailleurs la synthèse *in situ* limite la possibilité d'adapter les puces aux différentes expériences. Il est difficile de rajouter de nouvelles sondes pour de nouveaux gènes ou portion d'ADN à identifier puisqu'il est alors nécessaire de refabriquer l'ensemble des masques photolithographiques. Par ailleurs, les

chercheurs n'ont pas accès à la séquence des sondes utilisées et doivent travailler sur les annotations fournies par le fabricant (Quackenbush, 2001).

- ✓ **Les oligonucléotides longs**

Ces sondes, comme leur nom l'indique, sont plus longues que les oligonucléotides courts puisqu'elles se composent de 40 à 80 paires de bases. Contrairement aux oligo-nucléotides courts, leur synthèse ne s'effectue pas *in situ.* Ils sont généralement déposés par impression jet d'encre [85] (Blanchard et al., 1996). En 2001 Hughes *et al.* ont élargi cette technique initialement utilisée pour la fabrication des oligonucléotides courts à celle des oligonucléotides longs.

Les oligonucléotides longs comportent les même avantages que les oligonucléotides courts avec, en plus, la possibilité d'identifier les différents transcrits alternatifs (Barrett and Kawasaki, 2003) (Kane et al., 2000). Ils présentent moins de risques d'hybridation croisée que les oligonucléotides courts. Cependant Kane *et al.,* précisent que si un ARNm présente plus de 70% d'identité avec l'oligonucléotide long, il y a de grands risques d'hybridation croisée. La Figure 12 tirée de l'article de Barrett et Kawaski (Barrett and Kawasaki, 2003) résume les différents types de sondes disponibles et leurs différents modes d'obtention.

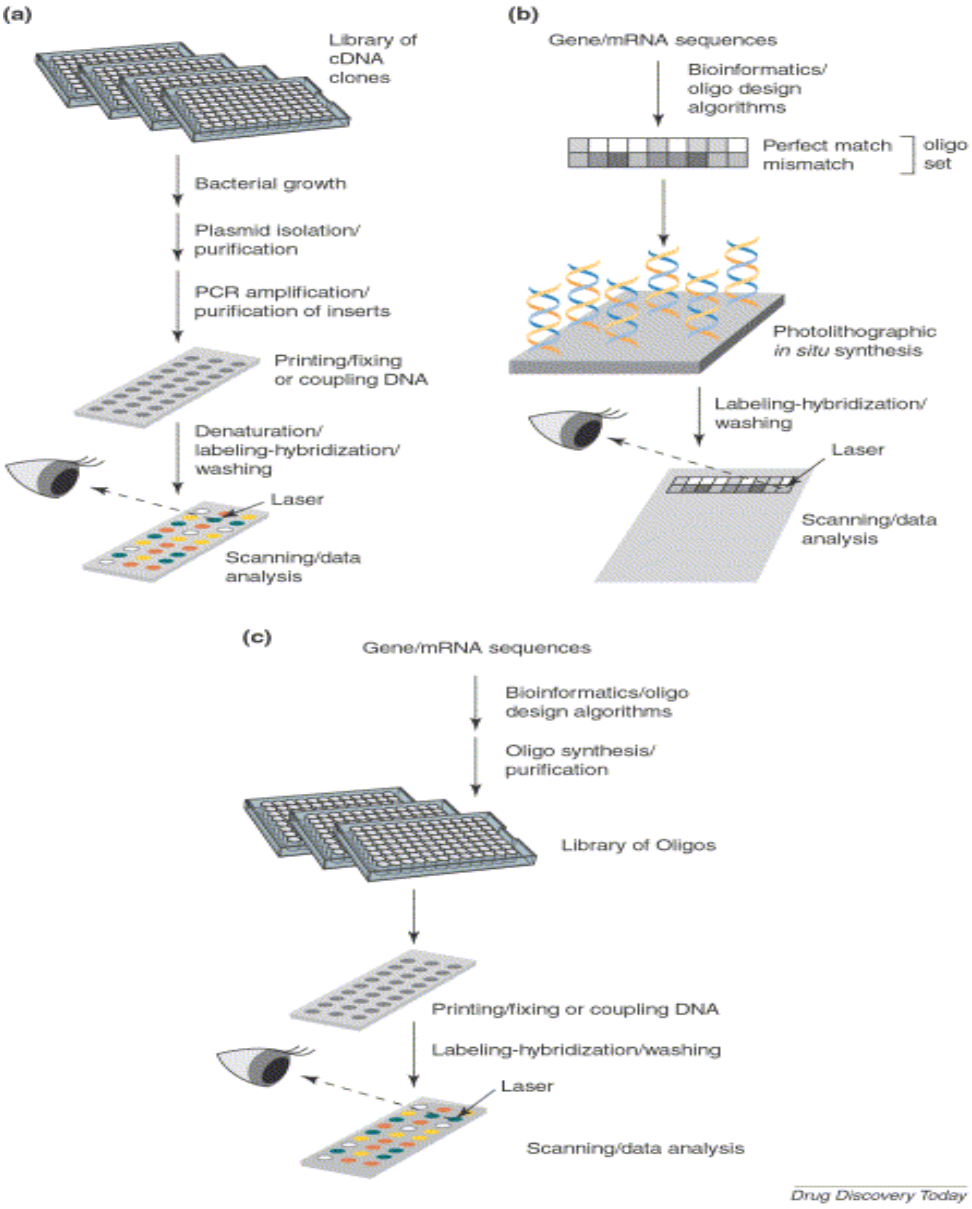


Figure 12 : Les différents types de sondes disponibles

- ✓ ***Comparaison des différentes sondes***

Le problème principal des oligonucléotides est leur conception. Il est nécessaire d'identifier des séquences spécifiques d'un gène, peu semblables à d'autres gènes, situées entièrement dans un exon, sans séquence répétitive ni la possibilité de palindrome et avec une composition en G/C constante pour toutes les sondes (Barrett and Kawasaki, 2003). Kane *et al.* ont comparé la différence de sensibilité entre des oligonucléotides longs (50 bp) et des produits de PCR (322 à 393 pb). Ces deux types de sondes présentent une sensibilité comparable. Leurs seuils de détection seraient de dix copies d'ARNm par cellules. Des études ont comparé les résultats obtenus à partir de sondes différentes. La plupart montrent que les corrélations entre les résultats obtenus avec deux sondes différentes sont relativement faibles (Kuo et al., 2002). Mais la suite de leurs résultats ne sont pas concordants. Certains précisent que les

oligonucléotides courts donnent des résultats plus fiables que les ADNc, d'autres démontrent l'inverse. Enfin Yuen *et al.* précisent que les oligonucléotides et les ADNc conduisent à une sensibilité et une spécificité équivalente.

Par ailleurs, l'incohérence des résultats se retrouve également pour le même type de sondes (ici des oligonucléotides) mais sur deux puces différentes. Nimgaonkar *et al.* ont montré que deux générations de puces Affymetrix pour l'Homme donnent des résultats relativement incohérents du fait d'un changement de densité, de sélection des oligonucléotides et d'autres changements dans la mise au point de la puce. Ces résultats ont été confirmés ultérieurement avec deux générations de puces pour l'étude du cancer (Jiang et al., 2004). Pour 25% des gènes, les résultats sont incohérents entre les deux types de puces.

b) Le type de support

Deux types de supports principaux sont actuellement utilisés : les membranes et les lames de verre. Les filtres ou membranes de nitrocellulose sont utilisées généralement avec un marquage radioactif. Ils peuvent être utilisés plusieurs fois, contrairement au support de verre. Les membranes avec marquage radioactif ont montré une plus grande sensibilité que les lame de verre avec fluorescence (Stillman and Tonkinson, 2001). Les lames de verre sont le support favori car elles ont une fluorescence résultante faible, elles sont transparentes et résistantes aux hautes températures. Comme le liquide ne peut pas pénétrer dans la lame, la surface, les sondes et les cibles sont directement en contact sans diffusion *via* des pores.

Bien sûr, cela demande d'agiter la préparation afin d'obtenir un haut taux d'hybridation (ceci est valable également pour les membranes) (Holloway et al., 2002)(Southern and Tonkinson, 2001). Contrairement aux membranes de nylon, leur relative « planité » autorise une lecture avec une précision de dizaine de micromètres. Leur rigidité permet une meilleure localisation des différents spots et plusieurs modifications chimiques de la surface sont disponibles afin de fixer les sondes (Holloway et al., 2002)(Southern and Tonkinson, 2001). Toutefois Stillman et Tonkinson ont montré que la longueur de la sonde avait une plus grande influence sur la qualité de l'hybridation que le type de support. Cependant, les lames de verre sont le support privilégié sans doute à cause de la possibilité de réaliser l'hybridation de deux échantillons à la fois et ainsi comparer directement deux conditions expérimentales.

Le type de marquage La plupart des puces actuelles utilisent un marquage fluorescent avec deux marqueurs de longueurs d'ondes d'émission différentes (une dans le vert et l'autre dans le rouge). Les cyanines fluorescentes Cy3 (vert) et Cy5 (rouge) mais aussi la fluorescéine et la rhodamine sont le plus souvent utilisées. La double fluorescence permet l'étude simultanée de deux échantillons sur la même puce. Cette possibilité explique l'engouement pour ce type de marquage puisqu'ainsi les biais entre différentes puces sont éliminés. Cependant Hoen *et al.*

ont montré que les valeurs obtenues ne sont pas influencées par la présence d'un seul ou de deux fluorochores. Ils conseillent, par ailleurs, d'utiliser les valeurs obtenues pour chacun de ces marquages plutôt que le ratio entre ces valeurs généralement utilisé. Aussi des marquages avec simple fluorescence sont également employés. Une solution alternative est le marquage radioactif au 32P ou 33P qui ne nécessite que de petites quantités d'ARN (Holloway et al., 2002). La lecture se fait à l'aide d'un *phosphorimager* couramment disponible au sein de laboratoires de biologie. Querec *et al.* ont montré que la radioactivité présente une plus grande sensibilité par rapport à la fluorescence et une reproductibilité supérieure. Cette plus grande sensibilité résulte de la nature du marquage radioactif : à des temps d'exposition suffisamment longs, l'émulsion sensible aux rayons X sera « activée ». Cependant, le temps d'exposition pour maximiser la détection des signaux faibles peut gêner la détection des gènes les plus exprimés à cause d'une saturation du signal. Les radiations excèdent alors la limite de détection du film ou du *phosphorimager.*

c) L'extraction des ARNm et le marquage

L'extraction est une des étapes clés de la mesure du transcriptome. Vu le temps de demi-vie variable des différents ARNm (renouvellement des ARNm de l'ordre d'une à deux minutes pour *E. coli* et *B. subtilis*), une extraction rapide est préférée. Il est également nécessaire de limiter le stress de l'extraction afin de limiter les perturbations du système et la synthèse d'ARNm des protéines de choc thermique. Les protocoles varient selon les organismes étudiés et leurs stades. Les mesures d'expression dépendent fortement de la technique d'extraction utilisée et de la sensibilité spécifique de chaque gène à la dégradation de son ARNm (Schuchhardt et al., 2000). Il est parfois nécessaire d'amplifier les ARNm obtenus, du fait du trop faible nombre d'exemplaires présents. Cette étape d'amplification PCR implique malheureusement des biais selon les gènes. Aussi, il est préférable de l'éviter si elle n'est pas nécessaire.

Après extraction des ARNm, il est préférable de synthétiser les ADNc correspondants, plus stables. Pour cela, il est possible, chez les eucaryotes, d'utiliser la queue polyA présente à l'extrémité 3' des ARNm pour ancrer une amorce polyT qui permet, par une transcriptase inverse, la synthèse d'ADNc marqués. Une autre possibilité lorsque la queue polyA est inexistante est la synthèse d'ADNc à parti d'un ensemble d'amorces aléatoires ou spécifiques des ARN à détecter. La longueur de l'ADNc synthétisé peut varier, ce qui influence également les niveaux d'hybridation mesurés (Schuchhardt et al., 2000). Par ailleurs, la quantité d'ADNc synthétisée n'est pas proportionnelle à la quantité d'ARN utilisée et pire, le rendement de la synthèse d'ADNc dépend d'un gène à un autre (Sekowska et al., 2001). Une des hypothèses est que, pour certains gènes, la synthèse d'ADNc peut être affectée par la

formation de structures secondaires dans l'ARNm ou par la présence de segments d'ARN largement biaisés par la composition en nucléotide (Sekowska et al., 2001).. Afin de révéler les ADN présents dans la cellule, une molécule radioactive ou fluorescente est incorporée (cf. types de puces). Cependant, l'incorporation de ces molécules peut varier selon l'ARNm. Ainsi la composition en nucléotides de l'ADNc fait varier la qualité du marquage radioactif (Schuchhardt et al., 2000).

L'hybridation La composition des sondes influence la stabilité de l'hybridation dans des conditions fixées. Ces impacts ont été relativement bien étudiés de façon à déterminer les conditions optimales d'hybridation. Le taux de GC, la structure des acides nucléiques ou la localisation de la sonde sur la séquence du gène influencent la température et les solutions nécessaires à l'hybridation. On peut se référer à l'article de Maskos et Southern pour l'étude des conditions pour les oligonucléotides (Maskos and Souther, 1992).

d) La lecture des données

La phase de lecture de données est loin d'être anodine. L'acquisition des images conditionne de façon majeure la précision des données et donc la pertinence des interprétations (Leung and Cavalieri, 2003). Suivant le type de marquage, l'acquisition des images diffèrent. Afin de détecter la fluorescence émise par la puce, les fluorophores sont excités à leur fréquence par un laser tandis qu'un scanner ou un microscope confocal couplé à un tube photomultiplicateur (PMT) permet l'analyse des photons émis par les marqueurs. Les canaux de lecture correspondant aux longueurs d'onde 635 nm et 532 nm sont utilisés pour lire la fluorescence de Cy5 et Cy3. Dans le cas d'un marquage radioactif, la radioactivité est révélée par autoradiographie grâce à une exposition d'un film à rayon X ou à un scanner *phosphorimager* qui permet de détecter la radioactivité présente pour chaque spot. Une image est alors obtenue pour chaque échantillon (2 images pour les puces avec deux fluorochromes). Les paramètres de réglages des appareils influencent grandement la qualité des images obtenues. En effet les réglages des photomultiplicateurs pour la fluorescence et le temps d'exposition pour la radioactivité influent grandement sur la sensibilité de la puce. Ainsi pour la radioactivité, si de grands temps d'exposition sont utilisés, il sera possible de détecter des ARNm présents à des niveaux très faibles. Cependant, on s'expose alors à des problèmes de saturation du film pour les gènes fortement exprimés ainsi. A cela s'ajoute le fait que l'intensité élevée d'un spot aura tendance à masquer l'intensité des spots voisins. Au contraire, des temps d'exposition courts écarteront tout problème de saturation et permettront une bonne évaluation de l'expression des gènes fortement exprimés. Cependant, les ARNm peu présents dans la cellule ne seront sans doute pas détectés. Aussi, utiliser différents temps d'exposition permet d'obtenir une mesure correcte, une grande gamme d'intensités et d'augmenter le nombre de gènes détectés

(Querec et al., 2000). Pour la fluorométrie, les scanners disposent généralement d'options diverses qui permettent d'améliorer la qualité du signal détecté : plus on augmente la tension du PMT ou la puissance du laser, plus les intensités mesurées sont fortes. Généralement, les intensités faibles sont détectées avec une précision plus faible. Il est donc tentant d'ajuster les paramètres afin d'obtenir des intensités fortes même pour les spots faiblement marqués. L'ajustement le plus couramment utilisé consiste à régler le gain du PMT de façon à ne conserver qu'un nombre minimal de spots présentant des pixel saturés (Lashkari et al., 1997). La puissance du laser est plus rarement utilisée du fait du risque de blanchissement de l'image (*photo-bleaching*) à trop fortes puissances (Leung and Cavalieri,2003). Cependant, des erreurs de mesure significatives peuvent être introduites lors de lectures à différentes tensions (Lyng et al., 2004) (Stoyanova et al., 2004) . Généralement, le signal mesuré est supposé proportionnel à l'intensité de la lumière émise à un voltage donné. Lyng *et al.* ont rappelé pour trois scanners l'influence non linéaire de la tension PMT sur l'intensité moyenne mesurée. La plupart des scanners montrent la relation log-linéaire désirée pour les moyennes d'intensités d'expression comprises entre 200 et 50 000 (Figure 13). Au dessus, il y a saturation, l'intensité croit plus lentement puis se stabilise. En dessous, l'intensité décroît rapidement avec un taux de décroissance différent selon les spots. Des problèmes de discrétisation des variables peuvent surgir pour des valeurs inférieures à 200. Lyng *et al.* [99] préconisent de faire différents scans pour des tensions différentes à partir d'une même puce en fonction de l'intensité des spots les plus faibles/forts.

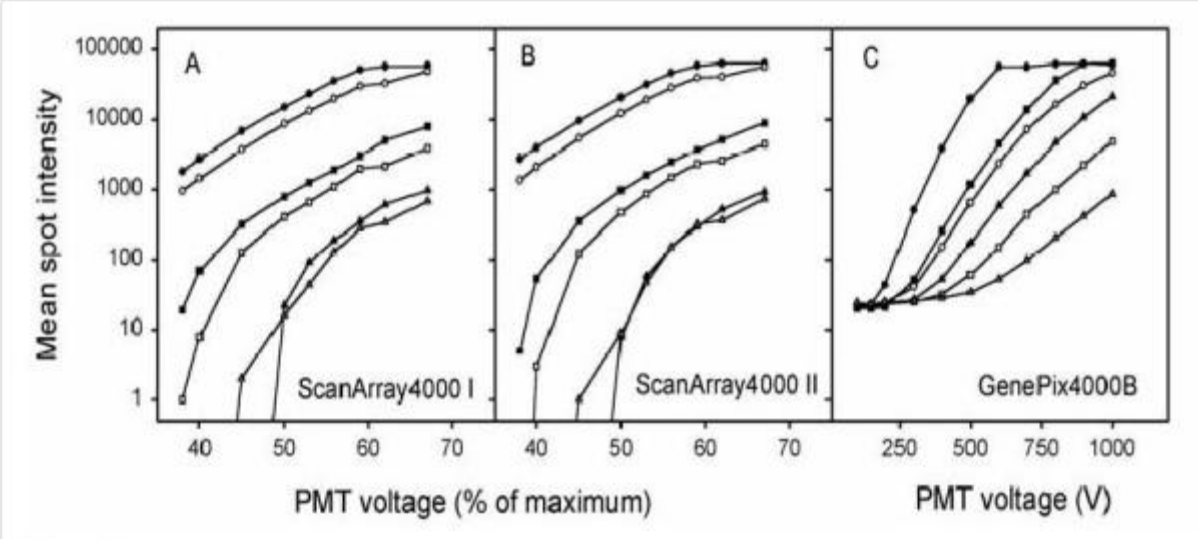


Figure 13 : Gamme de linéarité pour la mesure de niveaux d'expression (figure tirée de Lyng *et al.*)

A partir de l'image obtenue, les différents spots sont reconnus *via* un logiciel à l'aide d'une grille (Leung and Cavalieri, 2003). Les principaux logiciels utilisés sont ScanAlyze (Eisen and Brown, 1999), ImaGene et GeneSight de Biodiscovery inc, AtlasImage et AtlasNavigator de BD Biosciences Clontech et ArrayExplorer (Patriotis et al., 2001). Généralement, pour la mise en place de la grille, certains paramètres de la puce sont précisés comme la disposition de la puce, la taille et la forme des spots. Au départ, on peut croire que nous connaissons la disposition des spots et leur écartement et qu'ils sont tous circulaires avec le même diamètre. Cette vision idéale se révèle erronée: beaucoup de spots ne se trouvent pas dans la grille exactement à l'endroit attendu et certains présentent une forme non circulaire. Par conséquent, des interventions humaines sont souvent nécessaires à cette étape afin de correctement repositionner les spots ou affiner leur forme et leur taille (Hess et al., 2001). Aussi, deux chercheurs qui utilisent le même logiciel, par exemple scanAlyze, pour traiter la même image ne trouvent pas forcément des résultats identiques ou cohérents (Lawrence et al., 2004). Cela peut entraîner des variations fortes entre deux mesures, comme des ratios doublés simplement à cause du placement de la grille. L'intensité des spots est ensuite extraite. Globalement, le niveau d'expression d'un gène correspond à la moyenne ou la médiane des intensités des pixels du spot qu'il représente. Pour conclure, Stoyanova *et al.* [100] montrent que la plupart des effets non linéaires trouvés dans leur analyse correspondent à de mauvais réglages du scanner. Une attention toute particulière doit donc être apportée à ce paramètre.

2.6.5. La normalisation et la prise en compte du bruit

Le terme normalisation se réfère aux différents moyens d'éliminer l'effet des sources de variations systématiques qui affectent les mesures de transcriptome (Yang et al. ; 2002). Cette étape peut s'avérer nécessaire afin de distinguer les différences d'expression biologiques des différences liées au protocole expérimental utilisé. Les polémiques les plus importantes sur la normalisation concernent la soustraction du signal par un signal de référence ou la correction du bruit de fond ainsi que les techniques de lissage (Thygesen and Zwinderman, 2004).

a. Prise en compte du bruit

La prise en compte du bruit est généralisée dans la plupart des laboratoires. L'idée est qu'il existe une fluorescence résiduelle ou des hybridations non complémentaires dont il faut tenir compte afin d'obtenir le « véritable » niveau d'expression d'un gène. Cette prise en compte dépend du type de sondes utilisées.

La prise en compte du bruit pour les puces à oligonucléotides courts Affymetrix est réalisée à partir de la détection d'un signal sur des sondes mutées (MM). Il était au début préconisé de soustraire tout simplement le signal de la sonde mutée (MM) du signal de la sonde exacte (PM). Toutefois, cette méthode présente l'inconvénient d'obtenir de nombreux signaux

négatifs : la sonde mutée hybride quelques fois plus que la sonde exacte. Par ailleurs, cette soustraction est inadéquate pour les intensités fortes ou faibles (Chudin et al., 2002). Ainsi, pour les intensités fortes, le gène à détecter s'hybride sur les PM mais aussi sur les MM de manière non négligeable. En conséquence, les valeurs PM-MM deviennent faibles comparées à l'intensité d'expression réelle. Cet aspect du problème est décrit par Kothapalli *et al.* : parfois le signal MM masque le signal PM. Ce qui fait qu'un gène peut être considéré comme non exprimé dans une condition alors qu'il l'est après vérification par d'autres techniques. En bref, la soustraction du signal MM au PM n'est pas justifiée et ce d'autant plus que la température d'hybridation actuelle dans le protocole expérimental est plus basse que la température qui, idéalement, permettrait l'hybridation des cibles avec le PM et non le MM (Sasik et al., 2002). Pour les ADNc, généralement, l'intensité résiduelle autour de chaque spot est évaluée et soustraite à l'intensité du spot. Là encore, les données résultantes comprennent des valeurs négatives. Par ailleurs, outre les questions de la méthodologie utilisée afin de déterminer les alentours du spot, d'autres problèmes sont soulevés. Dans la discussion de leur article, Lyng *et al.* remarquent qu'en raison de l'étape d'acquisition de l'image en fluorométrie, il y a des erreurs importantes pour toutes les intensités faibles. Il en découle que la mesure du bruit est elle-même bruitée. Toutes les mesures dans lesquelles le bruit est pris en compte finissent par être biaisées à cause de l'évaluation du bruit. L'impact sera relativement faible pour les valeurs fortes mais il sera beaucoup plus important sur les valeurs faibles.

Pour être complet, il convient de se poser également la question du type de bruit de fond pris en compte. Pour les puces Affymetrix, le bruit que l'on tente de mesurer correspond à l'hybridation croisée entre des séquences homologues, il dépend donc du gène étudié. En revanche, pour les sondes ADNc, le bruit correspond à une hybridation non spécifique sur la surface du support (Vrana et al., 2003). La prise en compte du bruit n'est utile dans ce cas que si l'hybridation fluctue selon la localisation sur le support. Par ailleurs Vrana *et al.* soutiennent que ce bruit de fond n'est pas le même au sein d'un spot (présence de sondes) ou à coté d'un spot (hybridation directe sur le support). Ils suggèrent de considérer le bruit comme l'intensité du spot le plus faible correspondant à une fixation non spécifique. Cependant, la soustraction du même bruit de fond pour l'ensemble des gènes risque de se révéler inutile. Enfin, d'autres approches plus élaborées se sont développées comme une approche bayésienne qui utilise les mesures du bruit autour du spot (Kooperberg et al., 2002). Ces approches élégantes ne résolvent pas les problèmes liés au manque de précision dans la mesure de ce bruit ni la pertinence de sa prise en compte.

Pour conclure, la prise en compte du bruit est fortement sujette à débat. Ainsi, la correction du bruit basée sur la médiane de l'intensité des pixels autour du spot n'a pas été prouvée comme bénéfique (Thygesen and Zwinderman, 2004). La Figure 14 montre que la prise en compte du bruit risque de rajouter du bruit aux données initiales. Deux mesures du même échantillon marqué en radioactivité sont confrontées. A gauche, sont représentées les mesures sans prise en compte du bruit, à droite celles dont on a soustrait le bruit estimé à partir d'hybridations non spécifiques. Ce graphique démontre clairement le rajout de bruit pour les valeurs faibles en tentant de prendre en compte le problème d'hybridation non spécifique.

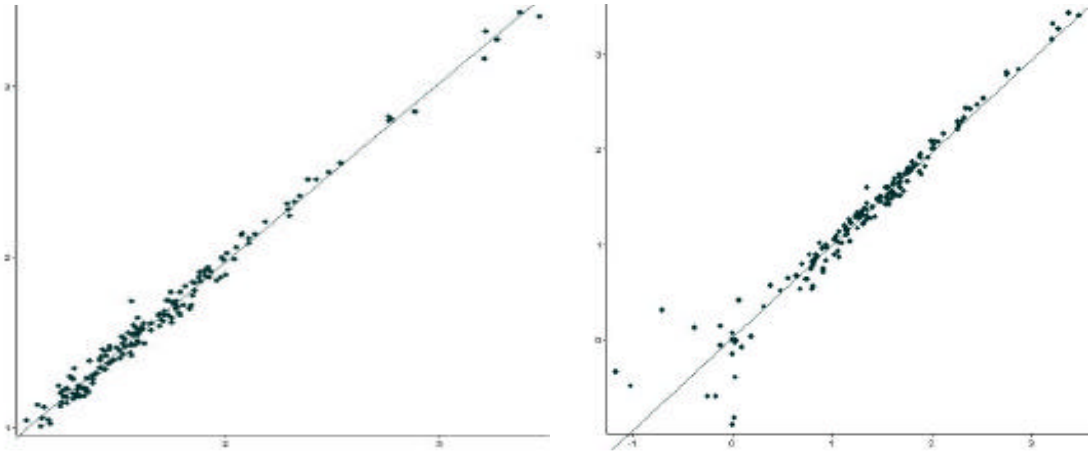


Figure 14 : La prise en compte du bruit rajoute du bruit

Les sources de variations sont multiples : de l'effet des différents fluorochromes à l'effet du réglage du scanner. Généralement, afin de pouvoir comparer les mesures effectuées sur deux puces ou avec deux fluorochromes différents, il est utile de remettre à la même échelle les mesures et d'éliminer les biais selon les intensités d'expression (Quackenbush, 2002). Plusieurs types de normalisation peuvent être distingués.

a) Différentes normalisations

✓ <u>Normalisation afin de limiter les effets de facteurs extérieurs comme le marquage et les réglages du scanner</u>

Une des premières stratégies a consisté à normaliser les données à partir des mesures de gènes de ménages (*housekeeping gene*) dont le niveau d'expression est supposé invariant quelles que soient les conditions (Vrana et al., 2003) (Whitehead andn Crawford, 2005) (Stoyanova et al., 1999). Il a été cependant révélé que leurs niveaux d'expression varient quand même pour certaines conditions expérimentales ce qui rend toute normalisation fondée sur leur expression aléatoire (Selvey et al.,2001) (Leeet al., 2002) (Kepler et al., 2002). Depuis, il existe un lot de méthodes qui identifient un jeu de gènes supposés de ménage dans les données (Stoyanova et al., 1999), (Kepler et al., 2002).. Ces méthodes présentent l'avantage de ne pas faire d'hypothèse sur les variations d'expression réelles. Elles peuvent être

intéressantes, notamment sur les puces dédiées. Cependant, elles dépendent fortement du lot de gènes identifié et requièrent un certain nombre de puces pour être valides.

Certaines expériences comparent plus de deux conditions expérimentales. Avec des puces à marquage fluorométrique, une référence marquée d'une couleur est généralement utilisée lors de chaque hybridation. La mesure du signal d'intérêt est parfois « normalisée » en soustrayant le signal de référence. Thygesen *et al.*ont analysé l'efficacité de cette soustraction. Pour leurs données elle n'apporte rien, voire pire, elle dégrade le signal initial. La plupart des méthodes de normalisation globales se fondent sur la moyenne des mesures de l'ensemble des gènes (Stoyanova et al., 1999). L'hypothèse sous-jacente suppose que l'expression de la majorité des gènes ne varie pas. Par ailleurs, les variations d'expression des gènes restants devraient s'équilibrer : autant de gènes seraient sous et sur-exprimés dans des échelles de variation proches (Yang et al., 2002) (Quackenbush, 2002). Enfin, entre deux conditions expérimentales, les niveaux moyens d'expression restent identiques (pas plus d'ARNm en moyenne). Une des normalisations les plus simples est de centrer et réduire l'ensemble des valeurs pour chaque condition. L'utilisation de la médiane et des quantiles au lieu respectivement de la moyenne et de la variance permet de se libérer de l'influence d'éventuelles valeurs extrêmes (Vrana et al., 2003)] (Stoyanova et al., 1999). Enfin Schuchhardt *et al.* ont divisé les intensités par la moyenne de chaque condition expérimentale. Dans certaines conditions, ce type de normalisation est totalement inadéquat. Ainsi, si l'étude porte sur la stabilité des ARNm au cours du temps, le niveau d'expression moyen va diminuer au fil des prélèvements du fait de la destruction des ARNm. Pour les puces dédiées à une thématique, où les gènes sont sélectionnés selon leur fonction, les hypothèses du maintien du niveau d'expression global et d'un faible nombre de gènes différentiellement exprimés peuvent également être invalidées (Quackenbush, 2002). Par ailleurs, cette normalisation globale ne prend pas en compte d'éventuels bruits qui dépendent des intensités mesurées (Kepler et al., 2002). Kerr *et al.* ont introduit l'ANOVA (*analyse of variance*) afin de normaliser les données sans les modifier *a priori*. La seule transformation réalisée avant cette normalisation est le passage au logarithme. D'autres modèles ont été également développés pour les puces à oligonucléotides (Li and W2001) (Stoyanova et al., 1999).

Afin d'observer l'effet différentiel des fluorochromes sur les niveaux d'expression mesurés, il suffit de confronter les résultats d'un même échantillon marqué par les deux fluorochromes. Les mesures observées montrent que le fluorochrome affecte les mesures selon les intensités d'expression. Ce biais résulte de différents facteurs dont les propriétés physiques des marqueurs (sensibilité à la lumière et la chaleur, temps relatif de demi-vie), l'efficacité de l'incorporation de ces marqueurs, les différences entre les techniques d'hybridation et les

paramètres du scanner (Yang et al., 2002). Même si ce biais systématique a une influence relativement faible, il peut mener à de mauvaises interprétations pour l'étude de subtiles différences biologiques [104] (Yang et al., 2002).. Différentes méthodes sont employées afin de corriger cet effet dont la méthode « lowess » (*Locally Weighted regression Scatterplot Smoothing*) ou encore des régressions locales (Vrana et al., 2003). Dans tous les cas, les hypothèses sous-jacentes sont identiques aux

méthodes globales : peu de gènes varient et leurs variations se compensent globalement (Kepler et al., 2002). Le lissage lowess repose sur l'idée que toute tendance non-linéaire entre les niveaux d'expression de deux échantillons est un artefact à éliminer (Thygesen and Zwinderman, 2004). Il est nécessaire de définir le pourcentage de points utilisés afin de lisser les données localement. Plus le pourcentage est grand, plus le lissage est important (Yang et al., 2002).Tout comme la prise en compte du bruit, il faut utiliser ce type de « correction » avec parcimonie de peur d'introduire du bruit supplémentaire. Ainsi, si l'effet non linéaire est faible, le bénéfice de lisser peut ne pas valoir son coût en terme de bruit rajouté par le lissage (Thygesen and Zwinderman, 2004). L'idéal non atteint serait de distinguer la non-linéarité artificielle des effets biologiques (Thygesen and Zwinderman, 2004) (Stoyanova et al., 2004). Différentes comparaisons entre des méthodes de normalisation ont été effectuées. Il est à noter qu'il s'agit souvent de montrer que la méthode décrite dans l'article est aussi bonne ou surclasse les méthodes couramment utilisées. La comparaison des méthodes lowess, de normalisation globale et celle développée par Zhao *et al.* sur des données simulées montre que, lorsque la majorité des gènes ne sont pas différentiellement exprimés ou que les nombres de gènes sur et sous exprimés sont égaux, la normalisation globale est la meilleure. En revanche, lorsqu'il y a beaucoup de gènes différentiellement exprimés et/ou de façon non symétrique, leur méthode est plus efficace.

Il faut néanmoins prendre ces conclusions avec prudence, puisqu'elles dépendent fortement du modèle de simulation des données ainsi que du critère de comparaison.

- ✓ <u>Normalisation afin d'adapter la distribution des données à l'analyse</u>

Certaines analyses statistiques classiques reposent sur des hypothèses de distribution de données. Par ailleurs, une distribution fortement asymétrique des données de puces (Chen et al., 2004) avec un petit nombre de valeurs fortes et beaucoup de valeurs faibles peut poser des problèmes pour l'analyse ultérieure. Une première étape, souvent négligée avant toute normalisation, est l'analyse de la distribution initiale des données. Hoyle *et al.* (Hoyle et al., 2002) ont analysé une variété de données de transcriptome de différents organismes, obtenus à partir de différentes plate-formes et supports. Ils ont montré que les données de transcriptome obtenues suivent toutes une même distribution mixte, proche de la distribution

lognormale pour la plupart des mesures ; les extrémités (queues) suivent cependant une distribution du type puissance. La variance de la distribution dépend de la taille du génome étudié. Ces observations sont valables pour toutes les puces globales, c'est à dire pour les puces qui ne présentent pas de biais de sélection des gènes comme les puces dédiées.

Comme la distribution des données est un mélange de deux distributions, il n'existera pas de transformation unique satisfaisante puisque le modèle change selon la valeur des données.

Cependant, comme la majorité des données suit une loi log-normale, la normalisation généralement utilisée est le passage à un logarithme. Cette transformation limite les effets de valeurs extrêmes sur la suite de l'analyse. La distribution obtenue est alors plus proche d'une distribution gaussienne. D'autres transformations sont possibles, comme l'utilisation de la racine-carrée. Elle ne correspond pas à la distribution des données observées. Sapir et Churchill (Sapir and Churchill, 2000) ont comparé la distribution résultante de la transformation en logarithme et en racine-carrée. Leur conclusion favorise la transformation en logarithme. De plus, cette dernière transformation a l'avantage d'avoir une interprétation biologique. Ainsi, certains phénomènes biologiques ont des effets multiplicatifs sur les niveaux d'expression. Le passage au logarithme permet un passage à des effets additifs qui sont plus facilement modélisés lors des analyses statistiques ultérieures (Kerr et al., 2000). Cependant, le logarithme pose un problème pour les valeurs négatives obtenues suite à la prise en compte du bruit (Kerr and Churchill, 2001) (Kerr et al., 2000).. Enfin, il est nécessaire de préciser que la recherche de gènes ayant le même profil d'expression peut nécessiter une étape de normalisation particulière suivant la distance utilisée. Si la distance est euclidienne, le niveau d'expression des gènes influencera beaucoup plus les résultats que le profil d'expression. Deux gènes avec de forts niveaux d'expression mais avec des profils différents se retrouveraient, sans normalisation, plus proches que deux gènes exprimés à des niveaux différents mais aux profils identiques. Quackenbush (Quackenbush, 2002) fournit une information plus détaillée sur les différents types de normalisation.

c) Les problèmes inhérents aux données utilisées.

- ✓ <u>Les valeurs faibles ou bornées.</u>

Le problème de fiabilité des valeurs faibles existe pour tous les types de puce à ADN. Les appareils de lecture sont nettement moins précis pour les valeurs faibles (Lyng et al., 2004). Du coup, les biologistes n'ont que peu de confiance dans ces petites valeurs et les traitent donc, généralement de manière différente. Elles peuvent être soit purement supprimées soit remplacées par une valeur seuil. Ainsi Tschentscher *et al.*, ont remplacé les niveaux d'expression mesurés inférieurs à 50 par la valeur 50 dans des données issues de puces Affymetrix. Cette pratique est également relativement courante dans les puces à lame de verre

avec double fluorométrie. De nombreux jeux de données comprennent, en outre, des valeurs seuils minimales artificielles ou non, des valeurs seuils maximales liées à la saturation du scanner. Lorsqu'un fort pourcentage de gènes présente des intensités seuils, toutes les analyses ultérieures se trouvent biaisées par cette distribution des données. Il arrive parfois qu'il faille réattribuer, à ces valeurs bornées, des valeurs aléatoires afin de pouvoir procéder à des analyses relativement robustes (Chiappetta et al., 2004).

- ✓ Les valeurs manquantes

Comme nous venons de le voir, un jeu de données peut comprendre des valeurs manquantes issues soit de problèmes sur la lame (poussières) soit, dans la majorité des cas, de méthodes d'acquisition d'images et de normalisation. La fréquence des valeurs manquantes n'est pas négligeable. Ainsi dans les données de Gash *et al.* 39% des gènes ont au moins une mesure d'expression manquante. On peut atteindre 72,5% pour les données de Garber *et al.* ou 73,5% pour Bohen *et al.* . Or, la plupart des méthodes d'analyse des données requièrent des jeux complets (Oba et al., 2003). Afin d'obtenir ce jeu, deux voies sont possibles : soit enlever tous les gènes présentant au moins une valeur manquante, soit réattribuer des valeurs aux données manquantes. Au vu des proportions de valeurs manquantes citées ci-dessus, il est préférable d'utiliser la deuxième voie. L'algorithme de clustering d'Eisen *et al.*présente l'alternative de ne pas tenir compte de la condition expérimentale qui présente une valeur manquante lors du calcul de distance entre l'expression de deux gènes. Toutefois, cela revient à considérer (pour une distance euclidienne) que l'expression de ces deux gènes est identique dans ces conditions. La distance entre des gènes avec des valeurs manquantes tend donc à être plus faible que les distances entre des gènes qui présentent le jeu de mesures complet (Oba et al., 2003). Dans les cas les plus extrêmes, certains clusters obtenus correspondent aux gènes dont les mesures d'expression sont incomplètes. Généralement, les chercheurs remplacent les valeurs soit par le seuil qu'ils avaient défini au préalable, soit par zéro ou plus rarement, par la moyenne des mesures pour le gène (Troyanskaya et al., 2003) (Kaski rt al., 2003) . Ces méthodes ne sont pas optimales puisque aucune ne prend en compte les corrélations existantes avec les autres gènes et que pire, les deux premières incorporent un nombre de valeurs identiques important ce qui biaise toute analyse future. Chiappetta *et al.* proposent de réassigner ces valeurs par des valeurs faibles avec un bruit aléatoire. Cependant, la réattribution des valeurs manquantes n'est pas une nouveauté propre aux données de puces à ADN. Ainsi, des méthodes classiques comme la réattribution des valeurs par régression (Zhou et al., 2003), par décomposition en valeurs propres (SVD ou singular value decomposition) (Hastie et al., 1999) ou par maximum de vraisemblance sont également utilisées. La plupart de ces méthodes calculent, à partir du plus grand jeu de données complet disponible, des

estimateurs des valeurs manquantes. Pour la SVD, les estimations sont fondées sur les valeurs propres ; pour l'algorithme des K voisins les plus proches (K nearest neigbour ou KNN) la valeur est complétée par la moyenne d'expression pour la condition manquante des K gènes les plus proches du gène à compléter (distance euclidienne sur les
valeurs d'expression présentes). Beaucoup de méthodes tendent à améliorer le KNN simple : notamment en changeant la distance utilisée, en faisant une moyenne pondérée par la distance des voisins (Troyanskaya et al., 2003), en utilisant les valeurs prédites au fur et à mesure (KNN séquentiel ou SKNN) (Kim et al., 2004). La plupart de ces méthodes nécessitent la définition de paramètres comme le nombre de valeurs propres prises en compte pour la SVD ou encore le nombre de voisins utilisés. Plusieurs études ont tenté de définir des gammes de paramètres adéquats. Pour le KNN entre 10 et 20 voisins semblent donner les meilleurs résultats (Oba et al., 2003) (Troyanskaya et al., 2003) (Kim et al., 2004).. Toutefois, cette valeur dépend du jeu de données et notamment du degré de similarité entre les profils d'expression ainsi que du pourcentage de valeurs manquantes. Pour la SVD, les meilleurs résultats obtenus se situent avec la prise en compte d'environ 20% des valeurs propres (Troyanskaya et al., 2003).

Plusieurs comparaisons de l'efficacité de ces méthodes ont été effectuées. La plupart reposent sur un jeu complet de données existantes, auquel un certain pourcentage de valeurs est supprimé aléatoirement. Plus une méthode procure des estimations proche des valeurs réelles, plus elle est caractérisée comme performante (généralement fondée sur le *RMSE root mean square error*). Les méthodes par maximum de vraisemblance présentent des performances bien moindres que le KNN et le SKNN (Kim et al., 2004).. Pour un pourcentage de valeurs manquantes situé entre 1 et 20%, Troyanskaya *et al.* ont montré que le KNN avec la moyenne pondérée est plus efficace que le SVD et bien plus efficace que la méthode d'attribution de la moyenne du gène. La SVD est plus sensible au type de jeu de données utilisé que le KNN. Pour des jeux de données où plus de 30¨% des gènes présentent au moins une valeur manquante, le SKNN se révèle meilleur que le KNN (Kim et al., 2004).. Pour un pourcentage de valeurs manquantes égal à 5% l'ACP bayésienne est plus performante que le KNN ou le SVD (Oba et al., 2003).

La question est de savoir si ces résultats sont extrapolables sur un jeu de données réelles. Un premier point est de définir la taille du jeu de données minimale afin d'avoir des estimations correctes des valeurs manquantes. Le KNN donne des résultats corrects pour un nombre de conditions expérimentales supérieur à 6. En dessous, les estimations sont moins fiables voire mauvaises en dessous de 4 (Troyanskaya et al., 2003).. Comme nous l'avons précisé plus haut, la plupart des données manquantes ne sont pas dues à des poussières sur la puce mais

plutôt à un manque de confiance dans les valeurs faibles mesurées. Or, la plupart de ces comparaisons font l'hypothèse de la répartition aléatoire des valeurs manquantes, indépendantes du niveau d'expression du gène. L'efficacité des estimations des valeurs manquantes devrait être moindre dans les conditions réelles (Oba et al., 2003)..

Aussi, la meilleure solution est encore de limiter le nombre de données manquantes, notamment en réduisant ou en éliminant les seuils de détection définis *a priori*. Ainsi, pour toute analyse, le traitement des données doit être réduit au minimum avant analyse : par exemple, on ne devrait pas enlever les bruit de fond des valeurs ou utiliser des ratios de valeurs car cela ne peut être réalisé sans introduire de biais spécifiques. Similairement, l'introduction de normalisation fondée sur des hypothèses biologiques (moyenne sur des gènes de ménage) introduit un biais systématique (Sekowska et al., 2001). Enfin, dans les expériences d'étude du profil d'expression de l'ensemble des gènes, les variations d'intérêt sont souvent cachées par le bruit (Sekowska et al., 2001).. L'interprétation des expériences est gênée par la précision des mesures physiques, de telle manière que l'on néglige souvent l'importance de fluctuations inhérentes aux expériences biologiques. La normalisation des données ne doit pas prendre le pas sur une analyse approfondie qui tient compte des différents facteurs biologiques interférant.

- ✓ <u>Précisions sur les données de puces</u>

Les intensités des spots ne peuvent pas être prises comme une estimation précise du niveau d'expression mais plutôt comme une mesure relative à l'ensemble des mesures de niveaux d'expression dans l'échantillon étudié (Querec et al., 2004).

Les données de transcriptome présentent peu de mesures, souvent quelques dizaines de conditions expérimentales, comparées au grand nombre de variables étudiées, les niveaux d'expression de milliers ou de dizaines de milliers de gènes (Vrana et al., 2003). Par ailleurs, les réplications sont en nombre limité ce qui rend difficile l'évaluation des erreurs et donc, un nombre de faux négatifs et positifs non négligeable (Vrana et al., 2003).. Les niveaux d'expression des gènes ne sont pas indépendants les uns des autres puisqu'ils peuvent participer à un même réseau de régulation ou à une même voie métabolique. A cela se rajoutent des mesures relativement peu reproductibles en raison des différentes sources de variations décrites précédemment.

- ✓ <u>La confrontation à des données externes</u>

En 1997, une étude sur les effets du stress lié à la température (chaud ou froid) (Lashkari et al., 1997) montre que si une minorité de gènes identifiés comme différentiellement exprimés correspondent aux gènes attendus, la majorité ne devrait pas être *a priori* différentiellement exprimée. Afin de savoir si les différences d'expression mesurées correspondent réellement à

un changement physiologique, il est nécessaire de prendre en compte des données externes (autres techniques de mesure ou connaissances préalables).

Les résultats obtenus doivent être confrontés à d'autres sources de données avant de se lancer dans des recherches futures consommatrices de temps et d'argent. En effet, les effets observés par l'analyse des puces à ADN ne correspondent pas forcément à l'effet direct escompté. Par exemple, dans une étude de l'influence de mutants de la DAM méthylase sur l'expression des gènes, on s'attendrait à voir des gènes qui présentent des clusters de GATC à l'intérieur de leur séquence ou dans leur région promotrice/régulatrice (Riva et al., 2004). Les résultats obtenus ne présentent aucune corrélation entre les gènes différentiellement exprimés et la présence ou non de GATC. L'expérience de transcriptome ne permet donc pas de conclure. Les gènes identifiés peuvent être différentiellement exprimés à cause de l'observation d'effets indirects comme l'activation de voies métaboliques annexes. Ils peuvent également être identifiés comme différentiellement exprimés en raison de changements de conditions expérimentales incontrôlés. Parfois, des phénomènes transitoires au début de l'expérience produisent une empreinte sur la culture pour toute l'expérience et se reflètent dans la mesure des niveaux d'expression (Sekowska et al., 2001). Cette étape, fortement négligée au début de l'engouement pour les puces à ADN, est désormais obligatoire avant toute publication de résultats. Kothapalli *et al.* ont montré que les résultats des puces étaient fortement dépendants de la technique utilisée. Parmi dix-sept gènes supposés différentiellement exprimés, seulement huit (47%) ont été confirmés par Northern blot. Aussi, il s'avère souvent nécessaire, avant de publier ou de se lancer dans une analyse plus précise des gènes détectés, de confronter les résultats des puces avec d'autres sources de données. Une des possibilités consiste à vérifier la sur ou sous expression du gène d'intérêt à l'aide de mesures par Northern blot ou RT-PCR quantitative (Querec et al., 2004).

Par ailleurs, les conditions expérimentales, jamais totalement maîtrisées, peuvent révéler le déclenchement de voies métaboliques indépendantes de l'objet de l'étude. La confrontation des données avec les connaissances existantes et notamment les bases de données fonctionnelles permet de valider certains résultats. Des études couplent également analyse du transcriptome et du métabolome chez *Arabidopsis thaliana* (Hirai et al., 2004).

2.6.6. L'analyse des puces

Cette partie présentera de manière non exhaustive les différents types de méthodes d'analyse de données. Tout comme les méthodes de normalisation, de nombreuses publications portent sur l'analyse des puces avec la création d'algorithmes spécifiques à ces données. A cela se rajoutent les méthodes de statistiques issues d'autres domaines, désormais appliquées au

transcriptome. Vu le nombre important de publications à ce sujet, nous aborderons certaines grandes classes de méthodes. Pour une vue plus exhaustive, on peut se référer au site de Li (http://www.nslijgenetics. org/microarray/) qui recense environ 1300 publications sur ce sujet depuis 1993. La première remarque à faire sur l'analyse des puces est qu'il n'existe, à ce jour, aucun consensus sur la méthode à utiliser (Vrana et al., 2003). En effet, aucune méthode d'analyse ne permet d'aborder l'ensemble de l'histoire présente dans les données de puces mais uniquement un chapitre de celle-ci (Leung, 2002). Il est donc nécessaire d'utiliser différentes méthodes plutôt qu'une seule si l'on souhaite tirer le maximum d'information des données disponibles (Quackenbush, 2001).

L'étape de l'analyse est une étape appréhendée par les biologistes, car elle est considérée comme la plus décourageante (Vrana et al., 2003). Souvent, la crainte de ne pas maîtriser les méthodes statistiques entraîne l'utilisation de méthodes d'abord plus facile comme l'utilisation du ratio et la définition d'un seuil audessus duquel un gène est défini comme différentiellement exprimé. C'est d'ailleurs la méthode la plus utilisée dans le monde du transcriptome. Le recours à des méthodes statistiques est cependant conseillé. Pour cela, la collaboration avec des experts de l'analyse de données est parfois nécessaire mais avant tout, rien ne peut remplacer une formation en analyse afin de mettre correctement en place les expériences (Leung, 2002).

a) Principes de l'analyse du transcriptome

Les puces à ADN servent généralement à étudier une voie métabolique particulière et à identifier les gènes impliqués dans un phénomène. Or, toute perturbation sur un gène particulier ou sur un composant du signal se propage rapidement à travers le réseau métabolique (Sontag et al., 2004). Il est en pratique impossible de mettre en oeuvre une expérience afin d'observer comment un changement à un nœud du réseau métabolique affectera un autre noeud. En effet, les interconnexions causeront des changements globaux du réseau. Le problème est d'utiliser ces perturbations globales pour retrouver les interactions entre des noeuds individuels. Le but de l'analyse du transcriptome n'est pas forcément de détecter les gènes avec de grandes variations d'expression, facilement repérables sans statistique. Il faut également rechercher les gènes qui ont des petites variations **reproductibles** (Kerr and Churchill, 2001). La reproductibilité des résultats est nécessaire avant de se lancer dans des expériences d'approfondissement des conclusions sous peine de perdre du temps et de l'argent dans des efforts non fondés. Dans ce cas, les scientifiques ont besoin d'un plan d'expérience et de méthodes d'analyse qui ne sont pas fondées uniquement sur l'étude relative des mesures mais qui permettent également une estimation des erreurs. Globalement, l'analyse de puces permet de détecter un grand nombre de gènes comme

différentiellement exprimés. Plusieurs articles détectent qu'environ 10% des gènes se trouvent différentiellement exprimés dans les puces, par exemple lors de l'étude du rythme circadien d'*Arabidopsis thaliana* (Davis and Millar, 2001) ou de l'influence de mutants GATC chez *E. coli* (Riva et al., 2004). Il semble que ce pourcentage de gènes différentiellement exprimés soit une propriété généralement observée dans les études du transcriptome. L'analyse des puces à ADN comporte deux volets : la recherche de gènes différentiellement exprimés et la recherche de profils d'expression communs. Chaque volet possède des méthodes d'analyse spécifiques.

b) Les ratios et le principe de seuillage

Dans le reste de ce manuscrit, nous ne traiterons pas des données fournies sous la forme de ratio. Beaucoup d'articles présentent des résultats issus de cette méthode non statistique mais intuitive. Il s'agit de l'utilisation simple du logarithme du ratio entre les mesures de deux conditions expérimentales. Au départ, le ratio a été utilisé pour le marquage par fluorescence avec un échantillon hybridé en vert et l'autre en rouge. Le ratio est désormais appliqué à des marquages différents et même des échantillons hybridés sur des puces différentes. L'idée est que si un gène est différentiellement exprimé entre les deux conditions, son ratio devrait s'écarter fortement de 1. La définition d'un seuil défini arbitrairement, généralement de 3 ou 1/3, permet de définir respectivement les gènes sur/ sous -exprimés dans une condition expérimentale (DeRisi et al., 1997). Elle présente l'avantage d'être intuitive avec une interprétation biologique immédiate : le gène s'exprime x fois plus dans cette condition que dans l'autre. Cependant, cette interprétation est erronée puisque la valeur mesurée comprend une valeur relative de l'intensité d'expression mais aussi des biais liés au protocole expérimental (Kerr et al., 2000). Les ratios requièrent donc des étapes de normalisation. Les ratios obtenus dépendent alors de la méthode utilisée. Si elle est aisément compréhensible, elle ne permet pas d'appréhender la complexité des données obtenues. Vu les biais décrits dans le chapitre précédent, les résultats de comparaison d'uniquement deux conditions expérimentales sont nettement insuffisants pour conclure à l'implication de gènes dans le phénomène étudié. Ainsi, si plus de deux conditions expérimentales sont comparées, la méthode ratio seuillage ne suffit plus. Cette méthode ne permet pas d'utiliser l'ensemble des données disponibles et notamment l'utilisation au mieux des éventuels réplicats. Dans le cas où il y aurait réplication, le ratio correspond généralement au ratio des moyennes des mesures. Par ailleurs, la valeur du ratio est fortement influencée par l'intensité globale de l'expression du gène. On accorde moins de confiance pour de faibles valeurs pour lesquelles le ratio varie plus rapidement (Newton et al., 2001). *A contrario,* pour les niveaux d'expression les plus hauts, de petits changements (petits ratios)

peuvent être réels mais ils seront rejetés par le seuil défini (Tusher et al., 2001). Comme une très grande majorité des gènes s'expriment à des niveaux faibles, les ratios d'expression sont aléatoires pour une grande partie de ces gènes. Comme le ratio est moins fiable pour les faibles valeurs, certaines analyses comprennent la suppression des valeurs faibles jugées non pertinentes (Quackenbush, 2002). On peut ainsi trouver des articles où l'on supprime les valeurs faibles pour ensuite les estimer. Les ratios sont également utilisés comme mesure de base dans beaucoup de méthodes statistiques appliquées au transcriptome. Si la distribution des intensités d'expression dans une condition peut se rapprocher d'une loi normale, cela n'est pas le cas avec les ratios dont la distribution est plus difficilement modélisable. Cette utilisation des ratios est donc fortement sujette à caution. Dans la suite de ce manuscrit, les données étudiées sont des niveaux d'expression mesurés pour une condition expérimentale.

c) Détection de gènes différentiellement exprimés

Généralement, on cherche à détecter les gènes différentiellement exprimés selon le facteur d'intérêt (2 ou plus modalités). Parfois, le plan d'expérience comprend, en plus du facteur d'intérêt, d'autres facteurs biologiques (ex. populations) ou expérimentaux (ex. protocole d'extraction). Dans ce cas, le but est de rechercher des gènes dont le niveau d'expression varie selon le facteur d'intérêt indépendamment des autres facteurs pris en compte.

Il existe de nombreuses méthodes qui servent à l'identification de gènes différentiellement exprimés. L'analyse repose sur les valeurs d'expression d'un gène dans les différentes conditions. Un critère de sélection, spécifique à la méthode employée, permet de déterminer si le gène est différentiellement exprimé ou si la valeur mesurée peut être obtenue par hasard. Les gènes sont étudiés indépendamment les uns des autres, ce qui entraîne un grand nombre de tests par analyse. Deux problématiques se posent : d'une part, déterminer la distribution des valeurs du critère si les échantillons sont répartis aléatoirement et, d'autre part, prendre en compte le fait que de nombreux tests multiples sont réalisés.

d) Problématique de la distribution aléatoire des valeurs du critère étudié

Afin de déterminer si un gène est différentiellement exprimé, la valeur du critère mesuré est comparée à la distribution des valeurs du critère lorsqu'il n'y a pas de différence d'expression. Avant l'utilisation massive des calculs *via* ordinateurs, cette distribution était approchée par des lois classiques comme la loi gaussienne ou la loi de Student. Certaines analyses reposent encore sur ce principe d'une loi *a priori* suivie par le critère de sélection. Cependant, les données de transcriptome s'éloignent généralement des hypothèses formulées (notamment des hypothèses de normalité des résidus). Beaucoup de méthodes d'analyse adaptées au transcriptome simulent la distribution des valeurs du critère de sélection *via* un rééchantillonage aléatoire des données existantes. Les méthodes diffèrent selon les groupes de

gènes utilisés pour l'estimation de la distribution. Certains permutent aléatoirement l'ensemble des données tandis que d'autres effectuent des permutations pondérées par la possibilité que ces gènes soient des *outliers* (Kutalik et al., 2004).

e) Problématique des tests multiples

L'analyse du transcriptome porte sur l'étude des niveaux d'expression de milliers de gènes simultanément. Pour chacun de ces gènes, on étudie la probabilité qu'il soit différentiellement exprimé. On fait appel aux statistiques pour répondre à la question : les différences d'expression observées sont-elles bien réelles ? La réponse est indirecte : les statistiques donnent la probabilité qu'il s'agisse d'un faux-positif. Un faux-positif correspond au cas d'un gène où le critère de sélection observé dépasse, par hasard, un seuil fixé à l'avance. Comme une expérience de transcriptome porte sur des milliers de gènes simultanément, l'analyse statistique est utilisée pour évaluer le pourcentage probable de faux-positifs au-delà d'un seuil donné : 40 gènes dépassent le seuil 1 % par hasard si l'expérience porte sur 4000 gènes alors que c'est le cas de 400 si elle porte sur 40 000 gènes. L'estimation du nombre de faux-positifs n'est qu'une première étape dans le raisonnement. En effet, on trouve, au-delà du seuil, des faux-positifs et des gènes pour lesquels la différence observée est bien réelle (on la retrouverait dans une autre expérience). L'information clé est la proportion de faux positifs dans l'ensemble des gènes identifiés comme différentiellement exprimés, FDR ou *False discovery rate* (Benjamini et al., 2001) car elle mesure le risque de se lancer dans une fausse piste si on décide de travailler sur un gène pris dans cet ensemble. Habituellement le seuil est choisi afin d'avoir moins de 5 % de faux-positifs dans le lot de gènes sélectionnés. Par exemple, prenons une expérience portant sur 4000 gènes et dont 80 gènes sont au-delà du seuil 0,1 %. Comme il y a en moyenne 4 faux positifs au-delà du seuil 0,1 % (4000 0,001), le pourcentage de faux positifs est de 4 / 80, soit 5 % des gènes sélectionnés.

f) <u>Description succincte de quelques méthodes</u>

SAM (significance analysis of microarrays) (Tusher et al., 2001) identifie des changements d'expression statistiquement significatifs de manière similaire à un test-t pour chaque gène. L'estimation de la distribution de l'estimateur est générée grâce à des permutations aléatoires des données. D'autres méthodes comme l'ANOVA sont employées de manière classique. Toutefois, comme la distribution de l'erreur obtenue s'écarte généralement fortement de la normalité (Kerr and Churchill, 2001), les probabilités peuvent également être estimées à l'aide de permutations des données.

g) Détection de profils d'expression semblables

Les méthodes de classification ou de *clustering* visent à identifier des groupes de gènes. L'hypothèse sous jacente est que des gènes co-régulés ou impliqués dans la même voie métabolique doivent avoir leurs niveaux d'expression corrélés.

h) Problématique du manque de mesures

La plupart des méthodes de classification renvoie systématiquement des groupes de gènes, quelles que soient les données fournies indépendamment de la pertinence biologique (Yeung et al., 2004). Or, les données de puces sont relativement bruitées à cause des erreurs de mesure et des variations techniques. Toute classification trouvera des profils d'expression dans le bruit autant que dans le signal. La pertinence d'une règle établie à partir des mesures obtenues dépend du rapport nombre d'échantillons par rapport au nombre de caractères étudiés (ici gène). Les résultats sont généralement estimés robustes pour un rapport d'au moins 5-10 (en fonction des données et de la complexité du classificateur) (Somorjai et al., 2003). Dans le cas des puces, le rapport est typiquement situé en 1/20 (cas rare où il y a beaucoup de conditions expérimentales) et 1/500. Les classifications obtenues ne seront sans doute pas robustes.

Ce problème est d'autant plus important que les résultats obtenus sont la base de diagnostics ou d'explorations plus poussées. Des comparaisons de méthodes de classification ont montré une caractéristique intuitive de la classification : plus le nombre de conditions expérimentales augmente plus les groupes sont significatifs et présentent des gènes co-régulés (Yeung et al., 2004). A partir de cinquante conditions expérimentales, les groupes sont relativement fiables. Il faut donc prendre garde aux conclusions réalisées à partir de groupes de gènes obtenus avec peu de données. Pour l'identification de gènes de diagnostic, il faut garder à l'esprit que, lorsqu'il y a un nombre réduit d'échantillons, il est facile d'identifier des classificateurs robustes qui donnent de bons résultats à la fois sur le jeu de données initial et celui de validation. Mais les résultats issus de ces jeux sont illusoires et les conclusions sont suspectes voire fausses (Somorjai et al., 2003). Les jeux de données doivent être suffisamment grands pour être représentatifs de la distribution de la population à étudier par la suite. Afin d'obtenir une classification pertinente, il est donc nécessaire d'obtenir le plus grand nombre de mesures possibles avec le problème de coût que cela implique et généralement de diminuer le nombre de gènes à classer. La solution conventionnelle est de réduire l'espace des variables/caractères en éliminant les informations redondantes ou les éléments bruités. Dans notre cas, il est possible d'éliminer les gènes dont les niveaux d'expression ne varient pas ou très peu lors des conditions expérimentales (Sapir and Churchill, 2000). Ce préalable est cependant souvent négligé. Une des idées proches de ce problème est d'identifier jusqu'à quelle profondeur un

arbre de classification est fiable. Cette profondeur de fiabilité dépend, contrairement au problème précédent, de la méthode de classification utilisée. Ainsi, la classification hiérarchique donne des résultats fiables pour des groupes de moins de dix gènes (Kaski et al., 2003). D'autres méthodes permettent d'obtenir des résultats fiables pour des groupes constitués de moins de cinquante gènes, rarement plus. Certaines études présentent une interprétation des résultats fondée sur une classification globale de l'ensemble des gènes. Les conclusions qui en découlent ne sont généralement pas fiables.

i) Problématique de la définition de similarité ou distance entre les gènes.

Une autre difficulté est de présenter la similarité entre les gènes (qui correspond à un espace d'une dizaine à plusieurs dizaines de dimensions) sur un espace à deux voire trois dimensions. Chaque méthode nécessite forcément des compromis (Kaski et al., 2003) et chaque méthode explorera le nuage des données de manière particulière. Cette exploration du nuage dépend de la distance entre deux gènes utilisée : distance euclidienne, angle, corrélation linéaire, … D'autres mesures de similarités ont été développées. Elles prennent en compte un phénomène d'apprentissage entre deux jeux de données : des données de transcriptome et des données auxiliaires comme des classes fonctionnelles ou des lieux d'expression des gènes (Kaski et al., 2003). La distance la plus utilisée est la distance euclidienne. En présence d'une normalisation sur le niveau d'expression de chaque gène, elle correspond à la corrélation linéaire. Dans le cas contraire, elle mesure à la fois la différence de niveau d'expression et le profil d'expression. Aussi, elle conduit généralement au rassemblement des gènes de même niveaux d'expression. Le coefficient de Pearson permet de mettre en évidence une similarité de profils d'expression sans tenir compte des niveaux moyens d'expression (Yeung et al., 2004) Yeung *et al.* ont comparé le nombre de gènes qui présentent le même facteur de régulation au sein des clusters obtenus. Quelle que soit la méthode utilisée, la corrélation offrait de meilleurs résultats que la distance euclidienne sans normalisation préalable.

2.6.7. Différents types de méthodes de classification

En plus de multiples possibilités de mesures de distance, il est possible de distinguer deux types de classification : la classification supervisée où l'on connaît déjà les conditions discriminatoires (par exemple différents types de cancer) et la classification non supervisée qui présente une démarche plus exploratoire.

Dans la classification supervisée, le but est d'obtenir :

- des groupes robustes (indépendants d'éventuel outliers) tels que les nouveaux échantillons soient classés correctement
- un nombre réduit de gènes discriminants.

Le but final est généralement de définir un jeu de gènes marqueurs d'un type d'échantillon (type de cancer) avec un nombre de gènes limités pour pouvoir réaliser une puce dédiée au diagnostic mais efficace pour le diagnostic. Malheureusement, il est assez difficile d'avoir ces deux critères réunis (Somorjai et al., 2003). Généralement, les gènes marqueurs sont identifiés indépendamment les uns des autres : les M meilleurs sont sélectionnés. Ces M gènes peuvent receler des informations communes et donc inutiles (Somorjai et al., 2003). Une autre solution est d'identifier un nombre important de jeux de gènes ayant un pouvoir de classification et, ensuite, de compter l'occurrence de chaque gène dans ces jeux et de prendre les plus fréquents. La classification non supervisée n'a pas d'information *a priori* sur les groupes de gènes à identifier. La classification peut être ascendante ou descendante. Certaines méthodes requièrent, cependant, le nombre de clusters à identifier comme les méthodes de K-means ou les SOM (*self organizing maps*) alors que d'autres produisent des arbres de classification sans définition de nombre de groupes. Une des méthodes les plus utilisées sans définir le nombre de groupe *a priori* est la classification hiérarchique qui rassemble les gènes les plus proches en un groupe, calcule les niveaux d'expression représentant ce groupe et poursuit ensuite la classification (Yeung et al., 2004). La méthode la plus utilisée est la méthode graphique de Eisen. Son interface graphique et sa facilité d'utilisation ont encouragé son utilisation. Le graphique comprend, d'un côté l'arbre de classification des gènes obtenu par classification hiérarchique avec comme distance la corrélation linéaire et, de l'autre le graphique des résultats obtenus sous forme de couleur rouge, vert ou jaune selon les conditions expérimentales (Figure 15). Les groupes sont définis en coupant arbitrairement à un niveau de l'arbre.

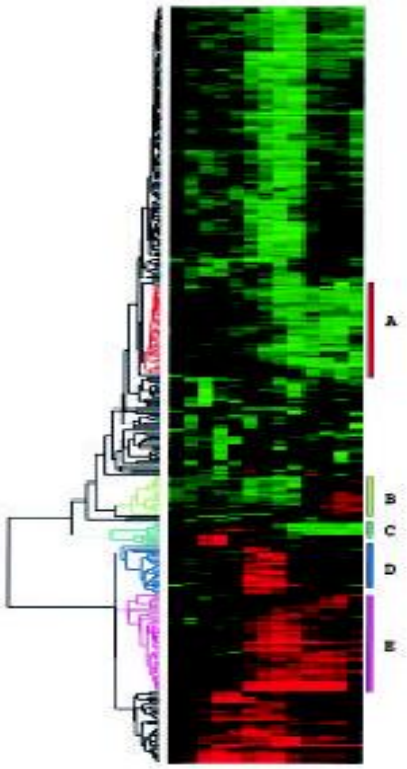


Figure 15 : Exemple de graphique de Eisen (tiré de (Eisen et al., 1998)**)**

Eisen *et al*. reconnaissent, à la fin de leur article, que la méthode de classification utilisée n'est pas forcément la plus adéquate. Depuis, de nombreux articles ont été publiés présentant des méthodes généralement importées d'autres domaines scientifiques. Ces méthodes vont de la plus simple comme la classification hiérarchique à la plus sophistiquée. Comme l'ont fait remarquer Somorjai *et al.,* la maxime « le plus simple est le mieux » a été la plupart du temps ignorée. En général, aucun effort n'a été réalisé afin de choisir le classificateur le plus approprié selon le type de jeu de données. Le choix s'effectue soit en prenant l'exemple d'un autre article, l'expérience et les préférences personnelles ou la disponibilité du logiciel. Et pourtant la complexité du classificateur et la taille de l'échantillon doivent être corrélées.

2.6.8. Problèmes de méthodes de classification et méthodes exploratoires

Un des problèmes des méthodes de classification est que chaque gène sélectionné est présent dans la classification finale (même si son profil d'expression est relativement éloigné des autres). Par ailleurs, il n'est présent que dans un seul groupe de gènes (Martoglio et al., 2002). Or, biologiquement, un même gène peut participer à plusieurs voies métaboliques ou de régulation définis dans différents clusters. On s'attend à ce que chaque gène soit influencé par plusieurs facteurs de transcription et qu'il puisse influencer différents gènes (Chiappetta et al., 2004).

D'autres méthodes sont donc utilisées afin d'identifier des groupes de gènes dont les profils d'expression sont proches : ce sont des méthodes généralement utilisées afin de visualiser le nuage des données. Ainsi, l'analyse en composante principale (ACP) (Alter et al., 2000) ou

l'analyse en composantes indépendantes (ACI) ont donné des résultats intéressants dans le cadre du transcriptome (Chiappetta et al., 2004). Ces deux méthodes recherchent des 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 et l'ACI des axes statistiquement indépendants sur lesquels les données s'écartent le plus de la loi normale qui caractérise généralement le bruit. Ces axes sont supposés représenter des processus biologiques indépendants. Comme ces méthodes ne sont pas dirigées, les axes ne représentent pas forcément les facteurs d'intérêt. Toutefois, dans notre expérience et dans les différents articles qui comprennent cette méthode, un axe ou un plan représentant les facteurs d'intérêt de l'expérience ont toujours été trouvés. Pour l'ACP, le premier axe représente toujours les niveaux d'expression relatifs de chaque gène (Alter et al., 2000), soit entre 80 et 90% de la variance des données. Cette proportion est d'autant plus forte que l'expérience porte sur de petites fluctuations des niveaux d'expression avec peu de gènes impliqués. L'ACP est également une méthode dédiée à l'étude de phénomènes temporels.

Ces méthodes peuvent également être utilisées pour réduire l'espace des données. Il suffit pour cela de filtrer les axes qui rassemblent du bruit ou des artefacts et de garder les axes qui contiendraient potentiellement l'information biologique. Il suffit d'appliquer ensuite des méthodes de classification ou, tout simplement, de visualiser les groupes de gènes discriminés par les axes restants. L'article qui suit complète cette approche bibliographique et détaille notamment cinq méthodes d'analyse utilisées sur les données de transcriptome : l'ACP, l'ACI, le t-test, l'ANOVA et SOM. Il présente également différentes questions souvent posées par les biologistes lors des formations permanentes auxquelles nous avons participé. Enfin, il souligne le fait qu'aucune méthode de normalisation et d'analyse ne peut compenser la nécessité d'un plan expérimental bien établi. Pour conclure cette partie, il ne faut pas perdre de vue que chaque méthode d'analyse adopte un point de vue particulier sur le nuage. Les différentes méthodes présenteront donc, généralement, des résultats différents mais complémentaires. Le Tableau 3, tiré d'une présentation réalisée à l'ASMDA (Applied stochastic models and data analysis) de Brest et détaillée dans le chapitre suivant illustre bien ce propos. Il présente les différents opérons détectés par des méthodes sur une expérience de transcriptome sur *B. subtilis*. L'étude porte sur les gènes activés par la mise en présence de différentes sources de soufre, le méthyl-thioribose ou la méthionine. Si quatre opérons sont détectés par l'ensemble des méthodes utilisées, trois ne sont pas détectés par une des méthodes, trois sont détectés par deux méthodes et six sont spécifiques d'une méthode d'analyse Il peut s'avérer intéressant d'employer différentes méthodes d'analyse des données afin d'aborder une vue un peu plus globale des résultats.

Tableau 3 : Comparaison de résultats obtenus par différentes méthodes

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
yqiXYZ	3	1	1	4	3	6
argCJBD carAB argF	7	15	28	29	201	56
argGH ytzD	3	1	1	6	6	2
ahpCF	2	46	7	85	11	13
lctEP	2	26			36	8
levDEFG sacC	5	316	220	287		
sunAT yolIJK	5		634			13
ydcPQRST yddABCDEFGHIJ	15				1313	116
ytmIJKLM hisP ytmO ytnIJ ribR hipO ytnM	12			45	92	
flgM yvyG flgKL yviEF csrA hag	8					509
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19					350
yxbBA yxnB asnH yxaM	5				15	
yvrPONM	4			494		
ycbCD	2			40		
comGABCDEFG yqzE	8			49		
Relevant detected operons		6	6	9	7	9

2.7. Le choix de la méthode adéquate pour identifier des gènes différentiellement exprimés : un critère biologique

L'engouement pour le transcriptome a conduit à un développement très important de méthodes spécifiques ou non du transcriptome. Chaque année, plusieurs centaines de publications portent sur l'adaptation ou la création de nouvelles méthodes d'analyse (cf. http://www.nslijgenetics.org/microarray/). Face à cette masse et cette diversité de méthodologie, les biologistes se retrouvent souvent démunis et ont des difficultés à choisir la méthode la plus adéquate pour traiter leurs données. En effet, aucune méthode n'a obtenu jusqu'à présent le consensus général et il n'existe donc pas de protocole idéal d'analyse. La nécessité de comparer l'efficacité des différentes méthodologies disponibles paraît évidente. Différents critères de comparaison ont déjà été appliqués dans le cadre de l'étude du transcriptome. Les gènes d'un même opéron présentent le même profil d'expression puisqu'ils sont généralement transcrits sur le même ARNm. Les résultats sont identifiés comme cohérents si, lorsque l'on détecte un même opéron comme différentiellement exprimé, les gènes qui le composent ont des rangs de détections proches les uns des autres. La sensibilité et la précision des méthodes d'analyse sont évaluées à partir de ce critère de rang.

2.8. Bioinformatique

2.8.1. Définition

La bio-informatique moderne est née de la convergence de deux aspects de la recherche en biologie : le stockage des séquences moléculaires sur ordinateurs sous la forme de bases données et l'application d'algorithmes mathématiques pour l'alignement des séquences d'acides nucléiques et protéiques. Discipline hybride en constante évolution, la bioinformatique et ses domaines d'applications se précisent.

La plupart des définitions de la bio-informatique suggèrent l'interaction entre la biologie, les technologies de l'information et les sciences informatiques (les mathématiques). D'après Claverie et al.(1999), « la bio-informatique est la discipline de l'analyse de l'information biologique, en majorité sous la forme de séquences génétiques et de structures de protéines …C'est le décryptage de la « bio-information » (« Computational Biology » en anglais) ». Andrade et Sander, dans Bioinformatics : from genome data to biological knowledge, Current Opinion in Biotechnology (1997), présentent une définition plus large de la bio-informatique. Selon ces auteurs, « Bioinformatics is a science of recent creation that uses biological data, completed by computational methods, to derive new biological knowledge». Cette définition, plus moderne, sous-entend que la bio-informatique ne se limite évidemment pas à l'analyse des séquences. Un objectif fondamental est la volonté d'intégration de données de différentes natures, celles relatives aux séquences mais aussi celles concernant les marqueurs moléculaires, les données phénotypiques, etc.La bioinformatique est une approche in silicode la biologie traditionnelle qui vient compléter les approches classiques in situ(dans le milieu naturel), in vivo(dans l'organisme vivant) et in vitro (en éprouvette).

La bio-informatique est une branche théorique et pratique de la biologie. Sur le plan théorique, sa finalité est la synthèse des données biologiques à l'aide de modèles et de théories en énonçant des hypothèses généralisatrices et en formulant des prédictions. Sur le plan pratique, son but est de proposer des méthodes et des logiciels pour la sauvegarde, la gestion et le traitement de données biologiques. Par souci de clarté, les Anglo-saxons, utilisent deux termes pour distinguer ces deux aspects de la bio-informatique. Associé au terme de "bioinformatics" pour l'aspect pratique, ils utilisent le terme générique de « biocomputing» ("computational biology" pour les Américains) pour désigner l'aspect théorique.

2.8.2. Historique

a) Emergence de la bio-informatique

Dès 1965, quelques dizaines de laboratoires dans le monde travaillent avec les biomathématiques, disciplines constituées pour répondre aux besoins de la phylogénie moléculaire, de la modélisation et de la génétique des populations. La même année est publiée le premier atlas sur les séquences et structures protéiques, par Margaret Dayhoff. Dès lors, les bases de données biologiques se développent. En 1979 et 1980, les premières bases de

données de séquences d'ADN apparaissent avec la Los Alamos Sequence librarydu DOE, qui devient en 1982 GenBank, et EMBL du Laboratoire Européen de Biologie Moléculaire (Brown, 2003). Enfin, la base théorique de la plupart des algorithmes qui constituent aujourd'hui le cœur de nombreux outils bio-informatiques (modèles de Markov, échantillonneur de Gibbs…) date également de cette époque. Cependant, le terme « Bioinformatics» n'est apparu dans la littérature scientifique qu'au tout début des années 1990 (Brown, 2003). Longtemps cantonné dans les articles aux matériels et méthodes, l'emploi du terme « bio-informatique » n'apparaît que très tardivement dans les bases de données bibliographiques. Avant 1985, le terme « bio-informatique » n'est pas indexé comme mot clé par la base de données de références bibliographiques médicales Medline. Jusqu'en 1992, il n'apparaît pas non plus dans les titres ou les résumés référencés. En 1993, le terme apparaît enfin 3 fois puis 9 et10 fois en 1994-95 pour ensuite augmenter de façon exponentielle. Les premiers articles dans le domaine ont le plus souvent été publiés dans les journaux The journal of Molecular Biology, Nucleic Acids Researchet Computer Applications in Biological Sciences. Ce dernier, fondé en 1985, devient en 1998 Bioinformatics, aujourd'hui journal de référence de la discipline. Désormais, plus d'une dizaine de journaux consacrés à la bio-informatique existent. Avec le développement de l'Internet, certains de ces journaux tel que BMC bioinformaticsne paraissent même plus sous format papier.

b) La communauté bio-informatique

La communauté bio-informatique s'est largement développée au cours de ces 10 dernières années. Des nombreux groupes de travail nationaux, européens et internationaux ont vu le jour.

En 1997, né le consortium ISCB ou The International Society for Computational Biology. Issu des conférences du ISMB (Intelligent Systems for Molecular Biology) initiées en 1993, le ISCB se consacre à l'avancement de la compréhension des systèmes vivants par le calcul. Cette organisation compte plus de1300 membres et a pour journal officiel Bioinformatics. De cette initiative sont nées de nombreuses conférences pour le développement de la bio-informatique, parmi lesquelles nous pouvons citer JOBIM (Journées Ouvertes à la Biologie, Informatique et Mathématiques), ECCB (European Conference on Computational Biology), Computer Society Bioinformatics Conférence ou encore Pacific Symposium on Biocomputing. Les thèmes abordés lors de ces conférences sont très variés allant de l'analyse de séquences aux traitements des données issues des technologies « haut débit » (SAGE, CGH, puces à ADN) en passant par la représentation des connaissances et les ontologies. Plus spécifiquement, dans le domaine des puces à ADN, nous pouvons mentionner le groupe de travail MGED[2] (Microarray Gene Expression Data Society) et la conférence CAMDA

(Critical Assessment of MicroarrArray Data Analysis). MGED est une organisation internationale composée de biologistes et bio-informaticiens dont le but est la standardisation et le partage des données issues des expériences de génomique fonctionnelle et protéomique. La conférence CAMDA, quant à elle, a vu le jour au département de ressources informatiques de l'université de Duke (Johnson et Lin, 2001). Elle vise à établir l'état actuel des connaissances concernant les méthodes d'exploitation des données de puces à ADN, identifier les progrès et définir les nouvelles orientations. Dans ce but, CAMDA a adoptée une approche originale. Elle propose une expérience à l'échelleinternationale, laissant les scientifiques analyser le même jeu de données avec différentes méthodes. Les techniques sont ensuite présentées et discutées.

2.9. Bio-informatique et puces à ADN

2.9.1. Besoins

Les techniques bio-informatiques sont essentielles à la mise en place des méthodes d'analyse du transcriptome ainsi qu'à la gestion et l'exploitation des données qui en résultent. Les paragraphes suivants font état de quelques uns des besoins dans le domaine des puces à ADN.

Le choix des gene reporters à déposer sur les puces à oligonucléotides n'est pas trivial. De leurs propriétés physico-chimiques dépendent leur sensibilité et spécificité. De nombreux outils bio-informatiques ont donc été développés dans lebut d'optimiser ce rapport. Les algorithmes s'attachent notamment à valider les alignements de séquences (pas de structures secondaires internes, ni d'hybridation croisée…), établir la distance par rapport à l'extrémité 3'UTR de la séquence ou encore calculer des paramètres thermodynamiques tels que l'enthalpie et l'entropie du complexe sonde-cible pour estimer la température de dénaturation (Stekel 2003).

Les bases de données sont devenues des outils informatiques indispensables pour sauvegarder, structurer, sécuriser et manipuler les données. En effet, les puces à ADN appartiennent à ces nouvelles technologies dites à « haut débit » qui génèrent une masse considérable de données qu'il faut savoir gérer. Les informations enregistrées font référence non seulement aux résultats mais aussi à l'ensemble des étapes mises en œuvre pour concevoir les puces. Il existe un grand nombre de bases de données dédiées aux expériences de puces à ADN (Dudoitet al., 2003) parmi lesquelles BASE (Saalet al., 2002), ArrayDB (NHGRI), Acuity® (Axon Inc.) ou encore Rosetta Resolver® (Rosetta).

Une autre finalité des bases de données est la standardisation des informationsà sauvegarder pour un meilleur partage des connaissances. Ainsi, le consortium MGED propose MIAME (Minimun Information About Microarray Experiment) qui correspond à la liste des informations minimales à enregistrer pour décrire une expérience de puces à ADN (Brazmaet

al., 2001). MIAME est aujourd'hui la référence pour diffuser les données de puces à ADN sur les banques de données publiques (repository) telles que ArrrayExpress[13] à l'EBI ou Gene Expression Omnibus au NCBI. Par ailleurs, ce mode de diffusion est devenu, pour de nombreux journaux la condition sine qua non à la publication des travaux issus de cette technologie.

Les méthodes mathématiques et statistiques sont aussi devenues incontournables pour le traitement et l'interprétation des données de puces à ADN. En effet, les matrices de données d'expression présentent généralement des caractéristiques atypiques : les données sont le plus souvent bruitées et les matrices sont généralement dissymétriques (plus de gène reporters que d'échantillons). Aussi, la nécessité de valider la qualité des données et la difficulté d'analyser des matrices dissymétriques sont à l'origine de nombreuses recherches et d'un grand nombre de développement mathématiques et statistiques. De plus, compte tenu de la quantité des informations générées, une analyse manuelle devient très rapidement fastidieuse et source d'erreurs. L'exploitation des données ne peut se faire sans l'aide de procédures automatiques, i.e.d'outils logiciels. Ainsi, de nombreux algorithmes et outils, à commencer par les logiciels d'analyse d'images, sont développés pour l'acquisition, le traitement et l'analyse des données de puces à ADN.

Enfin, le traitement et l'interprétation des données issues des expériences de puces à ADN (et de manière plus générale des donnéesgénomiquees) évoluent constamment. Les outils pour le développement des méthodes de traitement et d'analyse doivent donc être flexibles. Ce besoin a incité les chercheurs en bio-informatiques à s'orienter vers des logiciels possédant des environnements de développementtels Microsoft Excel®, SAS®, S-plus®, Matlab® ou R (Ihaka et Gentleman, 1996). Les principaux avantages de ces logiciels sont leur souplesse et leur interopérabilité avec d'autres outils informatiques comme les banques et les bases de données accessibles sur le Web. Ils offrent ainsi de nombreuses possibilités d'analyses avec une perspective d'intégration des diverses sources de données pour une meilleure interprétation.

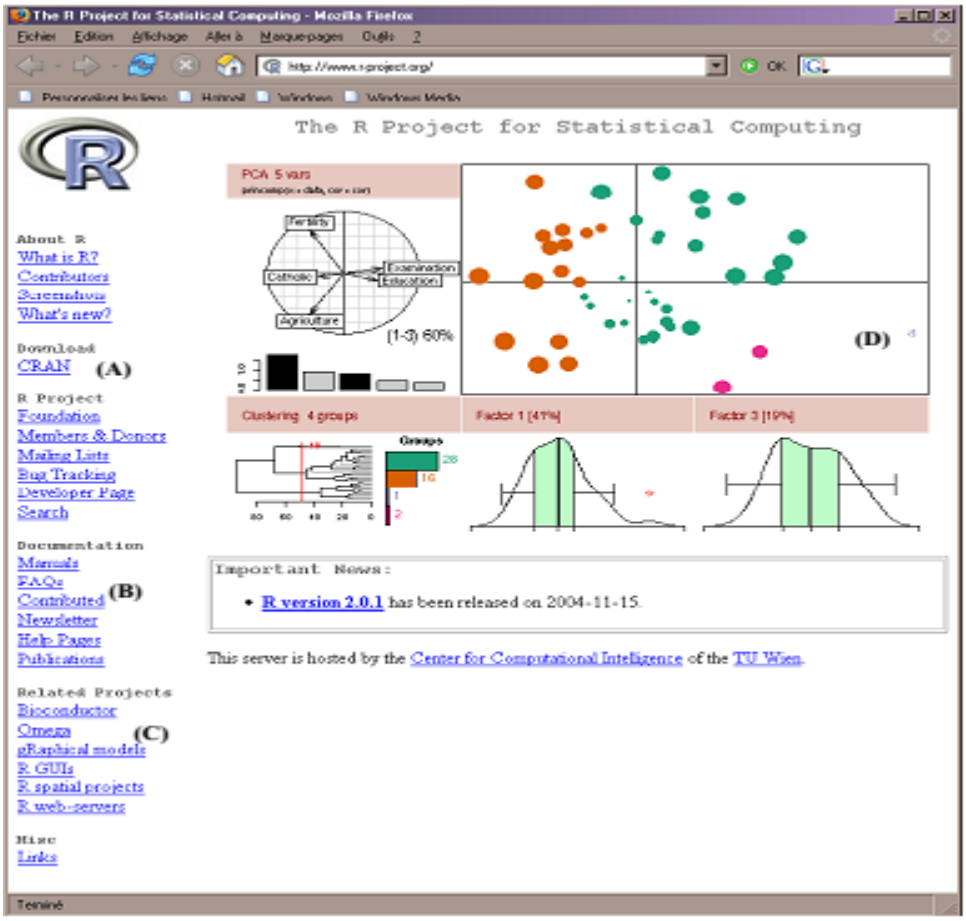


Figure 16. Site Web du projet R (www.r-project.org). La page d'accueil présente la version courante de R (ici R 2.0.1). De nombreux liens hypertextes (menu de gauche) donnent notamment accès aux (A) sites miroirs pour le téléchargement de R et de ses modules, (B) aux différentes documentations dont des manuels imprimables et des listes de diffusion, et (C) au projet Bioconductorpour l'analyse des données issues des expériences de génomique(puces à ADN, SAGE...). (D) Exemples des fonctionnalités graphiques de R.

2.9.2. R et BioConductor

D'après The Bioinformatics Organization, Inc., R est actuellement l'outil le plus utilisé pour le traitement numérique des données biologiques (soit 24% des 1136 votants contre 19% pour Matlab et 6% pour SAS®).

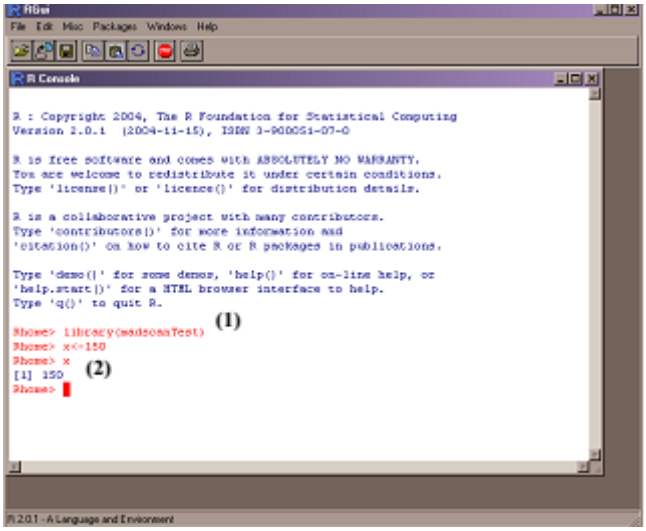
a) Historique

R est un outil d'analyses statistiques et graphiquesqui possède son propre langage de programmation. Nommé ainsi en référence à ses deux auteurs, Ross Ihaka et Robert Gentleman (1996), son nom est aussi un clin d'œil au langage S de AT&T Bell Laboratoires dont il est un dialecte. Contrairement au langage S et à l'outil d'analyse statistiques S-plus, commercialisés par Insightful®, R est distribué gratuitement suivant les termes des licences

publiques (GPL). Les codes sources et modules d'applications sont donc librement mis à la disposition de l'ensemble de la communauté scientifique.

Dans un premier temps développé pour les systèmes d'exploitation libres (et gratuits) à savoir UNIX et Linux, R est très vite devenu disponible pour les systèmes d'exploitation Windows et Mac-OS. Le noyau de R est implémenté essentiellement en langage C et FORTRAN. Ses versions sont distribuées sous la forme de codes sources binaires à compiler (UNIX et Linux) ou d'exécutables pré-compilés (Windows). Les fichiers d'installation sont disponibles à partir du site Web du CRAN (Comprehensive R Archive Network) (Cf. Figure. 16). Ce site répertorie également une importante source de documentation pour l'installation et l'utilisation de R sur chaque système d'exploitation. Depuis 1997, un groupe de développeurs (R Core Team), s'attache au maintien du bon développement des différentes versions de l'outil qui ne cesse de s'améliorer en termes de fonctionnalités graphiques et domaines d'applications (de l'exploitation des données géologiques à la génomique).

A)



B)

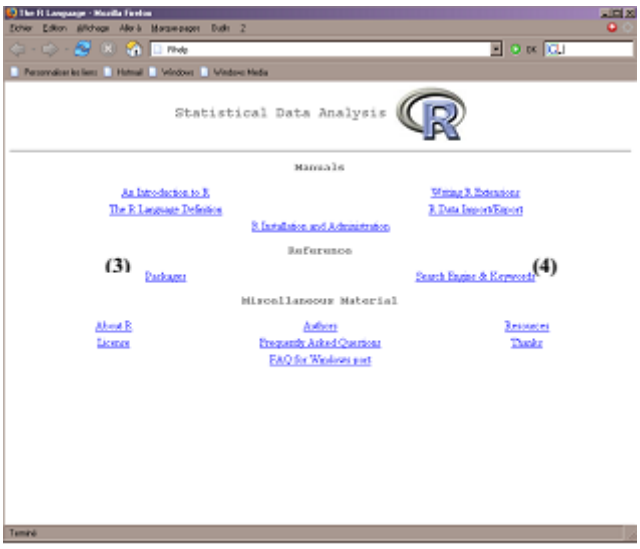


Figure 17 : L'ogiciel R : **(A)** Environnement de développement **(B)** Documentation électronique.(1) Appel d'une librairie de fonctions (2) définition et affichage d'une variable numérique. (3) liste des librairies installées localement (4) Moteurde recherche pour l'aide (en local)

b) Propriétés de R

R est un langage orienté objet ce qui signifie que les variables, les données, les fonctions, les résultats (etc.) sont stockés dans la mémoire del'ordinateur sous forme d'objets qui ont chacun un nom (Figure 17A). R est également un langage interprété, i.e. non compilé. Les commandes entrées au clavier sont directement exécutées et, contrairement à la plupart des langages informatiques (C, FORTRAN, JAVA...), la construction d'un programme complet n'est pas nécessaire. Cette propriété permet d'évaluer rapidement la qualité des algorithmes et de les déboguer. Cependant, l'exécution d'un tel programme peut être plus coûteuse en temps machine qu'un programme équivalent compilé.

Outil d'analyses statistiques et graphiques, R possède un environnement graphique d'applications qui permet l'exécution de commandes non seulement en mode interactif mais aussi sous forme de programmes (scripts). Cette fonctionnalité permet aux développeurs de créer des librairies de fonctions. Ces modules sontdédiés à des analyses spécifiques telles la librairie ctestqui proposent de nombreux tests statistiques ou la librairie blightyqui permet de dessiner le contour des côtes britanniques. L'interface graphique offre donc également la possibilité de réaliser des représentations graphiques très sophistiquées (Fig. 15.). Toutefois, les fonctions associées sont généralement complexes et le résultat peu interactif. Un autre atout de R est son interopérabilité. R peut dialoguer et interagir avec d'autres logiciels open-sourceécrits dans des langages différents. L'initiative Omegahat a notamment contribué à promouvoir et développer cette interopérabilité. De nombreux scripts et API (Application

Programming Interface) permettent ainsi une interaction bidirectionnelle de R avec les langages S, PERL, Python, Java et Visual Basic. Enfin, R possède une importante communauté de développeurs et une documentation très riche (Fig. 17B). Les documents fournis pour l'installation et la création de librairies sont généralement très détaillés. Chaque librairie s'accompagne également d'une documentation qui décrit chaque fonction (paramètres d'entrée, format de sortie des résultats) et présente le plus souvent des exemples d'exécution.

c) R et la génomique : le projet BioConductor

Les projets BioPerl, BioJava, BioPython…etc. proposent différentes solutions pour le traitement et l'analyse des données biologiques. La plupart des algorithmes ont été développés pour les analyses de séquences et très peu pour les analyses de données quantitatives telles que les données de bio-puces (ADN, protéines). R, langage et outil d'analyses statistiques, s'est avéré plus puissant pour le traitement des données numériques.

L'analyse des données de puces à ADN avec R a été initiée par un groupe de statisticiens dirigés par Terry Speed. Leur première librairie, nommée sma (statistical microarray analysis), a été développée pour répondre aux problèmes de normalisation des données de puces à ADN deux couleurs. Cet outil et les fonctions associées ont eu un impact considérable dans le domaine de l'analyse des puces à ADN. Compte tenu des résultats, des propriétés de R (fonctions et puissance de calcul) et du besoin croissant d'outils mathématiques pour l'analyse des données biologiques, des développeurs au sein de la communauté R ont proposé le projet BioConductor.

BioConductor est une initiative de collaboration entre statisticiens, mathématiciens, biologistes et développeurs afin de créer des outils informatiques (algorithmes, logiciels) pour résoudre des problèmes de biologie et de bio-informatique (Gentlemanet al., 2004). Les principaux buts de ce projet sont le développement, en collaboration, de logiciels innovants ainsi que leur vaste diffusion et utilisation, pour une reproductibilité des résultats de recherche.

Né en 2000, BioConductor, associé à R, reçoit en 2002 le titre de Insightful Innovation Award Open Source & Open Development Software Project. Insightful® commercialise aujourd'hui S+ Analyser, un outil relativement convivial qui reprend en majorité les librairies de BioConductor. De même, GeneTrafic® d'Iobion, logiciel dédié à la gestion et au traitement des données de puces à ADN, utilise de nombreuses librairies de BioConductor.

Enfin, dédiées à l'analyse des données de génomique, les librairies disponibles sur le site de BioConductor permettent non seulement l'analyse des données de puces à ADN (e.g.librairies Affy, marray, limma) mais aussi des expériences SAGE (SAGElyzer), de la spectrométrie de

masse (PROcess) ou encore l'annotation des gènes (GOstats).Issu du projet R, BioConductor en possède les avantages et les inconvénients. Des interfaces utilisateurs sont disponibles pour quelques librairies, telles limmaGUI ou affylmGUI, et facilitent leur emploi. Un système de « vignettes »documente le fonctionnement de certaines librairies et permet parfois l'exécution interactive d'exemples. Néanmoins, l'utilisation de la majorité des librairies nécessite une certaine expertise en R. De plus les modes de visualisation graphiques restent encore peu interactifs.

3. Les réseaux biologiques

3.1. Qu'est-ce qu'un réseau biologique ?

Un réseau biologique est une représentation abstraite d'un système biologique. Lorsque l'on parle de réseau, on se place dans une démarche de modélisation. La notion de réseau biologique est une notion très large, elle représente plutôt un niveau d'étude qu'une réalité biologique. L'idée principale qui motive l'utilisation de cette abstraction est la suivante : pour comprendre un processus biologique, il ne suffit pas de donner la liste des éléments qui y participent, cette liste ne constitue qu'une première étape à laquelle il faut ajouter l'étude des interactions entre ces éléments. Le terme de réseau en biologie est donc lié au terme d'interaction. De fait, la notion de réseau se retrouve dès lors qu'on veut modéliser des interactions en biologie, et ce, à différents niveaux de détails : depuis les interactions atomiques dans un repliement protéique jusqu'aux relations entre organismes dans une population ou un écosystème. Dans cette thèse, nous allons nous concentrer plus précisément sur les réseaux d'interactions moléculaires, qu'on peut définir comme un ensemble de nœuds, représentant des gènes, des produits de gènes ou des métabolites,
et un ensemble de connections représentant les interactions entre ces entités [Alm et Arkin, 2003]. Au sein même des réseaux d'interactions moléculaires (on parle aussi de réseaux cellulaires), il apparaît nécessaire d'opérer des subdivisions afin d'accéder à un niveau de description satisfaisant. En effet, si on analyse les réseaux d'interaction moléculaire dans leur ensemble, on se condamne à ne faire que des remarques très générales. Le terme de réseau biologique a en particulier été utilisé pour désigner un des processus suivants :
– le métabolisme ;
– la régulation des gènes ;
– la transduction de signaux.
Chacun de ces processus met en jeu différents types de molécules : les gènes (fragments d'ADN) ; les transcrits (ARN) ; différents types de protéines : facteurs de transcription, enzymes ; et enfin des petites molécules : les métabolites.

Dans une approche qui n'est pas centrée sur les processus, on peut aussi définir les réseaux d'interaction de protéines. Ils rendent compte de toutes les interactions physiques entre protéines. Ils regroupent aussi bien la formation de complexes que des cascades de phosphorylation.

On peut noter que les interactions indirectes, comme des enzymes catalysant des étapes successives d'une voie métabolique ou un facteur de transcription agissant sur la régulation transcriptionnelle d'une protéine, ne seront généralement pas couvertes par ce type de données. Cependant, certaines enzymes ou facteurs de transcription sont des complexes de plusieurs polypeptides. En outre, certains facteurs de transcription interagissent directement (de façon synergique) pour influer sur la transcription d'un gène. Les réseaux d'interaction de protéines recouvrent donc différents types de processus.

Quand on modélise ces processus par des réseaux, on opère généralement des simplifications, d'une part parce qu'on ne sait pas quelles sont toutes les molécules qui y sont réellement impliquées (les reconstructions sont incomplètes), et d'autre part, pour pouvoir faire des calculs, on construit des modèles qui sont le plus souvent des simplifications des réelles interactions moléculaires. Ainsi, un réseau de régulation de gènes va indiquer des interactions directes entre gènes alors que le mécanisme est indirect : l'un des gènes code pour un facteur de transcription, qui va être transcrit puis traduit pour pouvoir agir sur la transcription d'un second gène. Ces simplifications sont parfois acceptables mais parfois insuffisantes, suivant l'application qu'on considère.

3.2. Des protéines aux réseaux biologiques

3.2.1. Données

On a pu voir que les données étaient amenées à jouer un rôle central dans une démarche d'acquisition de connaissances en bioinformatique. Aussi, il est nécessaire de bien comprendre de quelles données on dispose en pratique et de savoir comment elles ont été obtenues pour mieux décider du niveau de confiance qu'on peut leur accorder.

Dans cette partie, nous parlerons essentiellement de données disponibles publiquement dans des bases puisque c'est ce type de données que nous avons principalement utilisé. Certaines bases de données sont généralistes (KEGG), d'autres sont spécifiques d'un organisme (EcoCyc). Il existe également des jeux de données disponibles publiquement qui sont spécifiques à une question biologique et qui ont été constitués précisément pour y répondre.

Au sein des données disponibles dans les bases, on distingue ainsi 3 types :

– les données issues de la littérature (ou données bas débit) ;

– les données issues d'expériences à haut débit ;

– les données inférées.

On peut également faire la différence entre données qualitatives (quelles sont les molécules qui interagissent) et données quantitatives (quelles sont les constantes d'association). On parlera dans cette partie essentiellement de données qualitatives. Ces données permettent de travailler avec des modèles qualitatifs afin de poser des questions d'ordre structurel. On note qu'en pratique, les données quantitatives ne sont disponibles que pour des voies très étudiées.

Enfin, parmi les types de données dont nous ne parlerons pas, on peut mentionner les données issues de la métabolomique, nouveau domaine en expansion. Une expérience de métabolomique consiste à mesurer tous les métabolites présents à un instant donné dans une cellule. Ces données peuvent être qualitatives ou quantitatives. Pour une introduction à ce domaine, voir [Nobeli et Thornton, 2006].

3.2.2. Interactions protéine-protéine (IPP)

Les protéines sont l'un des principaux composants de la matière vivante. En effet, elles constituent la majeure partie de la masse sèche des cellules (Alberts, 1998) et sont impliquées dans de très nombreux processus allant de la protection de l'organisme à la réplication de l'information génétique, en passant par la transduction de signaux cellulaires.

Les protéines ne travaillent pas seules. En effet, la majorité des processus biologiques font intervenir plus d'une dizaine d'entre elles, chaque protéine interagissant avec une ou plusieurs autres protéines et formant ainsi des complexes protéiques transitoires ou permanents. Ainsi, on estime l'interactome humain (l'ensemble des interactions protéine-protéine) à environ 130 000 interactions (Venkatesan et al., 2009). A une échelle moindre, la base de données SynSysNet spécialisée dans les protéines de la synapse recense 4638 interactions connues au sein des synapses (von Eichborn et al., 2013). Au sein de ce réseau d'interactions, toutes les protéines ne sont pas également connectées. En effets certaines n'interagissent qu'avec une protéine, alors que d'autre interagissent avec plusieurs centaines de protéines. Par analogie avec les réseaux de télécommunications, ces protéines centrales sont dénommées "hub" et sont particulièrement importantes pour le fonctionnement des cellules de par leur rôle central dans la formation de complexes (Jeong et al., 2001, Pang et al., 2010).

3.2.3. Mise en évidence d'interactions physiques entre protéines

Il existe de nombreuses techniques expérimentales pour mettre en évidence les interactions physiques entre protéines. L'une des premières méthodes haut débit développée est la méthode du "double hybride". Les paires des protéines dont on veut tester l'interaction sont exprimées sous forme de protéines chimériques. Sur l'une des deux protéines on ajoute un domaine de fixation à l'ADN et sur la seconde protéine un domaine activateur de la

transcription. Si les deux protéines interagissent, la présence de ces deux domaines entrainera la transcription d'un gène rapporteur et donc la détection de la transcription (Ito et al., 2001). Cette technique est puissante car elle se déroule in vivo et permet de détecter des interactions même transitoires.

Contrairement à la méthode du double hybride qui n'identifie que des couples de protéines interagissant, la méthode TAP-MAS (Tandem Affinity Purification - Mass Spectrometry) permet de mettre en évidence les complexes multiprotéiques (Puig et al., 2001). Cette technique s'appuie sur la création d'une protéine chimère formée d'une séquence tag et de la protéine d'intérêt. Cette séquence tag permettra de retenir la protéine d'intérêt dans une colonne d'affinité. Ainsi, lors du passage des protéines à tester dans la colonne, les protéines formant un complexe avec la protéine chimère seront retenues. La purification des complexes permettra ensuite d'identifier leurs composants par spectrométrie de masse.

3.2.4. Mise en évidence d'interactions fonctionnelles

Il existe également des méthodes dites indirectes pour détecter des interactions entre protéines. On parle alors plutôt d'interactions fonctionnelles au lieu d'interactions physiques.

Une méthode indirecte utilisée pour les organismes procaryotes est la notion de voisinage génomique. Cette méthodologie est possible grâce à l'organisation en opérons des génomes procaryotes. Les opérons sont des ensembles de gènes voisins qui sont régulés par le même facteur de transcription et impliqués dans les mêmes voies biologiques. Ainsi, en observant que deux gènes sont très fréquemment voisins dans le génome de plusieurs organismes, il est probable que les protéines issues de ces deux gènes aient une interaction fonctionnelle (Overbeek et al., 1999).

Pour les organismes dont les gènes ne sont pas organisés en opérons, il est possible d'étudier les co-expressions de gènes. En effet, une conservation de la co-expression de 2 gènes dans de multiples organismes indique un avantage sélectif lors de l'évolution et donc que les protéines codées par ces gènes interagissent (Stuart et al., 2003).

Une autre manière de détecter des interactions fonctionnelles est d'étudier les évènements de fusion de gènes. En effet, deux protéines d'un organisme peuvent être en interaction si elles sont également présentes dans un autre organisme sous la forme de deux domaines d'une seule protéine (Yanai et al., 2001).

Les interactions protéine-protéine peuvent être conservées entre les organismes proches (Walhout et al., 2000), on parle alors d'interologues (issue de la combinaison d'interaction et d'orthologue). En utilisant cette notion, on peut alors prédire des interactions en recherchant les interactions existantes dans des organismes proches.

3.3. Bases de données d'interactions protéine-protéine

Les différentes méthodes mentionnées ci-dessus permettent de mettre en évidence de plus en plus d'interactions protéine-protéine. Devant ce nombre croissant de données, de nombreuses sources de données se développent afin de mettre les informations sur les interactions protéineprotéine à la disposition des biologistes. L'une des plus importantes est certainement la base de données IntAct[2] qui recense les interactions protéine-protéine décrites dans la littérature (Kerrien et al., 2012). Une autre source de données est la base STRING[3] qui reporte les interactions protéine-protéine élucidées expérimentalement ou prédites (Figure 2.1).

3.3.1. Autres réseaux biologiques

L'interactome peut être divisé en modules. Ainsi, Alberts (1998) compare ces sous-réseaux à des machines, dans lesquelles les protéines sont organisées en modules de manière à réaliser des fonctions précises. Au sein d'un module, les protéines sont fortement connectées entre elles tandis que les interactions avec des membres extérieurs au module sont plus rares.

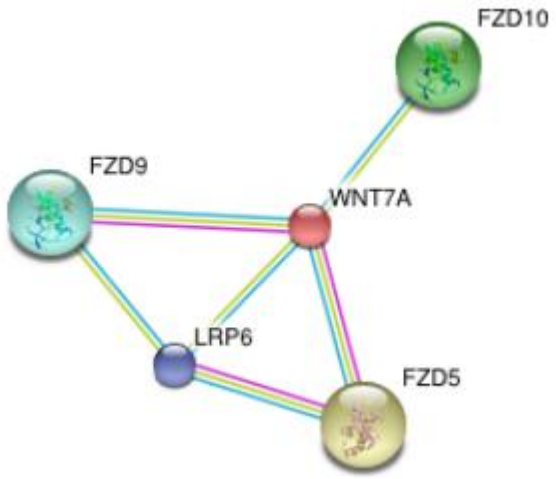


Figure 18 : Réseau d'interaction de la protéine WNT7A, extrait de la base de données String. Les relations en rose ont été élucidées expérimentalement, les relations bleues sont extraites d'autres bases de données et les relations vertes ont été prédites par fouille de texte.

Par exemple, on peut remarquer sur la figure 18 qu'au sein du module correspondant au métabolisme du galactose il existe seulement 5 connections avec d'autres modules. Ainsi, cette organisation peut être utilisée pour définir des réseaux biologiques. Ces réseaux peuvent être définis comme un ensemble d'interactions entre des composants physiques ou génétiques de la cellule, décrivant un processus de cause à effet ou dépendant du temps, et expliquant des phénomènes biologiques observables (Demir et al., 2010). Différents types de processus sont décrits par ces réseaux : la régulation de gènes, le transport de molécules, la transformation de petites molécules ou une interaction entre protéines entrainant la modification de l'une d'entre elle (Schaefer et al., 2009). Ainsi, trois groupes de réseaux sont couramment définis : voies métaboliques, voies de signalisation et voies de régulation.

Les voies métaboliques décrivent le métabolisme de la cellule, comme par exemple le métabolisme du galactose (Cf. Figure 18), et font intervenir des petites molécules qui sont modifiées par les réactions chimiques réalisées par des protéines. Les voies de signalisation correspondent à la perception de signaux cellulaires ou extracellulaires, puis à la transduction du signal dans la cellule. Le troisième groupe de voies est associé à la régulation des gènes. Le plus souvent ces 3 voies sont connectées : la cellule perçoit un signal qui sera transmis jusqu'au noyau et entrainera une modification de la régulation de gènes, ce qui pourra provoquer des changements du métabolisme de la cellule.

3.3.2. Les ressources sur les réseaux biologiques

Il existe plusieurs ressources contenant des réseaux biologiques connus. Les informations qui suivent correspondent à l'état de ces ressources en avril 2013. L'une des plus connues est certainement KEGG PATHWAY[4] contenant 496 réseaux construits manuellement et repartis en 6 catégories : métabolisme, traitement de l'information génétique, traitement des informations environnementales (signalisation), processus cellulaires, systèmes physiologiques et maladies humaines (Ogata et al., 1998). Reactome[5] est une autre source disponible pour décrire les réseaux biologiques (Joshi-Tope et al., 2005). Elle contient 1402 réseaux regroupés en 22 grands groupes et est peuplée manuellement. t

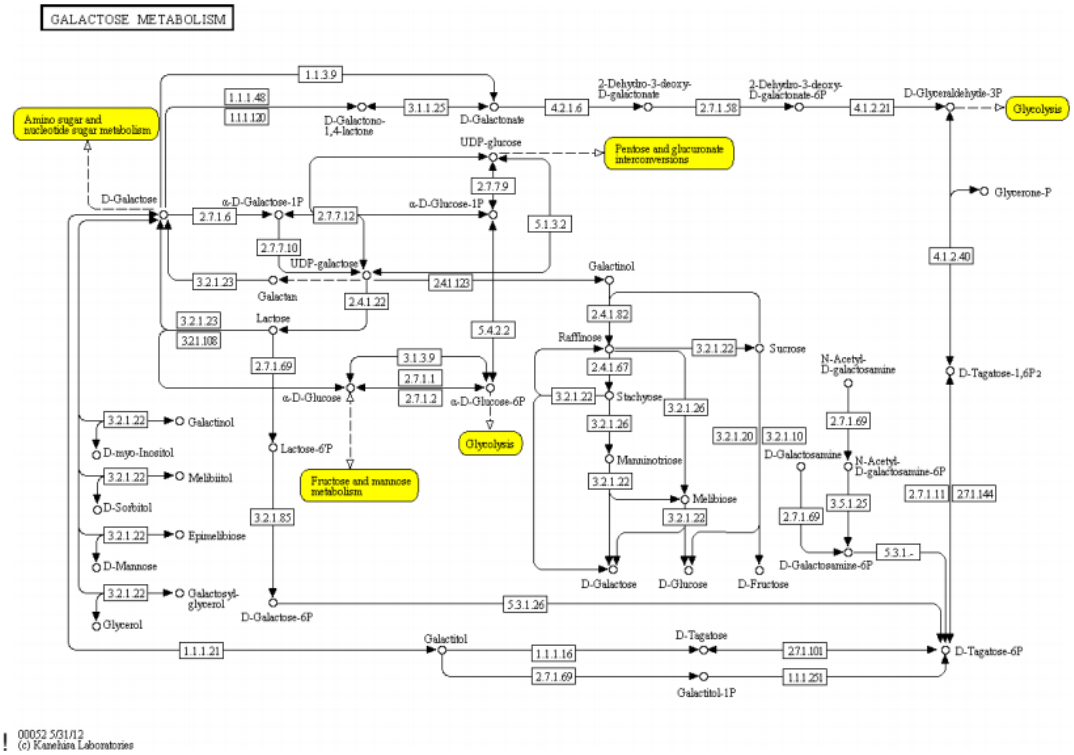


Figure 19 : Voie métabolique KEGG de référence correspondant au métabolisme du galactose. Les protéines sont représentées sous la forme de rectangles contenant le numéro EC

("Enzyme Commission") de la protéine et les petites molécules sont représentées par des cercles. Les éléments en jaune sont les voies métaboliques d'autres sucres.

Une troisième source, BioCarta[6], répertorie 354 réseaux vérifiés manuellement et repartis en 22 fonctions. La Pathway Interaction Database (PID) est une base de données intégrée. En effet, elle contient 1367 réseaux vérifiés par le groupe NCI-Nature ainsi que 322 réseaux provenant de Reactome et BioCarta. Contrairement aux autres bases qui sont plus générales, PID se concentre sur les voies de signalisation et de régulation (Schaefer et al., 2009) et n'intègre donc que ces deux types de réseaux. Il existe également la ressource BioModels[7] qui contient des modèles informatiques de processus biologiques (Li et al., 2010). C'est à dire des représentations formalisées (par exemple dans le langage SBML) et quantifiées (en vue d'une simulation dynamique de processus) tels que le cycle cellulaire ou les cascades MAP kinases.

Les protéines notamment via leurs interactions ont un rôle majeur dans le fonctionnement des cellules. Les interactions protéine-protéine sont donc particulièrement étudiées et les méthodes de mise en évidence expérimentales et informatiques de ces interactions produisent de plus en plus de données. L'organisation de ces interactions permet de définir des réseaux biologiques qui expliquent des phénomènes cellulaires tels que la transduction de signaux au sein de la cellule ou la régulation des gènes. Il est important d'aider l'utilisateur biologiste à exploiter la richesse et la profusion des données disponibles. De nombreux formats permettent de décrire ces réseaux et des outils de visualisation des réseaux sont développés pour faciliter l'analyse de ces réseaux d'interactions.

3.4. Représentation et visualisation des réseaux biologiques

3.4.1. Formats de représentations des réseaux

Il existe plusieurs formats permettant la représentation de réseaux biologiques. Certains ont été développés spécifiquement pour les réseaux biologiques (BioPAX, SBML, ...) alors que d'autres sont plus généraux pour représenter des graphes quels qu'ils soient (OXL, GraphML, ...). Notons que la plupart de ces formats ne se limite pas à la représentation des réseaux en tant que tel mais permet la représentation des données sur les réseaux soit sous forme d'attributs des nœuds, soit sous forme de nœuds décrivant les propriétés. Dans ce dernier cas, nous parlerons alors de graphe étendu.

3.4.2. Formats de réseaux biologiques

a) BioPAX : BioPAX est un modèle XML spécifiquement développé pour permettre l'intégration, l'échange, la visualisation et l'analyse des données de réseaux biologiques (Demir et al., 2010). Il est divisé en trois niveaux : le premier décrit seulement les voies métaboliques, le deuxième prend en compte les voies de signalisation et les interactions

moléculaires en plus des voies métaboliques et le troisième niveau décrit les réseaux de régulation génique et les interactions génétiques.

BioPAX est basé sur une ontologie de concepts avec des attributs, ce qui permet d'avoir des relations entre concepts plus explicites que pour les autres formats (Pavlopoulos et al., 2008). Ainsi, le type d'interactions des entités physiques (en ajoutant les gènes) est particulièrement bien renseigné. Par exemple, il est possible de faire la différence entre une interaction de type dégradation et de type transport (Figure 20). De plus, BioPAX a été développé de manière à être compatible avec les formats existants comme SBML dans leurs domaines d'applications communs (Demir et al., 2010).

b) SBML : Le Systems Biology Markup Language (SBML) est un modèle XML décrivant de manière qualitative et quantitative les modèles de réseaux biochimiques (Finney et Hucka, 2003). En effet, il est orienté vers la description des systèmes dans lesquels des entités biologiques sont impliquées et modifiées par des processus au fil du temps. Il contient notamment des éléments permettant de décrire la fonction mathématique associée au modèle ainsi que les réactions et leurs paramètres entre les espèces réagissant. Ainsi, SBML est particulièrement bien adapté pour modéliser les voies de signalisation cellulaire, des voies métaboliques et les régulations géniques.

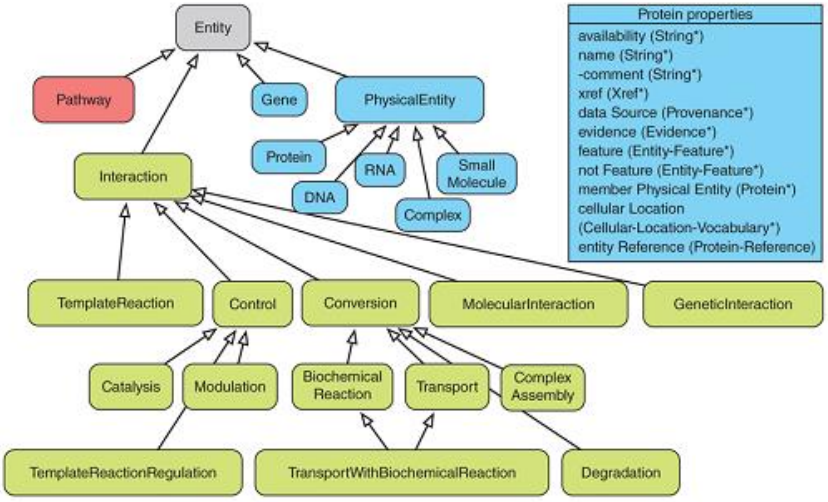


Figure 20 : Entités composant BioPAX. Les quatre types de classes composant BioPAX sont les réseaux biologiques (en rouge), les interactions (en vert) et les entités physiques avec les gènes (en bleu). Les flèches représentent les relations entre les entités BioPAX. La figure est extraite de Demir et al. (2010).

c) Formats de graphes

- ✓ **GraphML : GraphML** est un modèle XML développé spécifiquement pour les graphes (Brandes et al., 2001). Ainsi, ce modèle a été pensé de façon à décrire tous les types de graphes (orientés, non orientés, mixtes, hiérarchiques et hypergraphes). De

manière générique, GraphML est composé d'entités nœuds reliées entre elles par des arcs, un arc étant caractérisé par un nœud "source" et un nœud "cible".

- ✓ **OXL** : Le format OXL est un format spécifique à Ondex. Bien que le programme Ondex soit plutôt dédié à l'analyse de données biologiques (Kohler et al., 2006), le format OXL a été développé de manière à couvrir un grand nombre d'applications. Il est ainsi suffisamment flexible et extensible pour combiner différents types de données (Taubert et al., 2007). OXL est défini comme un schéma XML composé de deux grands types d'éléments : ondexdataseq et ondexmetadata. Ondexdataseq décrit les éléments composant le graphe, tandis qu'ondexmetada contient la liste de tous les types de métadonnées utilisées dans le graphe. Ondexdataseq est composé de 2 groupes d'éléments : les concepts et les relations, représentant respectivement les nœuds et les arcs du graphe (Figure 21). Les concepts sont caractérisés par un identifiant unique (id), un identifiant textuel alternatif (pid), des annotations et une description. La base de données source du concept est représentée par elementOf et son type par of Type.

La méthode de mise en évidence du concept est également stockée. Le(s) nom(s) du concept sont collectés dans concept_name et la liste de leurs différents identifiants dans des bases de données dans concept_accession. Il est également possible de définir des attributs comme la date de collecte ou l'organisme source. Cette liste des attributs "personnalisés" correspond à concept_gds. Les relations entre concepts sont décrites comme allant d'un concept source fromConcept vers un concept cible to Concept. Ces relations sont précisées par leur type ofType (interaction, traduction,...). Comme les concepts, il est possible de définir des attributs

supplémentaires (relation_gds) pour mieux caractériser ces relations.

```
<ondexdataseq>
 <concepts> <!-- Noeuds -->
  <concept> <!-- Premier concept -->
   <id>1</id>
   <pid>
    Aldo-keto reductase family 1 member C1
   </pid>
   <annotation/>
   <description/>
   <elementOf> <!-- BD source -->
    <idRef>Uniprot</idRef>
   </elementOf>
   <ofType> <!-- Type de donnée -->
    <idRef>Protein</idRef>
   </ofType>
   <evidences> <!-- code d'évidence -->
    <evidence>
  <idRef>IMPD</idRef>
    </evidence>
   </evidences>
   <cnames> <!-- Noms du concept -->
    <concept_name>
     <name>
      Aldo-keto reductase family 1 member C1
     </name>
    </concept_name>
   </cnames>
   <coaccessions> <!-- Identifiants dans BDs -->
    <concept_accession>
     <accession>Q04828</accession>
     <elementOf>
      <idRef>Uniprot</idRef>
     </elementOf>
    </concept_accession>
   </coaccessions>
   <cogds> <!-- Attributs personalisables-->
    <concept_gds> <!-- Organisme -->
     <attrname>
      <idRef>Organism</idRef>
     </attrname>
      <literal>Homo sapiens</literal>
     </value>
    </concept_gds>
   </cogds>
   <contexts/>
  </concept>
  <concept> <!-- Second concept -->
   <id>2</id>
   <!-- Autre concept similaire au concept précédent (par exemple une autre protéine) -->
  </concept>
 </concepts>
 <relations> <!-- Relations -->
  <relation>
   <fromConcept>1</fromConcept>
   <toConcept>2</toConcept>
   <ofType>
    <idRef>int_with</idRef>
   </ofType>
   <evidences>
    <evidence>
     <idRef>IMPD</idRef>
    </evidence>
   </evidences>
   <relgds>
    <relation_gds> <!-- Méthode de mise en évidence -->
     <attrname>
      <idRef>methode</idRef>
     </attrname>
     <value java_class="java.lang.String">
      <literal>two hybrid</literal>
     </value>
    </relation_gds>
   </relgds>
   <contexts/>
  </relation>
 </relations>
</ondexdataseq>
```

Figure 21 : Extrait d'un fichier OXL représentant les données Ondexdataseq et décrivant une interaction entre deux protéines

Les métadonnées (ondexmetadata) permettent de décrire les types de données utilisées dans le graphe. Elles sont caractérisées par un identifiant unique id, un nom fullname et une description description en texte libre et sont entièrement personnalisables par l'utilisateur. Il existe 6 types de métadonnées : les éléments de type cvs ("controlled vocabularies") représentent le vocabulaire décrivant les sources de données utilisées pour décrire les concepts et les relations. Les éléments units correspondent aux unités des propriétés des concepts et des relations. Les façons dont les concepts et les relations ont été mis en évidence (mise en évidence expérimentale, ...) sont représentées par les éléments de type evidences. Les éléments de types attrnames ("attribut names") qui sont des attributs définissables par l'auteur les concept classes et les relation types correspondent respectivement aux différents types de concepts et aux types de relations qui sont utilisées.

- ✓ **RDF :** Le Resource Description Framework (RDF) est un modèle d'échange de données sur le Web développé par le W3C8 . Un graphe RDF est composé de triplets (sujet, prédicat, objet) ou le sujet est la ressource à décrire, le prédicat est la propriété associée au sujet et l'objet correspond à la valeur de la propriété. Ainsi, le triplet correspondant à une interaction entre la protéine A et la protéine B est (A, interagit avec, B) ou (B, interagit avec, A).

3.5. Outils de visualisation

L'ensemble des interactions d'un organisme forme un réseau d'interaction protéine-protéine. Ce réseau est un outil important pour la compréhension des mécanismes cellulaires (Agapito et al., 2013). Étant donné que la visualisation des réseaux peut permettre de mettre en évidence des sousstructures intéressantes, comme des complexes protéiques, la visualisation de réseaux sous forme de graphes est particulièrement répandue (Ciofani et al., 2012, Cheng et al., 2012, Agapito et al., 2013, Campillos et al., 2008). Ainsi, de nombreux outils sont développés afin de filtrer et d'analyser les réseaux biologiques. Ces outils varient notamment sur les formats de données qu'ils utilisent, le mode de visualisation, la possibilité d'interroger des sources de données distantes, la possibilité d'annoter un réseau existant et la capacité de l'utilisateur à développer de nouvelles fonctionnalités au programme via des extensions.

3.5.1. Arena3D

Arena3D est un programme de visualisation de réseaux biologiques en trois dimensions (3D) (Secrier et al., 2012). Afin de faciliter la visualisation en 3D, chaque type d'élément (protéine, gène, maladie, ...) est affiché à une hauteur différente sur le graphe (Figure 22). Ce programme permet d'importer des fichiers SBML, PSI-MI et TXT. Et il est également capable d'interroger des bases de données distantes, telles que STRING pour les interactions protéine-protéine, OMIM pour les informations concernant les maladies génétiques, PDB pour les informations sur les structures et Gene Ontology pour les annotations fonctionnelles. Il est cependant impossible d'ajouter manuellement des annotations aux nœuds ou aux arcs. De plus, Arena3D ne permet pas le développement d'extensions afin de d'éviter des problèmes d'incompatibilité de technologies pouvant affecter les performances du programme (Agapito et al., 2013).

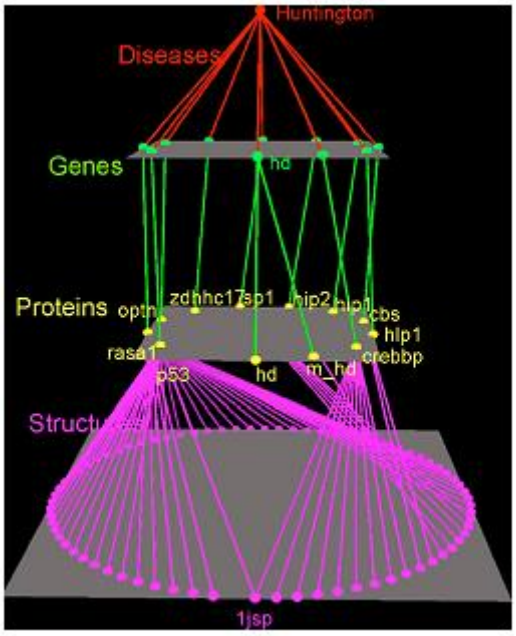


Figure 22 : Exemple de visualisation 3D avec Arena3D. Le réseau représente les gènes, protéines et structures protéiques associés à la maladie de Huntington et chaque type d'élément est représenté à une hauteur différente sur le graphe. La figure est extraite de Pavlopoulos et al. (2008).

3.5.2. BioLayout express3D

BioLayout express3D est spécialement conçu pour la visualisation en 3D et l'analyse de grands réseaux de données biologiques (Theocharidis et al., 2009). Ce programme est compatible avec un grand nombre de format de fichiers tels que : le format OWL de la base de données Reactome, les fichiers EXPRESSION, MATRIX, GraphML, mEPN, OXL, LAYOUT, SIF et TXT. Il ne permet pas de formuler des requêtes vers des bases de données externes. BioLayout express3D permet à l'utilisateur d'annoter les arcs notamment pour permettre de relier des nœuds grâce à des arcs ayant différentes annotations Pour les même raisons qu'Arena3D, BioLayout express3D ne permet pas le développement d'extensions.

3.5.3. Cytoscape

Cytoscape est l'un des outils de visualisation 2D de réseaux les plus populaires (Smoot et al., 2011, Cheng et al., 2012, Agapito et al., 2013). Ce programme permet de visualiser des réseaux allant jusqu'à des centaines de milliers de nœuds et de liens (Figure 22). Le principal objectif de Cytoscape est la visualisation d'interactions moléculaires et leur intégration avec des profils d'expression génique ou d'autres données. Cytoscape permet d'importer des réseaux existant sous différents formats : XML, RDF, OWL, GraphML, XGMML et SBML. Cytoscape permet également d'importer des réseaux à partir de bases de données distantes notamment via IntAct et les bases du NCBI. Cytoscape est capable d'importer des données

locales ou distantes (notamment GO) pour annoter les réseaux affichés. Il existe de nombreuses extensions développées par les utilisateurs, certaines permettant notamment de calculer l'enrichissement en termes GO du graphe étudié ou d'un sous graphe (Maere et al., 2005). Il existe également d'autres extensions permettant d'interroger des bases de données distantes afin d'enrichir le réseau affiché ou pour simplement créer un nouveau réseau, comme par exemple l'extension IntActWSClient qui permet d'interroger la base de données IntAct. De ce point de vu,

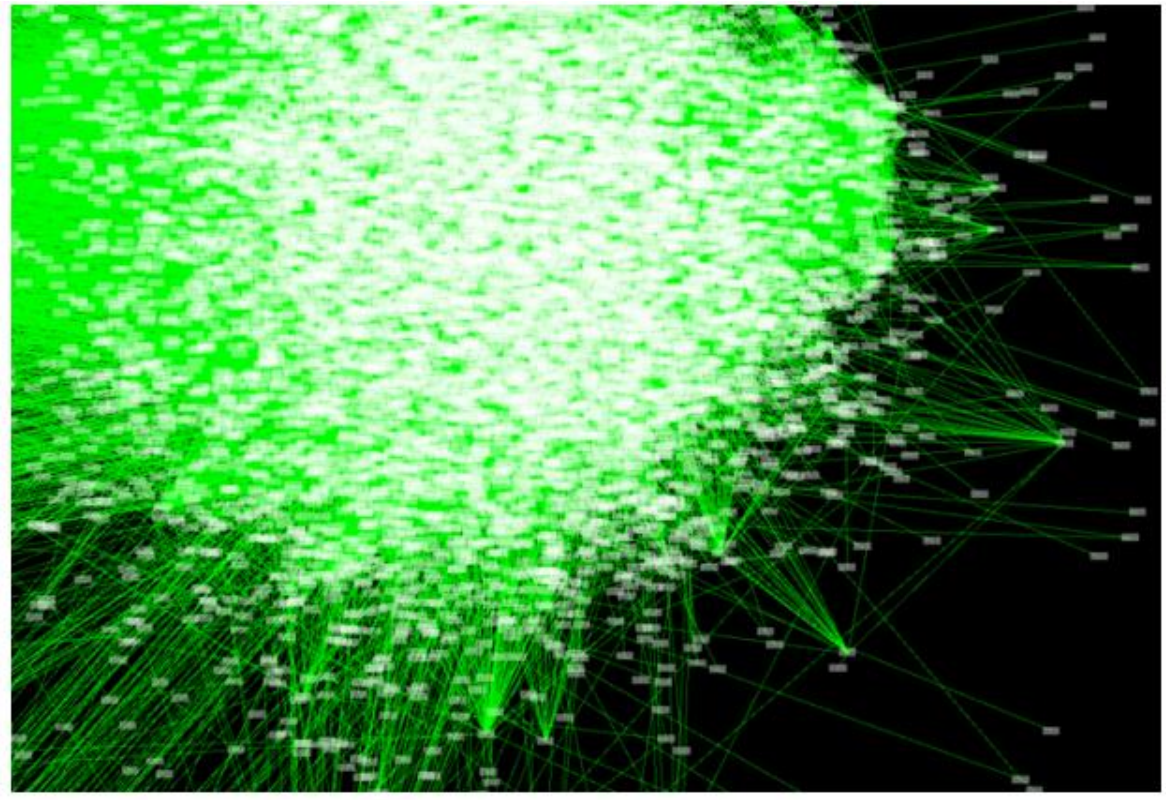


Figure 23 : Extrait de l'interactome humain (construit par Rual et al. 2005 et disponible à http://wiki.cytoscape.org/Data_Sets) sous Cytoscape. Au total, 10 203 protéines présentant 61 262 interactions sont affichées.

l'une des extensions les plus utiles est certainement BisoGenet (Martin et al., 2010) puisqu'a partir d'une liste de gènes ou de protéines, cette extension va interroger la base de données distante SysBiomics[9] qui intègre le contenu des bases de données d'interactions DIP, HPRD, IntAct, BioGRID, MINT et BIND. En utilisant BisoGenet à partir d'une liste de protéines, on peut obtenir l'ensemble de leurs interactants connus de rang 1, 2 ou 3. De plus, pour chaque élément affiché, les termes GO ainsi que les réseaux biologiques KEGG associés sont également présentés. Il est ensuite possible de rechercher le plus court chemin entre deux protéines du réseau.

3.5.4. Ondex

Ondex est un outil d'intégration et de visualisation de graphe 2D dédié aux données biologiques (Kohler et al., 2006). Il est capable d'importer les fichiers au format OXL (qui lui est spécifique), SBML et BioPAX. Ondex est développé comme un outil d'intégration de données biologiques. Il permet donc d'importer des données provenant de nombreuses sources telles que : Aracyc2, ChEBI, ChEMBL, EXPASY ENZYME, Gene Ontology, KEGG, MedLine et UniProt. L'utilisateur peut annoter manuellement ou à partir de sources externes chaque nœud et chaque arc du graphe affiché. Il est possible d'ajouter des extensions à Ondex principalement pour de nouvelles fonctionnalités d'intégration de données.

3.6. Utilisation des réseaux biologiques

Les réseaux d'interaction biologiques forment la base des processus cellulaires. De plus en plus d'études les prennent donc en compte afin de mieux appréhender des phénomènes complexes. Ainsi, ils sont de plus en plus utilisés pour étudier les maladies génétiques et plus particulièrement pour rechercher des gènes candidats pour expliquer ces maladies. Ils sont également particulièrement utiles pour la conception de nouveaux médicaments. En effet, ces réseaux permettent d'identifier et de rechercher des cibles potentielles lors du développement de nouveaux médicaments ou pour repositionner un médicament déjà sur le marché dans une autre indication. Grâce à eux on peut également espérer obtenir une meilleure compréhension des effets indésirables des médicaments.

3.6.1. Étude des maladies génétiques

Les maladies dites génétiques sont causées par une ou plusieurs anomalies dans le génome. L'origine de ces maladies est assez variée : il peut s'agir d'une "simple" modification de la séquence d'ADN (substitution, insertion, délétion) d'un ou plusieurs gènes ou encore d'un réarrangement chromosomique (duplication de séquence génomique, anomalies du nombre de chromosomes ou de structure). Il est possible de séparer ces maladies en plusieurs groupes selon les mécanismes impliqués. Si la maladie à pour origine un gène unique, on parle de maladie monogénique. Les maladies monogéniques peuvent être reparties en 6 sous-groupes selon leur mode de transmission. Le premier d'entre eux concerne les gènes à transmission autosomique dominante. Dans ce cas, le gène est situé sur un chromosome non sexuel et un seul allèle muté est suffisant pour provoquer la maladie. Par exemple, le syndrome de Marfan (MIM :154700) est provoqué par la présence d'un seul allèle muté du gène FBN1. Le second mode de transmission est autosomique récessif, la maladie n'apparaissant alors que si les 2 allèles d'un gène situé sur un chromosome non sexuel sont mutés. Ainsi, deux versions

mutées du gène CFTR sont nécessaire pour provoquer la mucoviscidose (MIM :219700). Le troisième groupe correspond aux gènes à transmission dominante liée au chromosome X. Les maladies associées à ce mode transmission sont provoquées par la présence d'un seul allèle portant la mutation. Les garçons comme les filles peuvent être affectés par ce type de maladie. Cependant, dans certain cas comme le syndrome de Rett (MIM :312750), la maladie est létale chez la plupart des garçons ce qui provoque une prédominance des observations chez les filles. Le quatrième groupe concerne les gènes récessifs situés sur le chromosome X. Du fait de la présence d'un seul chromosome X chez les garçons, ceux ci sont plus fréquemment affectés. L'une des affections récessives liées au chromosome X les plus connues est le daltonisme (MIM :303800) qui implique le gène OPN1MW. Les gènes du chromosome Y sont associés au quatrième groupe. Du fait du faible nombre de gènes sur ce chromosome (50 gènes codant une protéine), seuls 4 phénotypes lui sont associés dans OMIM. Le dernier mode de transmission concerne les gènes mitochondriaux (transmission mitochondriale). Ce mode de transmission est également appelé transmission maternelle puisque les mitochondries sont uniquement transmises par la mère. A ce jour, 28 phénotypes ayant une transmission mitochondriale sont recensés dans OMIM. Il est important de noter qu'une maladie monogénique peut être provoquée par des mutations dans plusieurs gènes différents à condition qu'une seule mutation soit suffisante pour provoquer la maladie. Ainsi, l'amiloïdose VIII (MIM :105200) peut être provoquée par une mutation soit dans le gène FGA, soit dans le gène APOA1, soit dans le gène LYZ.

Une maladie génétique peut également être provoquée par plusieurs gènes ainsi que par l'environnement. On parle alors de maladies multifactorielles. Certaines maladies sont provoquées par la présence de mutation "simultanée" dans deux gènes. Ainsi, Schaffer (2013) recense une cinquantaine de maladies dites digéniques. La première à avoir été mise en évidence est une forme de rétinite pigmentaire (MIM :608133) et implique des mutations dans les gènes PRPH2 et ROM1 (Kajiwara et al., 1994). Dans la plupart des maladies multifactorielles, il existe un grand nombre de gènes pour lesquels certains SNP vont provoquer une plus grande susceptibilité d'apparition de la maladie. Par exemple, Chen et al. (2010), Neale et al. (2010) et Ricci et al. (2013) ont montrés les variations présentes dans 5 gènes sont associés à un plus grand risque de développer une dégénérescence maculaire liée à l'age. L'un des chalenges de l'ère post-génomique est la compréhension des fonctions biologiques de gènes isolés ou des réseaux de gènes qui conduisent à l'apparition de maladies (Chen et al., 2008). L'ensemble des maladies monogéniques peut être représenté sous forme de graphe où les nœuds représentent les maladies (Goh et Choi, 2012). Deux maladies sont alors connectées si elles ont un gène en commun (Figure 2.8). On peut ainsi remarquer que les

maladies humaines sont fortement interconnectées montrant ainsi une origine génétique commune à de nombreuses maladies. Cependant, si plusieurs maladies sont reliées par un gène commun, une maladie est rarement la conséquence de la perturbation d'un seul gène (Barabasi et al., 2011). Ainsi l'étude des réseaux biologiques peut permettre d'identifier de nouveaux gènes responsables de maladies et de mieux comprendre les mécanismes sous-jacents.

a) Recherche de gènes responsables de maladies

A ce jour, il existe environ 5500 maladies monogéniques répertoriées dans la base de données OMIM[11] . Pour environ 30% de ces maladies, les gènes responsables ne sont pas connus. La recherche des gènes responsables d'une maladie se fait par l'établissement d'une liste de gènes candidats. Un gène candidat peut être un gène situé dans la région suspectée d'être responsable de la maladie ou possédant une fonction pouvant expliquer le phénotype de la maladie (Freudenberg et Propping, 2002). Freudenberg et Propping (2002) étendent cette définition en considérant que les maladies similaires peuvent être expliquées par des gènes similaires. Yilmaz et al. (2009) proposent une définition étendue d'un gène candidat comme un gène ayant une relation directe ou indirecte avec la maladie. On considère qu'il a une relation directe si ce gène est "colocalisé" avec la maladie c'est à dire s'il est localisé dans la région chromosomique associée à la maladie, s'il est observé comme étant dérégulé chez les malades ou s'il a une annotation fonctionnelle similaire à celle de la maladie. Par contre, la relation est considérée comme indirecte si un gène intermédiaire intervient. Celui-ci pouvant soit être un interactant soit un orthologue. Ainsi, un gène candidat ayant une relation indirecte avec la maladie peut être un gène dont l'orthologue a une annotation fonctionnelle similaire à celle de la maladie. Des définitions plus complexes peuvent encore être développées. Par exemple, un gène candidat peut être un gène qui a une annotation

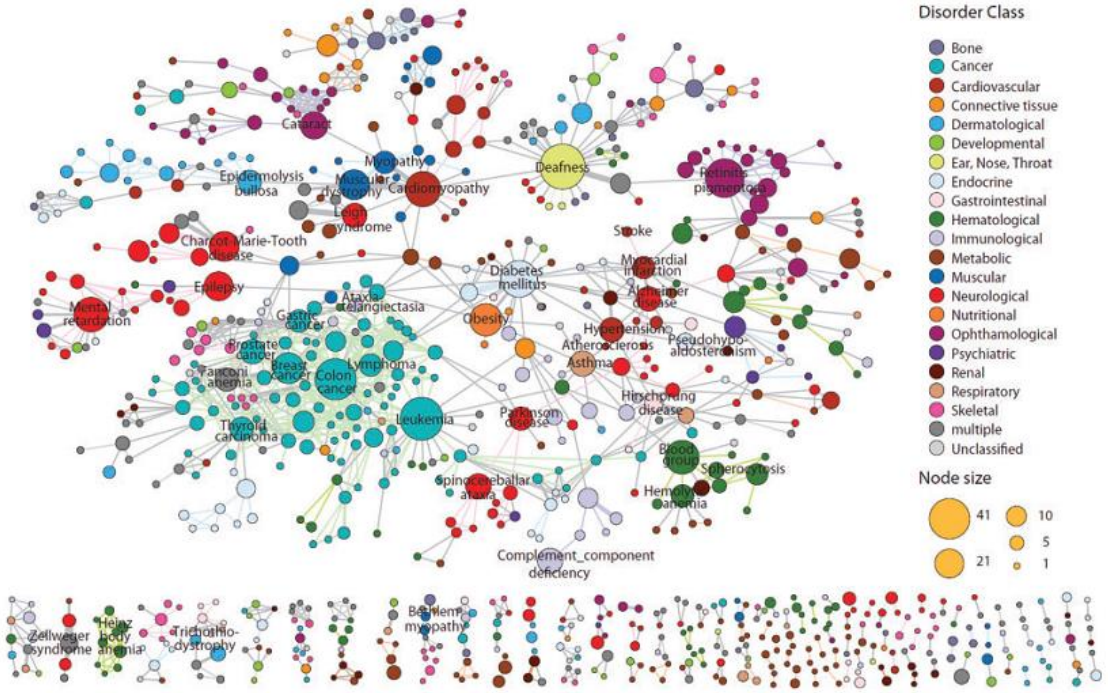


Figure 24 : Représentation du réseau des maladies humaines. Deux nœuds (maladies) sont reliés s'ils partagent un composant génétique selon la liste maladie-gène définie dans OMIM en 2005. La figure est extraite de Goh et Choi (2012).

fonctionnelle similaire à celle de la maladie et établie par des généticiens et qui interagit avec un gène dérégulé dans la maladie, ou encore un gène colocalisé sur le génome avec le locus de la maladie et qui est un orthologue d'un gène ayant une annotation fonctionnelle similaire à la maladie (Yilmaz et al., 2009).

Des méthodes de priorisation permettant de déterminer des gènes candidats ont été développées. Par exemple, le programme ENDEAVOUR utilise un grand nombre de sources de données différentes afin de proposer des gènes candidats (Aerts et al., 2006, Tranchevent et al., 2008). Ainsi, des données d'annotations fonctionnelles, d'interactions protéine-protéine, d'expression, d'orthologie et de régulation sont intégrées. Des scores sont calculés pour chaque type de données puis une étape de "genomic data fusion" est appliquée afin de fusionner les scores de chaque méthode et ainsi obtenir un résultat global. Cependant, ces études basées sur les réseaux d'interaction et les annotations fonctionnelles, sont grandement dépendantes de la qualité d'annotation des gènes (Piro et Di Cunto, 2012). Ainsi, on peut passer à côté d'un bon gène candidat si ce gène possède peu d'annotations. Pour éviter ce problème, Wagner et al. (2013) proposent d'intégrer différentes sources de données afin détecter les gènes associées aux rétinites pigmentaires. Pour cela, ils ont collecté des données de ChIP-seq, mRNA-seq et de microarray issues respectivement de la rétine de souris, de la rétine humaine et de profils d'expressions provenant de 10 tissus oculaires dont la rétine. Ensuite, à partir d'une liste de gènes comprenant des gènes connus pour être à l'origine de rétinite pigmentaires et d'autres gènes, Wagner et al. (2013) réalisent un apprentissage

automatique afin d'apprendre à séparer les deux types de gènes. Ils ont ensuite testé le classifieur ainsi obtenu sur un jeu de données de 13 gènes déjà connus et ont comparé les résultats avec ceux donnés par ENDEAVOUR. Ils montrent ainsi que leur système est plus performant et donc que l'utilisation de données expérimentales est également importante.
Ces deux méthodes de recherche de gènes candidats montrent bien l'intérêt d'utiliser des sources de données variées. Cependant, elles ne fonctionnent pas pour les maladies dont aucun gène responsable n'est connu. En effet, dans ces deux approches il est toujours nécessaire d'utiliser une liste de gènes associés à la maladie étudiée afin de pouvoir générer des candidats.

b) Réseaux biologiques au service de la compréhension des maladies génétiques

Alors que les méthodes de recherche de gènes candidats basées uniquement sur les polymorphismes nucléotidiques ne permettent pas de comprendre les mécanismes provoquant la maladie étudiée, les méthodes basées sur les annotations fonctionnelles ou sur des données d'expression permettent une meilleure compréhension des mécanismes associés à la maladie. En effet, on peut ainsi extraire des processus biologiques ou des gènes dont le fonctionnement est dérégulé lors de la maladie. Les études d'associations pangénomiques ("genome wide association studies" ou GWAS) sont une stratégie populaire pour essayer d'identifier les sources génétiques des maladies multifactorielles Humaines. Ces études consistent à étudier des centaines de milliers de SNP sur un grand nombre de patients et de personnes saines, afin de pouvoir établir des corrélations entre les SNP et la maladie étudiée. Cependant, dans un très grand nombre d'études, seul un petit nombre des corrélations les plus évidentes gènes-maladie sont expliquées (Bakir-Gungor et Sezerman, 2011). BakirGungor et Sezerman (2011) proposent d'utiliser les connaissances sur les réseaux biologiques dont les réseaux d'interactions afin d'enrichir l'analyse des données GWAS. Leur méthodologie est divisée en 3 étapes. La première consiste à associer les SNP aux gènes puis à calculer un score pour chaque SNP dépendant des propriétés fonctionnelles du transcrit, de l'effet du SNP sur la transcription, l'épissage, la traduction et les modifications post traductionnelles. Un score est calculé pour chaque gène en fonction de ses SNP. La deuxième étape consiste à rechercher des sous-réseaux actifs à partir d'un réseau d'interactions protéine-protéine. Un sous-réseau actif correspond à un ensemble de protéines (probablement associées à la maladie) qui interagissent physiquement et ont une forte chance d'appartenir au même réseau biologique. Cette étape est réalisée en utilisant l'extension jActiveModule de Cytoscape qui prend en compte à la fois les scores associés aux gènes et la topologie du réseau pour définir les sous-réseaux actifs (Ideker et al., 2002, Bandyopadhyay et al., 2006). La troisième et dernière étape correspond à l'enrichissement fonctionnel des sous-réseaux. Pour cela, ils recherchent les

termes GO ainsi que les réseaux biologiques KEGG et BioCarta qui sont plus fréquemment retrouvés dans le sous-réseau que dans la population générale. Bakir-Gungor et Sezerman (2011) ont appliqué cette méthodologie à la polyarthrite rhumatoïde. Ils ont ainsi pu associer de nouveaux réseaux biologiques à cette maladie, ainsi que des nouveaux gènes appartenant à des réseaux identifiés comme associés à la maladie. Au lieu d'intégrer interactions et réseaux biologiques, Curtis et al. (2012) ont récemment proposé une approche utilisant des données d'expression pour comprendre les relations qui existent entre les SNP détectés par GWAS et les maladies. Les résultats obtenus sur un ensemble artificiel de données ont montré qu'en plus d'augmenter la capacité de détection des relations gènes-maladies, cette approche permet de mieux comprendre les mécanismes liant les SNP aux maladies en associant les SNP à des dérégulations.

3.6.2. Médicaments, cibles et effets secondaires

a) Étude des cibles de médicaments

Depuis l'avènement de la biologie moléculaire, la recherche de médicaments s'est essentiellement fondée sur l'hypothèse qu'une molécule ne se fixe que sur un seul récepteur. Ainsi les nouveaux médicaments sont le plus souvent des molécules qui se lient spécifiquement sur la cible désirée (Apic et al., 2005). Cependant, cette approche est assez réductrice puisqu'il est maintenant admis qu'une molécule peut se lier à plusieurs cibles (Figure 25, Apic et al. 2005, Yildirim et al. 2007, Scheiber et al. 2009a, Pujol et al. 2010, Vogt et Mestres 2010). Ainsi, la recherche de nouvelles cibles pour des médicaments existants se développe sous le nom de "repositionnement moléculaire". Cette recherche intègre forcement des données sur les réseaux biologiques.

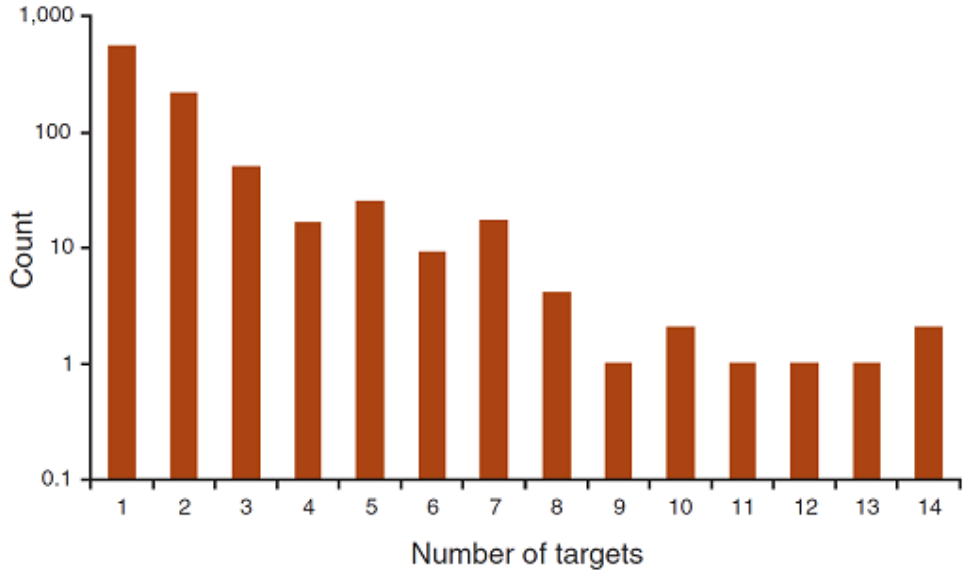


Figure 25 : Distribution des médicaments en fonction de leur nombre de cibles. Les médicaments et leurs cibles proviennent de DrugBank. La figure est extraite de Yildirim et al. (2007).

De nombreuses méthodes de prédiction de cibles de médicaments se basant sur différents types de propriétés ont ainsi été développées. Hopkins et Groom (2002) se basent sur la similarité de séquence entre les cibles pour proposer de nouvelles cibles à des médicaments. En partant du principe que la plupart des médicaments existants se fixent sur le site de liaison d'une molécule endogène, ils montrent que les domaines de liaison des cibles se regroupent en 130 familles et que près de la moitié des cibles représentent seulement 6 familles. Au sein d'une famille de protéines, la séquence et la fonction d'un site de liaison sont généralement bien conservées. Ceci suggère que si un membre d'une famille est capable de se lier à un médicament, alors les autres membres devraient également se lier à ce médicament ou à un autre structurellement similaire. Au lieu de se baser uniquement sur la séquence des cibles, Li et Lai (2007) prennent en compte les propriétés de cette séquence protéique. Chaque protéine est décrite par un vecteur à 146 dimensions représentants propriétés suivantes : composition en acide aminés, hydrophobicité, polarité, polarisabilité, charge, accessibilité au solvant, volume de van der Waals normalisé. Ce vecteur est calculé pour chaque protéine de 2 ensembles : un ensemble de protéines cibles et un autre ensemble de protéines non cibles. Ces deux ensembles servent de base d'apprentissage à un SVM (Support Vector Machine) qui permettra de reconnaitre les protéines qui peuvent servir de cibles à des médicaments. A l'issu d'un test de validation croisée (apprentissage sur 9/10 du jeu de donnée et test sur le dixième restant, répété pour les 10 dixièmes), le SVM obtient une précision de l'ordre de 84%. Les tests menés sur des jeux de données n'ayant pas servi à l'apprentissage ont confirmé les résultats obtenus montrant ainsi l'intérêt d'utiliser les propriétés de la séquence des cibles. Afin de mieux comprendre les caractéristiques des cibles de médicaments, Ma'ayan et al. (2007) ont notamment étudié leur enrichissement en termes GO. Ainsi, à partir de 485 cibles issues de la base de données DrugBank, ils ont remarqué que ces cibles étaient enrichies en termes GO associés aux protéines membranaires, aux récepteurs, aux facteurs de transcription et aux composants des voies de signalisation cellulaire. Ainsi, les propriétés fonctionnelles des cibles semblent partagées par un grand nombre d'entre elles pouvant ainsi servir à déterminer de nouvelles cibles à des médicaments existant. Yao et Rzhetsky (2008) se sont intéressés à d'autres propriétés des cibles. Ils se sont notamment posé la question de savoir si les cibles des médicaments correspondent à des gènes plus sujets à des polymorphismes que la moyenne des gènes humains. A partir d'une liste d'environ 16 000 gènes, ils ont montré que les gènes "cibles" possèdent moins de polymorphismes. Ils expliquent cette observation

par le fait qu'un grand nombre de polymorphismes dans une protéine cible pourrait diminuer la capacité du médicament à interagir avec cette cible. Yao et Rzhetsky (2008) ont également analysé les relations entre les médicaments et les tissus d'expression de leurs cibles. Ils ont ainsi remarqué que cinq tissus sont fréquemment ciblés (grandes endocrines, système nerveux central, l'appareil urinaire, les glandes excrétrices et les ganglions) alors que d'autres ne le sont que très rarement, notamment les tissus embryonnaires. Ce faible taux de ciblage de certains tissus pouvant s'expliquer par les risques importants d'effets secondaires liés à ces tissus. Par ailleurs, en étudiant le nombre de réseaux biologiques associés aux cibles des médicaments approuvés ou non par la "Food and Drug Administration" (FDA) (Cf. Figure 26), Sakharkar et al. (2008) montrent que le fait de cibler une protéine intervenant dans de multiples processus n'est pas une situation préférentielle pour un médicament Afin de prédire de nouvelles associations protéine médicament, Cheng et al. (2012) utilisent une méthode basée sur les réseaux d'interactions. Ils utilisent un score basé sur la topologie du réseau médicament-cible. De cette façon, ils obtiennent un score pour chaque couple médicament nouvelle protéine permettant ainsi d'obtenir une liste de cibles candidates à tester. Cheng et al. (2012) ont ensuite validé leur approche en testant et en confirmant in vitro leurs résultats pour 5 molécules et 2 cibles.

Les approches les plus originales pour la recherche de nouvelles cibles sont certainement celles de Campillos et al. (2008) et Takarabe et al. (2012). Ces deux approches se basent sur les effets secondaires des médicaments pour prédire de nouvelles cibles, leur hypothèse étant que le partage d'effets secondaires s'explique par des cibles communes. Campillos et al. (2008) se sont basés sur les notices des médicaments afin d'extraire les effets secondaires. Une valeur de similarité entre

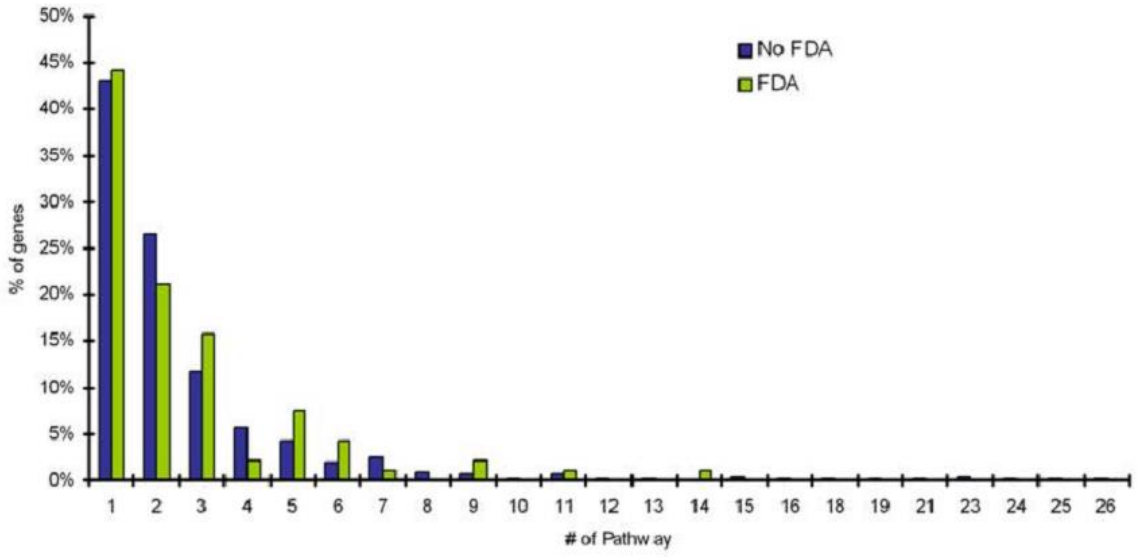


Figure 26 : Distribution du pourcentage de gènes cibles selon leur nombre de réseaux biologiques extraits à partir de SwissProt. La figure est extraite de Sakharkar et al. (2008).

les effets est ensuite calculée en se basant sur la co-ocurence des effets dans les médicaments (deux effets sont considérés comme similaires s'ils sont fréquemment associées aux mêmes molécules). Cette similarité en effets secondaires est ensuite combinée avec une mesure de similarité chimique à 2D afin de détecter de nouvelles cibles. Cette méthode a été testée sur 746 molécules et a permis de former 1018 couples de médicaments similaires fortement susceptibles de partager des cibles. Une vingtaine de ces couples découverts ont été vérifiés in vitro et pour 13 d'entre eux, des cibles communes ont été identifiées, confirmant ainsi l'intérêt de la méthode. L'approche utilisée par Takarabe et al. (2012) se base sur le système de signalement d'effets secondaire de la FDA (adverse event reporting system, AERS). AERS est un système permettant aux professionnels de la santé de rapporter les effets indésirables des médicaments observés chez des patients et ainsi d'mettre en évidence les effets secondaire de manière plus précoce que sur les notices. En utilisant

AERS on est capable d'observer la présence d'un effet indésirable pour une molécule quand au moins un rapport en fait mention. Contrairement à cela, l'industrie pharmaceutique ne fait mention d'un effet secondaire pour une molécule que si cet effet a été observé un grand nombre de fois. Ainsi, Takarabe et al. (2012) ont comparé les prédictions de cibles basées sur une similarité d'effets secondaires AERS avec une méthode de prédiction basée sur la structure des molécules. Les résultats de cette comparaison montrent que l'utilisation des effets secondaires donne de bons résultats mais moins bons que ceux basés sur la structure. Cette observation suggère que l'utilisation des effets secondaire pour prédire les cibles peut-être utile quand on ne possède pas de données sur la structure des molécules, notamment comme cela arrive pour les peptides utilisés comme médicaments.

b) Étude des effets secondaires

De même qu'un médicament peut avoir plusieurs cibles, une cible peut intervenir dans plusieurs voies biologiques. L'utilisation des réseaux biologiques permet de replacer les cibles dans leur contexte et donc de mieux comprendre les effets indésirables de ces molécules (Pujol et al., 2010). L'étude des effets secondaires est donc facilitée par l'utilisation de données des réseaux biologiques. Le lien entre la structure des molécules et leurs mécanismes d'action ayant été établi (Martin et al., 2002), les premières approches de prédiction des effets secondaires se sont basées sur les sous-structures des molécules (Bhavani et al., 2006, Scheiber et al., 2009b, Atias et Sharan, 2011, Pauwels et al., 2011). Dans ces approches, les molécules sont décrites à la fois par un vecteur contenant leur composition (sous-structure) et

un autre vecteur contenant leurs propriétés (effets secondaires). Ensuite, pour chaque effet secondaire, un apprentissage est effectué pour reconnaitre les molécules ayant cet effet. De plus ces méthodes ont permis d'identifier des sous-structures responsables des effets étudiés (Figure 27).

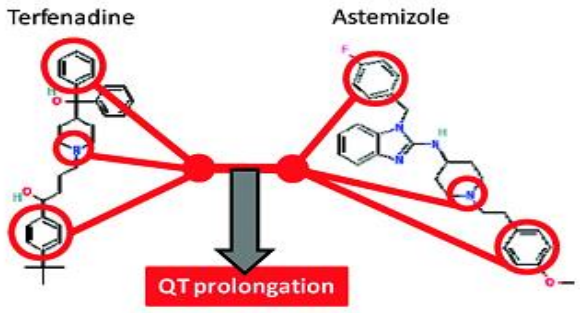


Figure 27 : Sous-structures moléculaires identifiées par Scheiber et al. (2009b) comme étant responsables d'une prolongation QT (arythmie cardiaque).

Comme Campillos et al. (2008) l'ont souligné, il existe des relations entre les effets secondaires et les cibles des médicaments. Afin de prendre en compte l'aspect molécule et l'aspect cible, Yamanishi et al. (2012) ont développé une méthode de prédiction des effets secondaires basée sur l'espace chimique et l'espace biologique. Comme précédemment, l'espace chimique correspond à une description des sous-structures des molécules. Quant à l'espace biologique, il est caractérisé par les cibles des médicaments (Figure 27). L'apprentissage nécessaire à la prédiction est réalisé sur 658 médicaments annotés par 969 effets secondaires et décrits par un vecteur chimique de 881 sous-structures et un autre vecteur de 1368 cibles. Les résultats de tests obtenus montrent l'intérêt relatif de combiner à la fois des données biologiques et chimiques pour prédire les effets secondaires (Tableau 4).

Au lieu de se baser sur la structure des médicaments, d'autres études ont préféré prendre en compte les processus biologiques associés à ces derniers (Lee et al., 2011, Huang et al., 2011). Ainsi, Lee et al. (2011), ont utilisé les processus biologiques associés aux médicaments par la ressource "Connectivity map" (Lamb et al., 2006). Ces données ont permis de déterminer quels processus étaient dérégulés par les médicaments. Lee et al. (2011) proposent ensuite que les médicaments partageant des effets secondaires partagent également des processus biologiques (Figure 28). Cette approche a été testée sur l'effet secondaire éruptions cutanées ("rash") et les résultats

Descripteurs	AUPR
Aléatoire	0.04
Chimie	0.18
Biologie	0.19
Chimie + biologie	0.21

Tableau 4 : Performance de prédiction basée sur une cross validation à 5 itérations. AUPR (Area Under the Precision-Recall curve) est la surface sous la courbe précision-rappel. La ligne aléatoire correspond aux résultats attendu si la classification est réalisée de manière aléatoire. Les données sont extraites de Yamanishi et al. (2012).

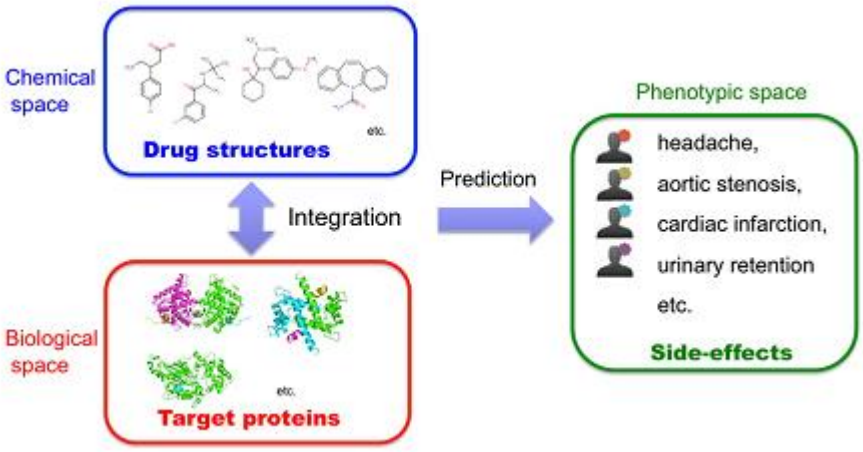


Figure 28 : Méthodologie proposée par Yamanishi et al. (2012) pour prédire les effets secondaires.

ont été comparés à ce qui était déjà connu dans la littérature. Cela a permis monter qu'un grand nombre des processus biologiques découverts par leur méthode étaient déjà reportés comme liés à l'effet secondaire dans la littérature. Au lieu de se baser sur les termes GO associés directement

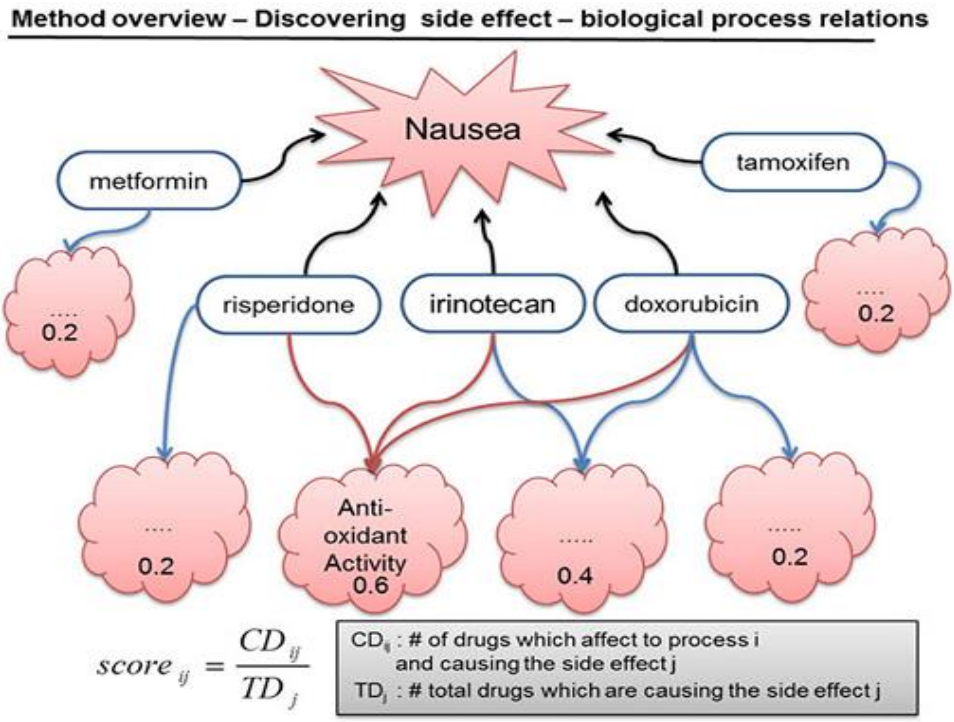


Figure 29 : Méthodologie proposée par Lee et al. (2011) pour associer effets secondaires et processus biologiques.

aux médicaments comme l'ont fait Lee et al. (2011), Huang et al. (2011) prennent en compte les cibles ainsi que leurs termes GO. Ainsi, de manière similaire à Lee et al. (2011), ils montrent que les fonctions associées aux cibles sont fortement corrélés l'apparition d'effets secondaires. Les travaux de Liu et al. (2012b) sont particulièrement intéressants car ils combinent les caractéristiques phénotypiques des médicaments (indication et effets secondaires autres que celui étudié), la structure chimique et des propriétés biologiques (cibles et réseaux biologiques). Cette étude a permis de montrer que la combinaison de toutes ces données permet d'augmenter la qualité des prédictions. Il est cependant important de noter que l'utilisation des effets secondaires pour en prédire d'autres augmente fortement les résultats de la prédiction.

Chapitre II: Démarche

I. Production de souris transgéniques et CD36

Bien avant le développement des techniques de génétique moléculaire, les pathologistes cherchaient à explorer les mécanismes biologiques associés à des maladies héréditaires touchant l'homme ou l'animal. Au début, cette exploration n'a pu se faire que par le biais de l'utilisation de souches mutantes que l'on peut qualifier de naturelles (Roths et al., 1999). Puis grâce au développement des techniques de transgénèse, le concept de créer des souches animales servant de modèle à des maladies pouvant toucher l'homme ou l'animal est devenu une réalité. Ainsi par l'utilisation de ces méthodes, l'étude in vivo de l'effet de gènes particuliers sur le métabolisme ou la physiologie a pu devenir une réalité.

Le terme de « transgénique » est devenu, ces dernières années, de plus en plus présent que ce soit dans la littérature scientifique ou dans l'actualité. Il est apparu pour la première fois dans une publication scientifique en 1981 où il était utilisé pour qualifier des souris obtenues suite à une micro-injection de gènes dans un des pronuclei (Gordon et Ruddle, 1981). Le terme de « transgénique » s'applique à des organismes vivants, qu'ils appartiennent au règne végétal ou animal, dans lesquels ont été transférés des gènes étrangers d'origine animale ou végétale à leur patrimoine héréditaire propre. Le gène étranger, appelé « transgène », se transmet à la descendance selon un mode mendélien (Kernbaum, 1998). Dans un récent rapport de l'ECVAM (European Center for the Validation of Alternative Methods ou centre européen pour la validation des méthodes alternatives) sur l'utilisation des animaux transgéniques dans l'Union Européenne, une définition des techniques de transgénèse est également donnée : « la technique de transgénèse implique l'introduction de matériel génétique fonctionnel (ADN) au niveau de cellules germinales d'un organisme » (Ben Mepham et al., 1998). Le fait qu'il soit dit que l'ADN est introduit au niveau des cellules germinales sousentend la notion de transmission héréditaire de cette nouvelle information génétique. Mais la définition la plus récente des animaux ou végétaux transgéniques est suggérée par Beardmore : ce sont des « organismes contenant des séquences intégrées d'ADN cloné (transgène), transférées en utilisant des techniques de génie génétique » (Beardmore, 1997). Cette définition est la plus générale et elle fait directement référence à la science à la base de cette révolution : le génie génétique.

L'obtention d'animaux transgéniques repose, quelque soit la technique utilisée, sur les quatre mêmes principales étapes (Cf. Figure 30)

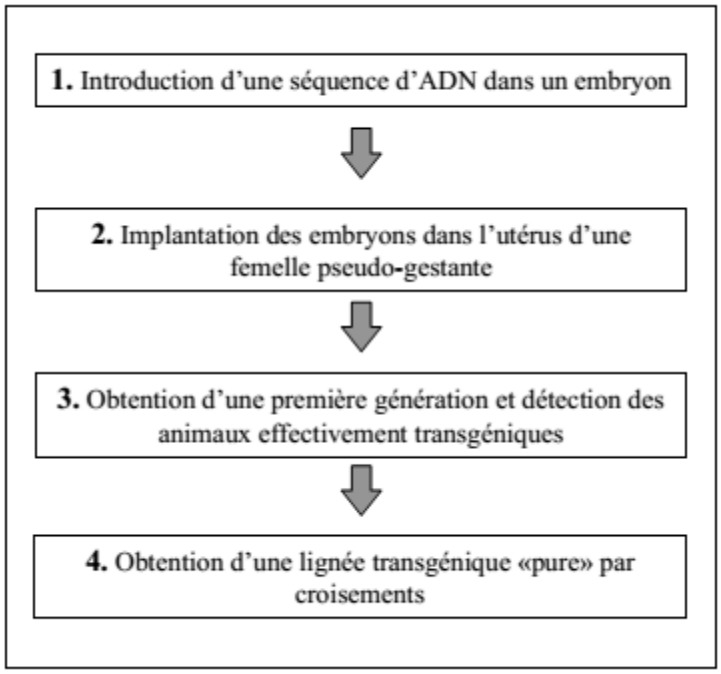


Figure 30: Les principales étapes de l'obtention d'animaux transgéniques

Ces quatre étapes sont bien sûr précédées d'une étape de biologie moléculaire qui consiste en une étape d'isolement du gène d'intérêt (identification, séquençage éventuel).

Chez la souris, trois techniques de transgénèse (ou plus exactement trois techniques d'introduction d'une séquence d'ADN dans un embryon) sont le plus communément utilisées (voir Figure 31) (Jallat, 1991) :

- la micro-injection pro-nucléaire d'ADN cloné dans des oeufs fertilisés
- utilisation de vecteurs rétroviraux
- transfert ciblé de gènes dans les cellules embryonnaires via la recombinaison Homologue

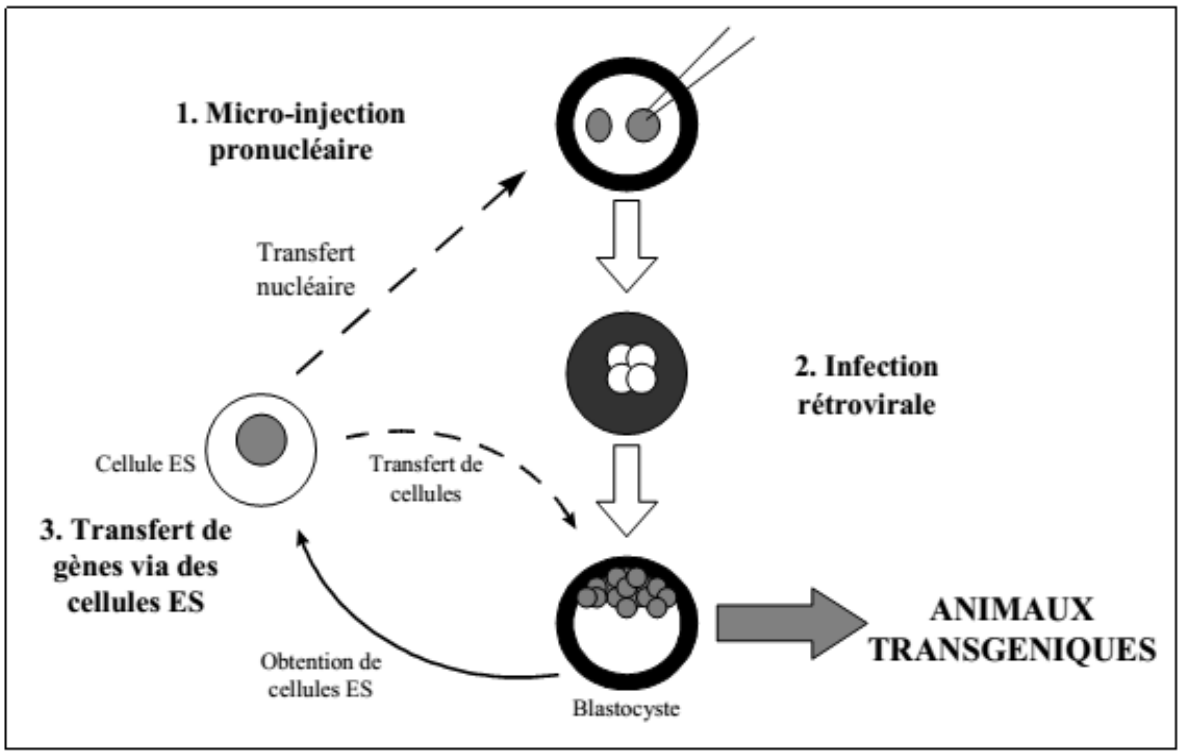


Figure 31: Les 3 principales méthodes de transgénèse (d'après (Janne et Alhonen, 1996)

D'autres méthodes sont également possibles mais elles sont encore peu utilisées (introduction de gènes dans les spermatozoïdes, injection de régions chromosomiques microdisséquées ou technique de ciblage par le système Cre-lox). Parmi les nouvelles méthodes, celle utilisant le système Cre-lox est la plus prometteuse car elle permet un ciblage tissulaire.

Nous étudierons en détail la méthode de la micro-injection pro-nucléaire d'ADN cloné dans des oeufs fertilisés utilisée pour la génération des souris GR

1. Microinjection pronucléaire

1.1. Principe

Cette méthode correspond à la méthode princeps, c'est-à-dire celle utilisée par l'équipe de Gordon lors de leur première expérience de transgénèse (Gordon et al,., 1980). C'est également la première méthode qui s'est révélée efficace chez les mammifères. Elle consiste à injecter des fragments d'ADN purifié (sous forme d'une solution) dans un des pronucléi (en général le pronucléus mâle car il est à la fois le plus grand et le plus proche de la surface) d'un oeuf fertilisé au stade unicellulaire. Ensuite, ce zygote est réimplanté dans l'utérus d'une femelle pseudo-gestante où il poursuit son développement.

1.2. Les différentes étapes (Jallat, 1991)

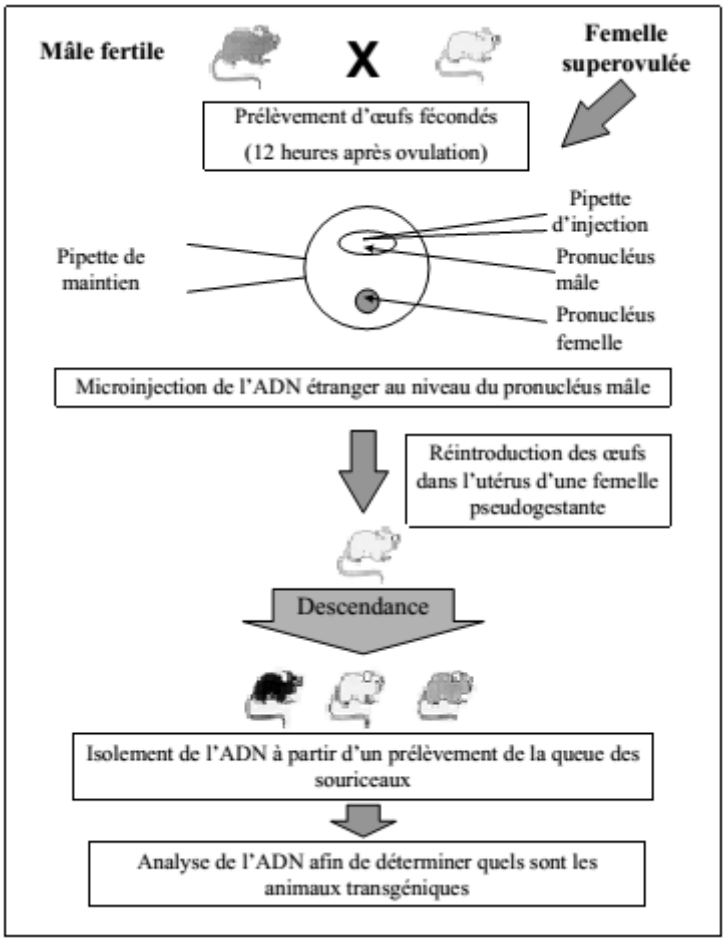


Figure 32: Les principales étapes principales de l'obtention de souris transgéniques par micro-injection pronucléaire(d'après (Jallat, 1991)

1.2.1. 1ère étape : Préparation de la solution d'ADN injectée

Une séquence d'ADN déterminée est clonée (c'est-à-dire reproduite à l'identique un grand nombre de fois) ce qui permet l'obtention d'une solution d'ADN. Cette solution sera ensuite microinjectée au niveau d'un des pronucléi. L'efficacité de la micro-injection dépend des propriétés et de la qualité (e.g. pureté) de la solution injectée (Gordon et Ruddle, 1981). Il a été montré que plus le nombre de copies d'ADN est grand, plus le taux d'intégration de l'ADN est important. Cependant, une préparation trop concentrée donc trop visqueuse peut augmenter le risque de mort de l'embryon au moment de l'injection. De plus, l'ADN linéaire s'intègre plus efficacement que l'ADN circulaire (Brinster et al., 1985). De plus, l'effet inhibiteur de séquences procaryotiques (provenant des vecteurs de clonage) a été démontré : ils n'ont pas d'effet sur l'intégration du transgène dans le génome hôte mais ils peuvent cependant inhiber l'expression de certains transgènes (CHADA et al., 1985),(Townes et al., 1985). Il est donc recommandé d'éliminer toute séquence inutile, notamment les séquences provenant du vecteur de clonage au moyen d'enzymes de restriction.

Mais certaines séquences, au contraire, se sont révélées importantes dans la régulation de l'expression du transgène : c'est le cas des introns (qui sont de larges séquences non codantes) qui ont une influence importante sur le fonctionnement des gènes introduits par micro-injection (Brinster et al., 1988). Les observations de Brinster suggèrent que les introns facilitent la transcription des gènes micro-injectés, notamment en aidant à maintenir une activité de transcription pendant le développement. Il est donc recommandé d'injecter l'ADN sous sa forme génomique plutôt que sous sa forme d'ADNc (ADN complémentaire qui est obtenu par transcription reverse à partir de l'ARN messager et qui ne comprend pas les introns).

1.2.2. 2ème étape : Récolte des œufs fertilisés

Le choix de la souris donneuse est important : il est nécessaire de choisir une lignée avec un bon rendement en œufs, mais il est également essentiel que ces œufs survivent bien à la micro-injection. Des femelles prépubères sont stimulées hormonalement par deux injections intrapéritonéales successives d'hormones à 46-48 heures d'intervalle (il est impératif de respecter ce délai) :

- une première injection de PMSG (Pregnant Mare Serum Gonadotropin : gonadotropine sérique). Cette hormone en se substituant à l'hormone gonadotrope hypophysaire à effet FSH (Folicle-Stimulating Hormone ou hormone follicotrope) stimule le développement des follicules mûrs.
- une seconde injection de hCG (human Chorionic Gonadotropin : gonadotropine chorionique humaine). Cette hormone, en mimant l'effet de la LH (Luteinizing Hormone ou hormone lutéinisante) parachève la maturation folliculaire et favorise l'ovulation. Suite à ce traitement, un accouplement entre ces femelles et des mâles fertiles est réalisé. L'apparition, après quelques heures, d'un bouchon vaginal, chez 80 à 90% des femelles, est le signe que l'accouplement a effectivement eu lieu. Il a été montré que le rendement est plus élevé dans le cas d'œufs hybrides (c'est-à-dire issus d'un croisement de 2 souches pures) plutôt que dans le cas d'œufs de souche pure. Brinster et son équipe ont montré que ce rendement était huit fois plus élevé avec des souris hybrides (C57BL/6xSJL) qu'avec des souris C57BL/6) (Brinster et al., 1985).

Douze heures après l'accouplement, les femelles sont sacrifiées et les oviductes contenant habituellement chacun entre 10 à 15 oeufs sont prélevés. Ainsi, entre 20 et 30 œufs peuvent être récoltés par souris. Ces œufs fécondés peuvent être conservés pendant 3 à 36 heures dans un milieu de culture. Ils sont incubés à 37°C dans une atmosphère à 5,4% de CO2 sur un milieu avec un pH équilibré. Les œufs subissent ensuite un traitement avec une hyaluronidase qui les libère de leur gaine de cellules folliculaires. Puis finalement ils sont « lavés » par

passages successifs dans le milieu de culture afin d'éliminer toute trace résiduelle d'enzyme pouvant avoir un effet néfaste pour eux.

1.2.3. 3ème étape : Micro-injection

Cette étape nécessite à la fois une technique particulière et un appareil spécial. Le coût de l'équipement d'un laboratoire de micro-injection (avec un poste de micro-injection) est d'environ 80 000 $ soit un peu plus de 87 000 euros (Lake et al., 1999). Cette technique nécessite de la dextérité et de l'entraînement pour la personne qui la réalise et se fait sous microscope inversé (Cf. Figure 33).

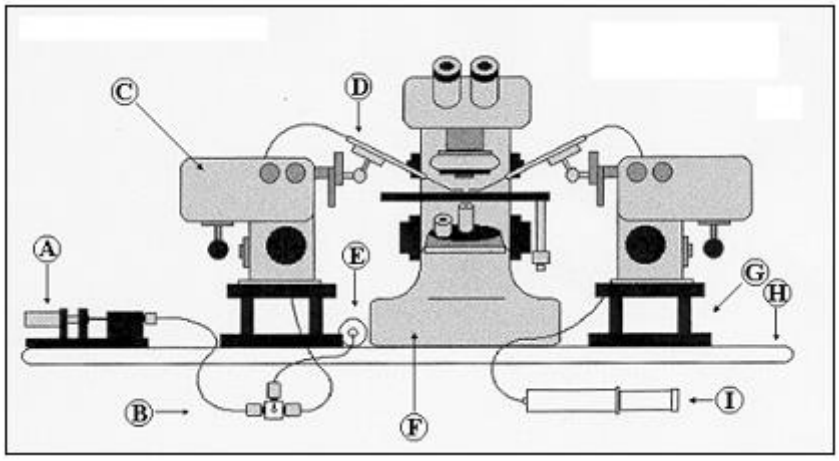


Figure 33: Schéma d'un appareil pour micro-injection

A. Micromètre avec vis hydraulique (Micrometer / Hydraulic Drive Unit)
B. Robinet 3 voies (3-way stopcock valve)
C. Micromanipulateur (Micromanipulator)
D. Porte-aiguilles avec micro-injecteur et micropipette de contention (Needle holder / Instrument collar)
E. Seringue de 60 ml avec embout (60 ml luer end syringe)
F. Microscope inversé (Inverted light stereomicroscope)
G. Support du micro-manipulateur (Micromanipulator stands)
H. Plateau de base ("Bread-board" instrument base)
I. Seringue en verre de 60 ml (Glass 60 ml syringe)

La micro-injection se fait généralement une fois que le pronucléus mâle est visible et observable. L'intégration du transgène se fait vraisemblablement lors de la réplication de l'ADN. Cependant en raison de la dégradation de l'ADN, il est préférable de ne pas réaliser la micro-injection trop tôt avant la réplication. En outre, il paraît judicieux de ne pas microinjecter après la première réplication (c'est-à-dire au stade 2 cellules soit 11 à 20 heures après la fécondation) (Houdebine, 1998), (Monnereau, 1997). Mais dans la pratique, aucun moment ne paraît nettement à privilégier pour le faire (Gagne et al., 1997). Ainsi, il semblerait préférable d'injecter avant la première réplication, c'est en tout cas le moment choisi par la plupart des expérimentateurs (Houdebine, 1998). Quelques œufs sont déposés dans une boite de Piètri contenant une goutte de milieu recouverte d'huile de paraffine, par

exemple, afin d'éviter toute évaporation. Une micro-pipette de contention (ou holding pipette) permet le maintien et le positionnement correct des œufs par aspiration. Une autre micro-pipette permet l'injection de l'ADN préalablement préparé. La manipulation de ces deux micro-pipettes se fait grâce à des micro-manipulateurs.

A ce stade de développement de l'œuf, les deux pronuclei sont bien visibles. Le pronucléus mâle étant le plus gros et le plus près de la surface, la micro-injection se fait à ce niveau (voir Figure 34). Un picolitre de liquide, contenant entre 100 et 1000 copies du transgène, est alors injecté : le nucleus se dilate avec doublement du volume.

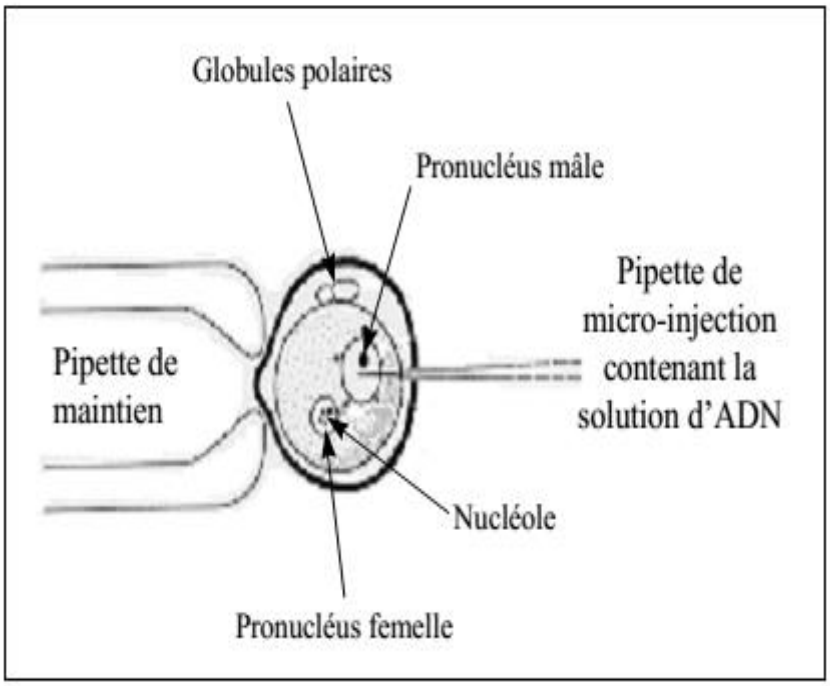


Figure 34: Micro-injection pronucléaire

1.2.4. 4ème étape : Réimplantation

Les œufs sont réimplantés dans une mère porteuse mais pour que le développement embryonnaire se déroule correctement, le terrain hormonal de ces femelles doit être adéquat. Pour cela, la veille de la micro-injection, des femelles en oestrus sontaccouplées avec des mâles infertiles. Ces mâles ont été préalablement stérilisés par vasectomie (section chirurgicale du canal déférent). Il est également possible d'utiliser des mâles génétiquement stériles. Les femelles présentant le lendemain un bouchon vaginal feront partie du pool de mères porteuses. La réimplantation est une opération chirurgicale, sous anesthésie générale, réalisée sous un microscope. Les œufs sont introduits au niveau de l'ampoule de l'oviducte (20 œufs si l'implantation se déroule le jour de la micro-injection, environ 15 œufs si ils ont été conservés in vitro). Le temps de gestation est de 20 jours à partir de la date de l'implantation. Mais le nombre d'œufs se développant à terme est faible et les mères porteuse

n’ayant pas mis bas au 20ème jour seront sacrifiées. Ensuite les souriceaux (environ 5 à 6 par portées) sont adoptés par des femelles nourrices (choisies en fonction de leurs bonnes caractéristiques maternelles).

1.2.5. 5ème étape : Identification des souriceaux transgéniques

Une fois que la première génération de souris est obtenue, il est nécessaire de savoir quelles sont les souris ayant intégré le transgène. Pour cela, une petite quantité d'ADN des animaux est nécessaire : en général, des fragments de queue sont prélevés sur les animaux encore jeunes. Ensuite les transgènes sont identifiés par la technique de transfert de Southern (ou Southern-Blot) ou par PCR (Polymerase Chain Reaction : réaction de polymérisation en chaîne (Houdebine, 1998)) sur ces fragments. Pour la technique de Southern, l'ADN est déposé sur un filtre de nitrocellulose ou de nylon et une hybridation avec une sonde marquée et spécifique du transgène est réalisée. Mais le taux de faux-positifs pour cette méthode est important car il est nécessaire que l'ADN utilisé soit d'une pureté suffisante. De plus, la détection des animaux mosaïques peut être difficile en raison du faible nombre de copies du transgène (Houdebine, 1998). Dans le cas de la PCR, la quantité d'ADN peut être plus faible : la salive (IRWIN, 1996) ou des fragments obtenus suite au marquage des animaux peuvent être suffisants (Ren et al., 2001). De plus, cette méthode peut ne pas être pratiquée à partir d'ADN purifié car une étape d’amplification de l’ADN est incluse dans la technique (SAIKI et al., 1985). Elle est également plus sensible et permet donc de détecter des animaux mosaïques. Ainsi la PCR apparaît comme la méthode la plus simple et la plus rapide.

1.3. Résultats

Le but poursuivi dans toutes les techniques de transgénèse est que le transgène se transmette à la descendance de manière stable et qu'il s'exprime. Comme nous l'avons vu précédemment, dans certains cas, le transgène n’était pas exprimé alors qu'il était bien présent ; c'est le cas dans l'expérience princeps de cette méthode (Gordon et al,., 1980). Il est donc nécessaire de comprendre le système d'intégration de l'ADN. Suite à la micro-injection, l'ADN linéaire injecté se circularise et puis se relinéarise de façon aléatoire et finalement se recombine par recombinaison homologue : il y a formation de concatémères en tandem (Cf. Figure 35).

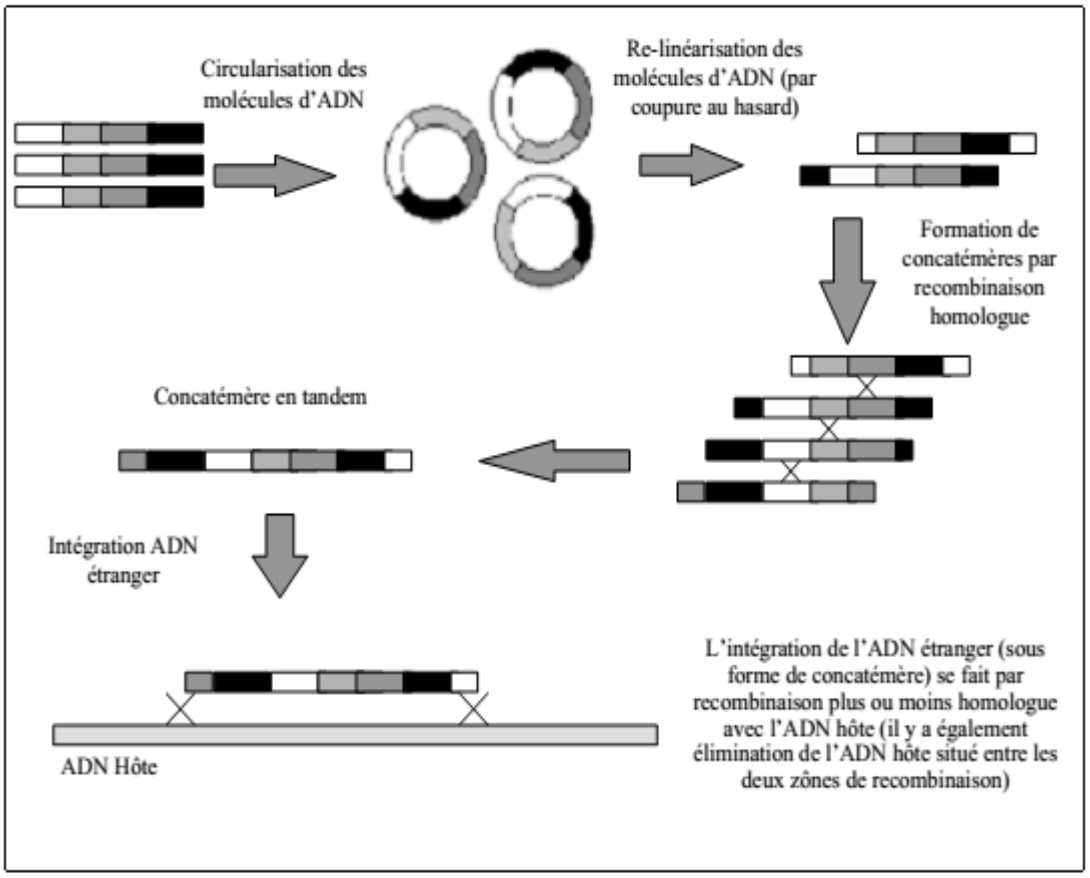


Figure 35: Le mécanisme de formation de concatémères en tandem(d'après (Houdebine, 1998),)

Ces concatémères sont ensuite soit dégradés, soit intégrés au génome en faisant intervenir un mécanisme d'intégration hétérologue. Ce phénomène se fait grâce à une reconnaissance partielle de l'ADN génomique par l'ADN étranger. Pour obtenir des lignées pures, il serait souhaitable que l'intégration de l'ADN se réalise au stade d'une cellule. Cependant, ce n'est pas toujours le cas, et si l'intégration se fait à un stade plus tardif, le taux de souris mosaïques sera relativement élevé (Houdebine, 1998). Il est possible que l'ADN micro-injecté s'insère aux niveaux de certains sites conduisant alors à différentes mutations pouvant être létales, notamment dans le cas où l'insertion provoque une interruption au sein d'un gène ayant une fonction vitale ou participant au développement (Bishop, 1997),(Krulewski, 1989).

En règle générale, sur 100 embryons manipulés, une à trois (cinq au grand maximum) souris transgéniques peuvent être obtenues. Ce faible rendement s'explique notamment par le fait que peu d'embryons survivent à la micro-injection et au transfert dans l'utérus de femelles pseudo-gestantes (environ 20 à 30%) (Houdzbine, 1999). Mais l'efficacité varie également selon la construction d'ADN utilisée (Gordon, 1997). Chez les autres espèces, le rendement est moindre, ce qui peut expliquer pourquoi la transgénèse animale est réalisée à l'heure actuelle en grande majorité sur des souris (Houdebine, 1998),

1.4. Avantages et inconvénients

1.4.1. Avantages

L'avantage principal de la micro-injection est son efficacité à produire des lignées transgéniques qui expriment la plupart du temps les gènes de manière prévisible (Jaenisch, 1988). Le travail de biologie moléculaire est assez simple (comparé aux autres techniques) : le gène que l'on souhaite insérer dans le génome doit être identifié puis cloné afin d'avoir un nombre suffisant de copies à injecter (Gordon, 1997).. En raison du faible temps de gestation de la souris et du peu de technicité de la phase de préparation de l'ADN, le temps de développement est réduit : entre 3 et 9 mois (Jaenisch, 1988).

1.4.2. Inconvénients

Cependant cette technique ne permet que l'addition d'un gène (ni la délétion, ni la substitution ne sont possibles). L'intégration se faisant au hasard, il est difficile de prévoir les effets secondaires. Il est donc nécessaire de prendre en compte les effets du locus d'insertion sur le patron d'expression. A cet effet s'ajoute également le fait que le nombre de copies d'ADN insérées est variable (en raison de la formation de concatémères) (Jaenisch, 1988). Même si les gènes micro-injectés s'expriment de façon efficace, il se peut que le niveau d'expression soit sous la dépendance de différents facteurs comme des facteurs procaryotiques provenant des vecteurs utilisés qui peuvent inhiber l'expression du transgène (Gordon, 1997). La dernière contrainte est la précision nécessaire à la technique de micro-injection, qui requiert de la dextérité de la part du technicien et donc un entraînement. Malgré ces quelques défauts, la technique de micro-injection pronucléaire est la plus efficace des techniques et également la plus utilisée pour obtenir des animaux transgéniques. Des fragments d'une longueur d'environ 50 kilobases peuvent être ainsi introduits dans le génome d'un œuf et ainsi ils seront exprimés à la fois dans les cellules somatiques et les cellules de la lignée germinale.

Rescue Of CD36 Expression In Skeletal Muscle Impacts Endurance During Physical Exercise

Imane Sabaouni[1], Nada A. Abumrad[2] and Azeddine Ibrahimi[1]

[1] Biotechnology Lab (MedBiotech), Faculté de Médecine et de Pharmacie de Rabat, University Mohammed V in Rabat, Morocco.

[2] Department of Physiology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, 6200 MD Maastricht, Netherlands.

* **Corresponding author :** Email: im_sabouni@yahoo.fr

Running head:
Role of CD36 protein expression in FA metabolism of skeletal muscle

Abstract

CD36 is a membrane protein that facilitates long chain fatty acid (FA) transport in muscle tissues. The role of its expression in overall metabolism and its impact on muscle and heart function was examined by studying a newly generated mouse model CD36-KO mice (GR model). This model harbors the CD36 gene under the control of the creatine kinase promoter, allowing its expression in muscles and heart.

Specific expression of the CD36 protein was confirmed by western blot analysis and immunofluorescence labeling. Unlike CD36 null mice, GR mice had no fasting hypoglycemia, and their plasma triacylglycerol and free fatty acid levels, which are increased in the KO mice, were restored to WT levels by muscle-targeted CD36 re-expression.

Moreover, CD36 null mice subjected to forced swimming tests showed a drastically reduced ability to endure exercise in comparison to wild type and GR mice. This suggests a decreased glycogen and triacylglycerol levels of CD36-deficient muscles and their impaired ability to oxidize palmitate. Fatty acids oxidation rates were significantly higher in GR muscles than in CD36-null muscles but comparable to the wild type rates. Stimulation by dipyridamole increased oxidative rates in wild type and GR muscles, but did not affect oxidation in KO muscle, confirming a role in metabolism.

The present study showed that muscle and heart performances were affected by CD36 expression. Results shed some light on the role of genetic and environmental alterations on CD36 expression and its impact on human performance and athletic ability.

Keywords: CD36, Fatty acid transport, Fatty acid oxidation, exercise, endurance

Introduction

Early work with isolated rat adipocytes indicated that beside the simple diffusion of long chain fatty acids (FA), the existence of a facilitated transport mechanism [1]. A heavily glycosylated 88-kDa integral membrane protein was identified as a candidate FA transporter using reactive sulfo-*N*-succinimidyl derivatives of long–chain FA [2]. The corresponding cDNA, isolated from a rat adipose tissue library, encoded FAT (for FA translocase) [3], which is the rat homologue of the human platelet CD36, also known as PASIV, GPIV and GPIIIb in earlier literature [4]. Tissues with active FA metabolism expressed CD36 with abundance, correlating with their presumed role in FA uptake [3, 5, 6,7]. In vivo, CD36 expression was regulated in altered lipid metabolism conditions such as type 2 diabetes and high fat diet [8, 9].

The physiological role of the protein CD36 was established by the generation of the CD36 KO mice model [10]. Indeed, this model [10, 12-13] exhibits more than 60% decrease of FA uptake and utilization by some tissues such as heart, oxidative skeletal muscle, and adipose tissues. On the other hand, Mice overexpressing CD36 in muscle [11] showed an increased FA oxidation in response to contraction.

In the CD36 deficient spontaneously hypertensive (SHR) rats, defective FA uptake is accompanied by a large shift towards glucose metabolism, hyperinsulinemia, insulin resistance and the development of myocardial hypertrophy [14, 15]. All these defects were improved in congenic and transgenic rats re-expressing CD36, and are thus directly attributable to the CD36 gene [14-17].

CD36 deficiency is also known to occur in humans, and several mutations resulting in this condition have been observed [18]. The prevalence of CD36 deficiency varies from 0.3 to 11% and the higher incidences are found in the Asian and African populations [19]. Humans with CD36 deficiency showed a myocardial FA uptake defect [19, 18] similar to the one observed in the CD36 KO mice [13]. In humans, CD36 deficiency may underlie some forms of cardiac diseases such as hypertrophic cardiomyopathy [17,20-21].

To be able to further dissect the direct and possibly secondary effects of CD36 deficiency, we generated a new transgenic mouse line using a genetic rescue approach. By restricting CD36 expression to heart and skeletal muscle, we were able to correct, and thus further specify the main abnormalities of FA

metabolism in the CD36 null mouse. We report that absence of CD36 severely impacts muscle energy production and leads to a lowered endurance during exercise.

Materials and Methods

Materials

Immun-Star Detection Kit was from Bio-Rad. Immobilon PVDF transfer membranes were from Millipore. Enzymes for DNA manipulation were from Roche Molecular Biochemicals. The polyclonal CD36 antibody F2-35 was generated in rabbits against rat adipocyte CD36. Kodak films, FA, and all other chemical products were from Sigma.

Generation of GR-mice

Animals received human care in compliance with the "Principle of Laboratory Animal Care" formulated by the National Society for Medical Research and the "Guideline for the Care and Use of Laboratory Animals" prepared by the National Academy of Sciences and the National Institutes of Health (NIH publication No. 85-23, revised 1985).

Male mice of different genotypes were housed in a barrier-free facility and maintained on a chow diet. They were placed on at 4 weeks of age and were followed for a period of 14 weeks. Mice were weight, age and sex matched.

Generation and identification of CD36 gene- rescued mice CD36 null mice were generated by targeted gene-disruption and backcrossed 4 times to C57Bl/6 [10]. Wild-type control littermates were bred from the same cross as the nulls and were, therefore of identical genetic background. CD36 gene rescued (GR) mice were generated on the CD36 null background using the same approach as previously described [11]. A linear SalI CD36 mini-gene was injected into the male pronucleus of fertile eggs from superovulated females of the CD36 null line. These females had been mated with CD36 nul males and microinjected eggs were transferred into the oviducts of surrogate CD36 null females. The mini-gene contained the full-length rat CD36 cDNA including the poly-adenylation signal, controlled by the regulatory sequences (3.3 kb) of the mouse muscle creatine kinase gene (MCK).

GR founders were identified by southern blot on genomic DNA from tail biopsies as previously described[11]. Transgenic offspring were identified by PCR amplification of a segment of the CD36 gene using the following primers: 5′[GGCTGCGATCGGAACTGTGGGC] -3′ and 5′-[CGATCGGAACTGTGGGCTCATTAC]-3′. These primers amplify a 400-bp segment of the transgene.

Protein analysis

CD36 protein expression was examined by Western analysis. Skeletal muscle was homogenized in 1 ml of ice-cold TES buffer (20 mM Tris-HCl, 1 mM EDTA, 250 mM sucrose, pH 7.4) and a microsomal pellet was obtained by centrifuging at 350,000 x g. Samles from this pellet (20-50 ⊠g of protein) were subjected to SDS-PAGE according to Laemmli et al. [21] followed by transfer to a PVDF membrane. The membrane was blocked with 5% non-fat dry milk and incubated with a polyclonal anti-CD36 antibody for 2h. After washing, immunodetection was carried out using the horseradish peroxidase conjugated goat anti-rabbit antibody (1h, 1:10,000 dilution) of an Immun-Star Detection Kit according to the manufactures manual.

Immunostaining.

Tissues were quickly rinsed in a saline solution and immersed for 48 hours in formalin followed by a sucrose solution [22]. Cryosections on glass slides were washed with PBS (PBS: 100 mM NaCl, 4.5 mM KCl, 3 mM Na2HPO4, 33 mM KH2PO4) and then blocked with normal goat serum at room temperature for 1h. Subsequently to PBS washing, primary polyclonal antibodies anti-CD36 (F2-35) were added. Incubations were conducted overnight at 4°C. The slides were then washed with PBS and secondary antibodies were applied in blocking buffer for 1h at room temperature. The secondary antibodies were fluorescein (FITC)-conjugated goat anti-rat IgG from Jackson Immunoresearch Laboratories Inc.. The slides were then washed with PBS, mounted with Vectashield (VECTOR), and examined using either a standard fluorescence microscope or a confocal microscope.

Analysis of Blood Parameters

Blood parameters were analyzed from overnight (16 h) fasted animals. Tail vein blood was collected in EDTA-rinsed tubes and plasma was prepared immediately. Triglyceride levels were determined using the Infinity enzymatic kit from Sigma Diagnostics (St. Louis, MO) and determination of plasma free FA levels was carried out with the Wako kit (Wako Chemicals, Richmond, VA). Plasma insulin was measured using a Linco kit (St Charles, Missouri, USA). Glucose was measured in whole blood with a MediSense glucose analyzer.

Triacylglycerol, Glycogen and ATP levels in tissues

Hind limbs and liver of fasted mice were quickly rinsed in saline, clamped between aluminum tongs precooled in liquid nitrogen and stored at -80ºC. Tissue glycogen was measured as glucose after

hydrolysis with KOH (30%) and HCL (0.6 N) [23]. Triacylglycerol content was determined enzymatically after lipid extraction as previously described [24]. Tissue protein was determined according to the method of Markwell et al. [25].

Fatty acid oxidation by isolated soleus muscles

The soleus muscles, including the tendons were quickly and carefully excised from the animals and placed in a capped 10 ml vial containing 5 ml KHB supplemented with 2% BSA. After 20 min incubation in a shacking water bath (150 strokes/min) at 30ºC, the muscles were quickly transferred to new vials (1 4 soleus/vial) containing 3 ml KHB with 0.2 mM palmitate and 0.45 Ci/ml [1-14C] palmitate (ICN, Irvine, CA) complexed to 2% BSA. An antibiotic/antimyocytic solution (Sigma, St. Louis, MO) was added to the latter medium to prevent bacterial growth. The vials were capped and 30ºC incubation was continued for 30 min. The incubation medium was quickly transferred to new vials (Kontes, Vineland, NJ) thereafter. The rubber caps of these vials were equipped with a center well filled with 200l of ethanolamine/ethylene glycol (1:2, vol :vol). Perchloric acid was added through the cap to a final concentration of 0.6 M and the vials were incubated overnight with light shaking. The amount of CO produced was determined by liquid scintillation counting of the ethanolamine/ethylene glycol mixture.

Dipyridamole (dissolved in DMSO at 100 x final concentration) was added after the initial 20 min incubation at 30ºC.

Forced swimming test

Mice were individually forced to swim in a transparent glass vessel by attachment of a weight equivalent to 6.67% of the mouse body weight. The water temperature was maintained at 30º C. To swim, mice need to move their legs permanently to keep their heads above the water, and they were considered to be exhausted when they showed the first signs of stopping this movement. Time to reach exhaustion was recorded. Mice were forced to swim 4 repetitive times with 15 min intervals of rest.

Statistical analysis

Data are expressed as mean ± SEM. The significance of the difference in mean values was evaluated using the two-tailed, unpaired Student's t test assuming equal variance. Significance as accepted at $p < 0.05$.

Results

Generation of a transgenic mouse model re-expressing CD36 in heart and skeletal muscle on the CD36 null background

The muscle creatine kinase promoter was used to generate the CD36 gene rescued Mice (GR) expressing this gene in heart and muscle CD36 knock-out mice (KO).

Mice were screened by Southern blot then subjected to PCR analysis. Using primers specific for the injected construct, no signal was detectable in DNA from CD36 null mice and their wild-type littermates (data not shown). Restoring CD36 protein expression in skeletal muscles from gene-rescued (GR) mice was examined by Western analysis. As shown by Febbraio et al. for KO and WT [10], Figure 1a shows that membrane associated CD36 levels in GR muscles were comparable to wild-type levels, while the CD36 protein was undetectable in CD36 null muscles. GR tissues other than heart and muscle were negative for the CD36 protein (results not shown). In order to confirm the cardiac expression of CD36 in rescue mice, we examined heart tissue sections from the different animals by immunohistochemistry. GR mice (Figure 1, panel b) showed a high expression level of CD36 that was comparable to the wild type levels. CD36 null hearts showed no expression of CD36 confirming the absence of the protein in these mice.

Rescue of the CD36 null phenotype in GR mice

Upon fasting; CD36 null mice showed hypoglycemia and elevated plasma levels of triglycerides (Figure 2). Rescuing the expression of CD36 in heart and muscle of GR mice normalized the levels of triglycerides. Eventhough GR glucose levels were significantly higher the CD36-null mice, they did not reach the levels of the WT mice (Figure 2). Fasting glucose and insulin levels of WT mice, were similar to those of GR mice but were significantly higher than in CD36-null mice.

Endurance during exercise

Since muscle ability to perform work is highly dependent on energy from FA-oxidation, we tested whether CD36 expression impacts the ability to exercise. Forced swimming tests were used to examine wild-type, CD36 null, and GR mouse tolerance to endurance exercise (Figure 3).

CD36 null mice's isolated soleus muscles, which have low energy stores (Figure 4), showed a drastically reduced ability to oxidize. As shown in Figure 5a, the palmitate oxidation rate of the GR

soleus was significantly higher compared to the CD36null muscle, and comparable to the rates of wild-type muscle. Significant differences in soleus weights were absent (results not shown).

Dipyridamole was shown to stimulate palmitate oxidation in a concentration-dependent manner in cells expressing CD36 [26] was used in the same set of experiments. Figure 5b shows that the oxidation increased in the WT and GR muscles (expressing CD36) in comparison to the KO animals lacking CD36.

Stimulation of FA uptake using dipyridamole increased palmitate oxidation by both the wild-type control and the GR soleus, but did not show an effect on the CD36 null muscle (Figure 5b). Wild-type and GR oxidation rates remained statistically indistinguishable after dipyridamole stimulation.

Glucose utilization

Tissue uptake of glucose in vivo was measured to determine the effect of CD36 rescue in the GR mice on glucose uptake and metabolism. When comparing uptake of F-2- FDG in GR mice to CD36Ko (Figure 6) , decreased levels were observed in hearts, diaphragms, soleus, gastrocnemius, and hind limb muscle , unaltered levels in adipose tissue, and increased levels in the liver, in a way comparable to the WT mice (Figure 6).

Discussion

The present study showed that CD36-facilitated FA transport in skeletal muscle plays a key role in the ability to exercise. The re-expression of CD36 in heart and skeletal muscle reversed the main abnormalities in FA metabolism previously described in CD36 null mice. Gene-rescued GR mice showed normal blood parameters, and FA oxidation rates by isolated soleus muscles were comparable to the WT rates. CD36 null animals displayed both significantly decreased muscle FA oxidation rates and a drastically lowered ability to undergo forced exercise. The muscle-targeted rescue of CD36 expression in GR mice resulted in a completely normalized endurance to forced exercise. The decreased tolerance to exercise of CD36 null animals is essentially due to a decreased ability to store metabolic energy, thereby impacting energy expenditure. This was demonstrated by the observation that muscle triacylglycerol and glycogen contents are significantly decreased in CD36 muscles, whereas these levels in wild-type and GR muscles were fully comparable. All these findings are consistent with the concept that cellular energy derived from CD36- mediated FA uptake is a determinant for the ability to exercise.

It is understandable that since most of muscle's fibers have a high dependence on energy from FA oxidation, CD36 expression , which mediates FA uptake and utilization, impacts the ability of muscle to perform work. Coburn et al. [12] showed at least a 60% reduction in FA uptake by the heart and red muscles of CD36 KO mice which was compensated by an increase of glucose utilization [12,27]. Nevertheless, this seems to be insufficient to maintain a normal metabolic energy storage and production and the effects became apparent at heightened cellular demands such as during forced exercise. In line with this, we recently demonstrated that CD36 null hearts have a drastically lowered tolerance to ischemia and reperfusion [28]. Bonen et al. [29] recently showed that the acute regulation of FA uptake by muscle activity involves the translocation of CD36 from intracellular stores to the sarcolemma. This molecular mechanism appears to be similar to membrane recruitment of GLUT -4 during the regulation of glucose uptake [30]. In our experiments, adding dipyridamole to the muscle incubations did not show an effect on the FA oxidation rate of the CD36 null soleus. However, our data also elegantly demonstrates the key role for skeletal muscle in the maintenance of FA homeostasis. Alterations in CD36 facilitated FA transport, and subsequently FA utilization in muscles, which were directly reflected in the blood as both, the defect in muscle FA

oxidation and the alterations in plasma metabolite present in CD36 null mice, could be reversed by muscle-targeted re-expression. Skeletal muscle, by being a key FA oxidizer, is the principal site for the removal of FA from the circulation, and this removal is mediated by CD36 protein. Interestingly, the findings also imply that the expression of CD36 by the structural muscle cells is sufficient for maintaining normal FA uptake rates and the plasma metabolite levels.

In conclusion, the present study demonstrate the role that could play CD36 expression in muscle function and performance. Experiments underway will help determine if genetic or environmental factors alterations affecting CD36 levels could impact humans muscular performance and athletic ability.

Author's contributions

All the authors participated in the Study conception and design, Acquisition of data, Analysis and interpretation of data, drafting of manuscript and Critical revision.

Acknowledgements

This work was conducted in part at the physiology and biophysics department of Stony Brook University (NY, USA). Work was supported by NIDKK Grant DK33301 to NAA and AHA grant 00303445. The authors would like to thank Dr. T. Rosenquist for technical help in generating the transgenic animals (Stony Brook University,NY, USA).

We acknowledge support from Volubilis (French-Moroccan Grant) to AI. This work was also supported by a grant from the NIH for H3Africa BioNet to AI.

References

[1] Abumrad, N. A., Perkins, R. C., Park, J. H., and Park, C. R: Mechanism of long chain fatty acidpermeation in the isolated adipocyte. J. Biol. Chem 1981, 256, 9183-9191.

[2] Harmon , B. A., et al: in AIP Conf. Proc. 280, 1st Compton Gamma Ray Observatory Symp., ed. M. Friedlander, N. Gehrels, & D. Macomb (New York: AIP) 1993, 314

[3] Abumrad, NA, el-Maghrabi, MR, Amri, EZ, Lopez, E & Grimaldi, PA: Cloning of a rat adipocyte membrane protein implicated in binding or transport of long-chain fatty acids that is induced during preadipocyte differentiation. Homology with human CD36. J Biol Chem 1993, 268, 17665-17668.

[4] Greenwalt, D. E., Lipsky, R. H., Ockenhouse, C. F., Ikeda, H., Tandon, N. N., & Jamieson, G. A: Membrane glycoprotein CD36: A review of its roles in adherence, signal transduction, and transfusion medicine. Blood 1992, 80,1105−1115.

[5]Van Nieuwenhoven, F.A., et al: Putativemembrane fatty acide translocase and cytoplasmic fatty acid-binding protein are co-expressed in rat heart and skeletal muscles, Biochem Biophys Res Commun 1995,18 207(2):p.747-52.

[6] Luiken, J. J., Schaap, F.G., van Nieuwenhoven, F. A., van der Vusse, G. J., Bonen, A., & Glatz, J. F: Cellular fatty acid transport in heart and skeletal muscle as facilitated by proteins. Lipids, 34 (Suppl.) 1999, 169−175.

[7] Luiken JJFP, Turcotte LP & Bonen A : Protein-mediated palmitate uptake and expression of fatty acid transport proteins in heart giant vesicles. Journal of Lipid Research 1999, 40, 1007−1016.

[8] Greenwalt, D. E., Scheck, S. H., & Rhinehart-Jones, T: Heart CD36 expression is increased in murin models of diabetes and in mice fed a high fat diet. Journal of Clinical Investigation 1995, 96, 1382−1388.

[9] Pelsers, M. M. A. L., Butler, P. J., Bishop, C. M. and Glatz, J. F. C: Fatty acid-binding protein in heart and skeletal muscles of the migratory barnacle goose throughout development. Am. J. Physiol 1999, 276,R637−R643.

[10] Febbraio, M, Abumrad, NA, Hajjar, DP, Sharma, K, Cheng, W, Pearce, SF & Silverstein, RL: A null mutation in murine CD36 reveals an important role in fatty acid and lipoprotein metabolism. J Biol Chem 1999, 274, 19055-19062.

[11] Ibrahimi, A, Bonen, A, Blinn, WD, Hajri, T, Li, X, Zhong, K, Cameron, R & Abumrad, NA: Muscle- specific overexpression of FAT/CD36 enhances fatty acid oxidation by contracting

muscle, reduces plasma triglycerides and fatty acids, and increases plasma glucose and insulin. J Biol Chem 1999, 274,26761-26766.

[12] Coburn, C.T., et al: Defective uptake and utilization of long chain fatty acids in muscle and adipose tissues of CD36 knockout mice [In Process Citation]. J Biol Chem 2000, 275(42): p.26761-6.

[13] Pravenec M, Zídek V, Šimáková M, Křen V, Křenová D, Horký K, Jáchymová M, Míková B, Kazdová L, Aitman TJ, ChurchilL PC, Webb RC, Hingarh NH, Yang Y, Wang JM, Lezin EM, Kurtz TW: Genetics of Cd36 and the clustering of multiple cardiovascular risk factors in spontaneous hypertension. J Clin Invest 1999, 103: 1651-1657.

[14] Pravenec M, Zídek V, Musilová A, Vorlíček J, Křen V, St Lezin E, Kurtz T: Genetic isolation of a blood pressure quantitative trait locus on chromosome 2 in the spontaneously hypertensive rat. J Hypertens 2001, 19: 1061-1064, .

[15] Hajri, T & Abumrad, NA: Fatty acid transport across membranes: relevance to nutrition and metabolic pathology. Annu Rev Nutr 2002, 22, 383-415.

[16] Tanaka R. Higo Y. Shibata T. Suzuki N. Hatate H. Nagayama K., Nakamura T. Accumulation of hydroxy lipids in live fish infected with fish diseases. Aquaculture 2002, 211:341–351.

[17] Nozaki, M., Ohishi,K., Yamada, N., et al: Developmental abnormalities of 1 glycosylphosphatidylinositol-anchor-deficient embryos revealed by Cre/loxP system. Lab. Invest 1999, 79:293–299.

[18] Tanaka T. Chemoprevention of human cancer: biology and therapy. Crit Rev Oncol/Hematol 1997, 25, 139-74.

[19] Tanaka T. Effect of diet on human carcinogenesis. Crit Rev Oncol/Hematol 1997, 25, 73-95.

[20] Uysal K, Wiesbock S et al. : Protection from obesity induced insulin resistance in mice lacking, TNT-alpha function. Nature 1997, 389, 610-614.

[21] Laemmli, U.K., E. Molbert, M. Showe, and E. Kelenberger: Form-determining function of genes required for the assembly of the head of bacteriophage T4. J. Mol. Biol 1970, 49:99-113

[23] Passonneau, J.V. and V.R. Lauderdale: A comparison of three methods of glycogen measurement in tissue. Anal. Biochem 1974, 60: 405-412.

[24] Folch, J., M. Lees, and G. H. Sloane-Stanely: A simple method for the isolation and purification of total lipids from animal tissues. J. Biol. Chem 1957, 226:497–507.

[25] Markwell, J. P., S. Reinman, and J. P. Thornber: Chlorophyll-protein complexes from higher plants: a procedure for improved stability and fractionation. Arch. Biochem Biophys. 1978, 190(1):136–141.

[26] Luiken JJ1, Coort SL, Willems J, Coumans WA, Bonen A, Glatz JF: Dipyridamole alters cardiac substrate preference by inducing translocation of FAT/CD36, but not that of GLUT4. Mol Pharmacol. 2004, 65(3):639-45.

[27] Ibrahimi A. and N.A. Abumrad: Role of CD36 in membrane transport of long chain fatty acids. Curr Opin Clin Nutr Metab Care 2002, 5: 139-145.

[28] Irie H, Krukenkamp IB, Brinkmann JF, Gaudette GR, Saltman AE, Jou W, Glatz JF, Abumrad NA, Ibrahimi A. Myocardial recovery from ischemia is impaired in CD36-null mice and restored by myocyte CD36 expression or medium-chain fatty acids Proc Natl Acad Sci U S A 2003, 27;100(11):6819-24.

[29] Bonen A1, Luiken JJ, Arumugam Y, Glatz JF, Tandon NN: Acute regulation of fatty acid uptake involves the cellular redistribution of fatty acid translocase. J Biol Chem. 2000, 12;275(19):14501-8.

[30] Bonen A1, Luiken JJ, Arumugam Y, Glatz JF, Tandon NN: Acute regulation of fatty acid uptake involves the cellular redistribution of fatty acid translocase. J Biol Chem. 2000, 2;275(19):14501-8.

Figures:

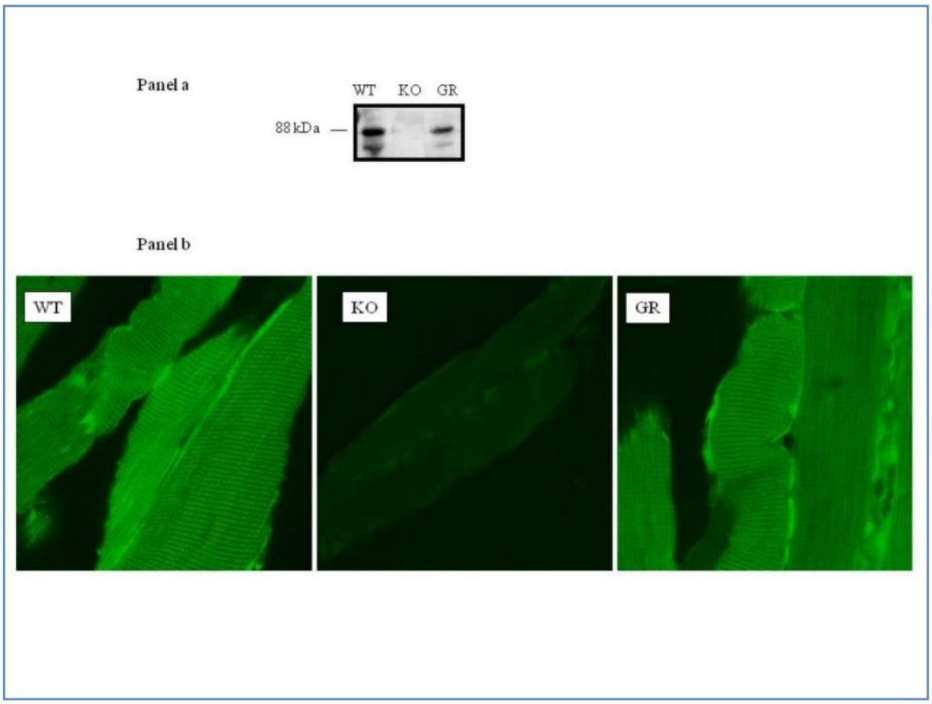


Figure 1: Generation of transgenic mice overexpressing CD36 in muscle tissues.

Panel a: CD36 protein expression was analyzed by western blot. Proteins were prepared as described in experimental procedures. CD36 protein levels from different mice WT (wild type); KO (CD36-null) and GR (CD36-rescue mice) were detected using a monoclonal antibody against CD36. Reaction with preimmune sera did not yield a detectable signal (not shown). Immunodetection was by Immun-Star Detection kit.

Panel b: Immunofluorescence studies of CD36 expression in muscle of different animal models. Tissue sections prepared from mouse muscle were incubated with rabbit polyclonal antibodies against CD36 followed by a fluorescein-conjugated goat anti-rabbit IgG. Sections were viewed under a confocal microscope.

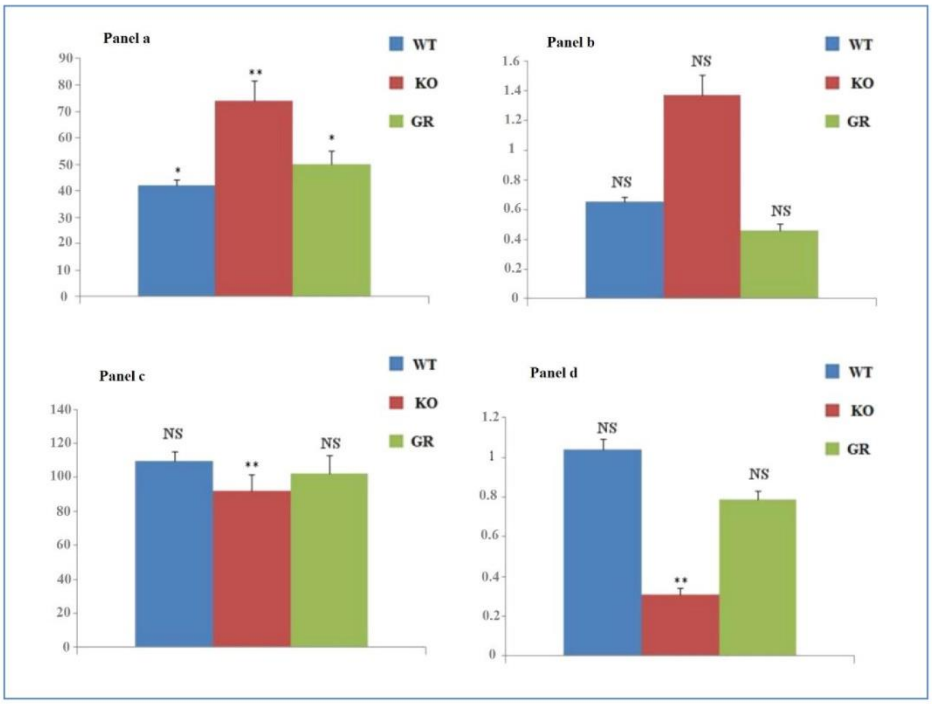


Figure 2: Fasting plasma levels of TGs (mg/dl), FAs (mM), glucose (mg/dl) and insulin (ng/ml) for WT, CD36-null and GR mice maintained on chow diet. Mice were fasted overnight. Blood was collected from the tail vein into EDTA-coated tubes and centrifuged to separate out plasma, which was used to determine triglyceride and insulin levels. Glucose levels were determined on whole blood. Data are mean ± SEM and n was 4 or more for each mouse group ($P<0.01$). ***** KO Significantly different from wild-type and GR mice. NS: No significant difference between wild-type and GR mice.

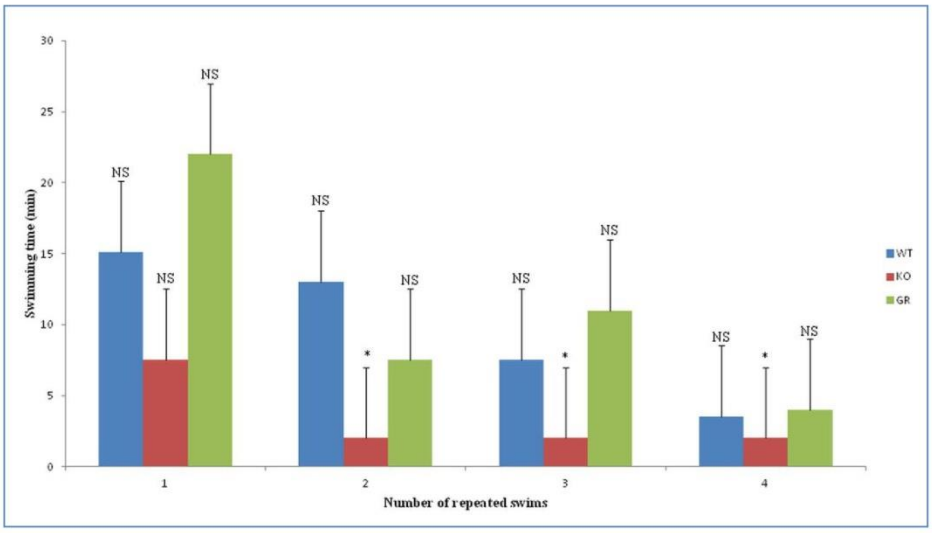


Figure 3: Forced swimming test of GR mice. Wild-type (WT), CD36 null and gene-rescued GR mice (n=6 per group) were forced to swim until exhaustion as described in experimental procedures. Time to reach exhaustion was recorded and swimming was repeated 4 times with 15 min intervals. Data are mean ± SEM. ***** KO Significantly different from wild-type and GR mice. NS: No significant difference between wild-type and GR mice.

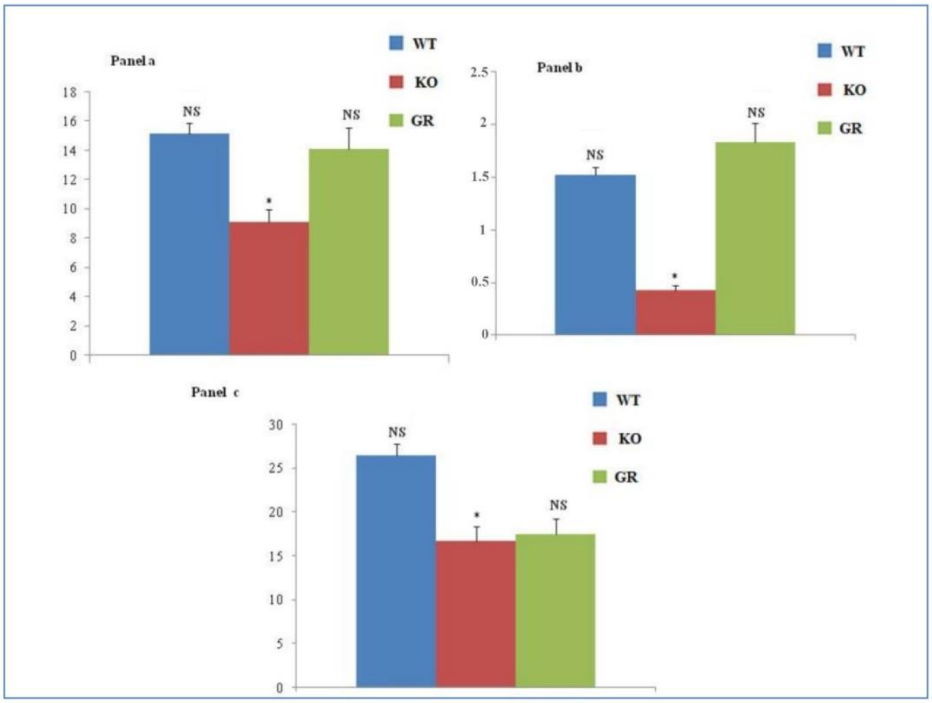


Figure 4: Glycogen, TG and ATP contents in muscle of WT, CD36-null and GR mice fed chow diet.

Muscle tissues were collected from different animal models and levels of glycogen, triglycerides and ATP were determined as described in experimental procedures. Data are mean ± SEM and n was 5 for each mouse group ($P<0.01$). * KO Significantly different from wild-type and GR mice. NS: No significant difference between wild-type and GR mice.

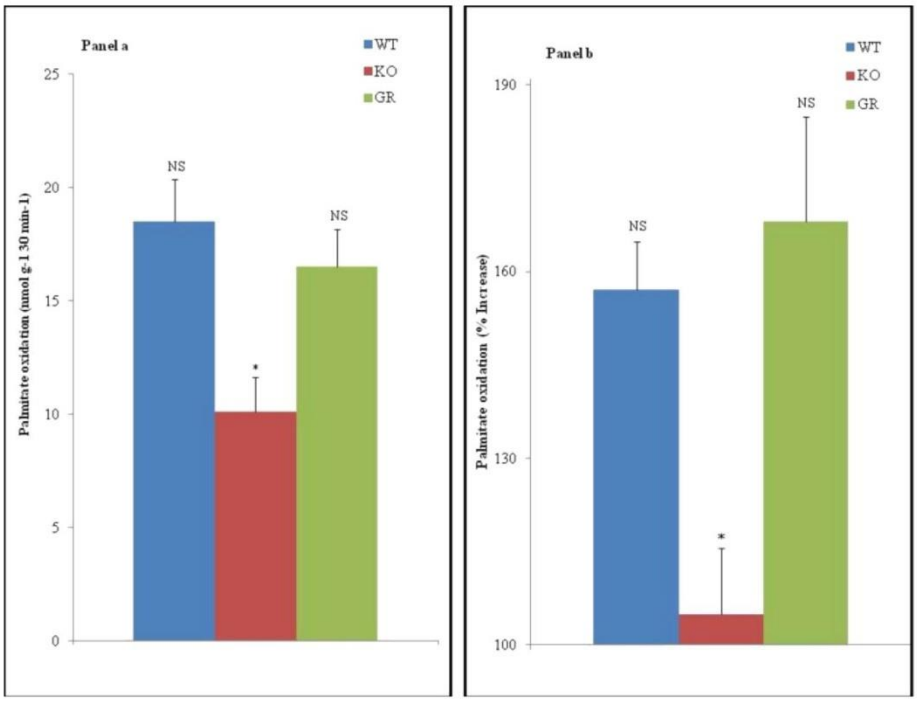


Figure 5: Palmitate oxidation by soleus muscle from GR mice. Palmitate oxidation rates of isolated soleus muscles from wild-type (WT), CD36 null and gene-rescued GR mice were determined at a low fatty acid/BSA ratio. **Panel a:** represents fatty acid oxidation of non-stimulated muscles (n=3 per group). In **panel b**, fatty acid oxidation was stimulated by addition of 100 ⊠M dipyridamole (n=3 per group). Data are mean ± SEM. ⁎ Significantly different from wild-type and GR mice ($P<0.01$). ***** KO Significantly different from wild-type and GR mice. NS: No significant difference between wild-type and GR mice.

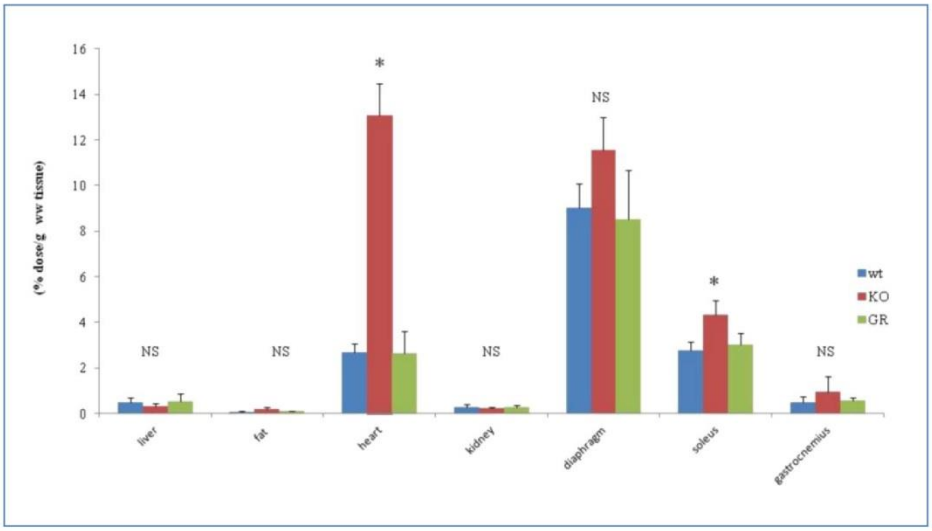


Figure 6: ^{18}F-2-FDG uptake by tissues of CD36-null and WT mice fed Chow diet. ^{18}F-2-FDG (12 µCi) was injected into a lateral tail vein of mice fasted for 16 hours that were maintained on a Chow diet. At the end of the experiment, tissues were removed, weighed, and counted for ^{18}F-2-FDG radioactivity; uptake rate is expressed per gram of wet weight tissue (ww). FDG uptake was determined as described in Methods. Data are means ± SEM (n = 4). *P <0.05. * Significantly different from wild-type and GR mice. NS: No significant difference between KO, wild-type and GR mice.

II. Caractérisation des classes de gènes régulés en absence de l'expression du gène CD36

Dans ce chapitre, nous avons souhaité décrire la régulation de l'expression des gènes en réponse à différents stress, et ce, pour voir ce qui se passe, principalement à l'échelle des gènes et des protéines correspondantes, lorsque qu'on supprime l'expression d'un gène ou de le surexprimé, nous avons notamment voulu caractériser la réponse transcriptionnelle de manière globale, c'est-à-dire savoir combien de gènes et lesquels étaient induits ou réprimés. Le but était ensuite d'identifier les gènes co-exprimés, qui sont exprimés de manière similaire dans différentes conditions. Pour cela, nous avons recherché les classes de gènes dont les comportements étaient proches en termes de réponse transcriptionnelle. Finalement, notre objectif était d'expliquer pourquoi certaines classes de gènes étaient induites alors que d'autres étaient réprimées. A cet effet, nous avons voulu caractériser les sous ensembles de protéines correspondantes par des signatures, telles que le métabolisme et l'angiogenese, afin de comparer ces classes de protéines.

5. Présentation du dispositif expérimental

Pour étudier la régulation de la transcription en réponse à l'absence d'expression de gène CD36, nous avons choisi de travailler sur trois modéles de souris. On s'intéresse dans le cadre de notre étude aux gènes qui suivent le même profile du gène CD36, en l'occurrence les gènes qui sont présents en présence de CD36 et les gènes qui sont absents en cas de son absence. Alors le groupe requête ici est un groupe formé du gène CD36 comparé aux autres gènes de la puce à ADN de différentes conditions :

- Cas de Wild Type (WT le sauvage) ; des souris qui ont une expression normale de gène CD36 ;
- Cas de knockOut (KO) ; souris en absence d'expression de gène CD36 ;
- Cas de GeneRescue (GR), souris où le gène CD36 à été ré-exprimé chez les Knockout (surexpression de gène) dans le muscle et dans le coeur.

Alors le but est l'extraction des gènes qui sont présents dans le WT, et absents chez les KO. Ainsi que pour confirmer les résultats on compare les résultats de la première analyse avec les résultats des GR c'est-à-dire comparé avec les gènes qui apparaissent après ré-exprission du gène CD36 chez les KO.

Dans cette étude nous avons utilisé deux technologie du puce à ADN (article 2), d'une part le cas des microarrays simples canaux utilisés est la technologie Affymetrix, dans laquelle des oligonucléotides de petite taille (25-mers) sont synthétisés in situ. Chaque gène est représenté par 11 à 20 paires d'oligonucléotides: des oligonucléotides complémentaires à la séquence du gène qualifiés de « perfect-match » (PM), et des oligonucléotides différant du PM par une mutation du nucléotide situé en position centrale (position 13), et qualifiés de « mis-match » (MM). L'objectif de ces doublons est de contrôler les hybridations non spécifiques. Le principe général est le même que pour les microarrays double canaux mais ici, un seul type de fluorophore est utilisé pour la cible (streptavidine couplée à la phycoérythrine). D'autre part la technologie Souris Whole Genome Microarray 4x44K (Agilent Technologies, Santa Clara, CA) dont ces microarrays sont constitués de lames de verre sur lesquelles des milliers d'ADNc ou oligonucléotides sont déposés: chaque gène (de fonction connue ou inconnue) est représenté par au moins un spot sur la lame. Deux échantillons d'ARN, l'un issu d'une population cellulaire contrôle et l'autre issu de la population à étudier, sont réverse- transcrits en ADNc, et marqués par des fluorophores différents (cyanine-5, rouge et cyanine- 3, vert). Les deux échantillons sont ensuite déposés simultanément sur le microarray de manière à s'hybrider avec les séquences complémentaires présentes sur la lame. Le microarray est ensuite analysé par un scanner à laser (microscope confocal). Selon la longueur d'onde émise par le laser, le scanner détectera le signal renvoyé par l'un ou l'autre fluorophore. Le rapport des fluorescences rouge/vert est ainsi déterminé pour chaque spot et permet de comparer le niveau d'expression relatif de chacun des gènes pour les deux échantillons de départ.

La préparation des puce à ADN est faite en collaboration avec la faculté de médecine de Washington, et nous avons reçu des fichiers sous trois formats (.dat, .cell , et .schip) pour l'analyse des données

6. Caractérisation de la réponse cellulaire

Nous avons d'abord voulu mener une analyse préliminaire de manière à dégager les principales tendances des réponses transcriptionnelles étudiées. Pour cela, nous avons normalisé les données brutes et utilisé une méthode heuristique pour déterminer si un gène donné était induit ou réprimé à un temps donné. Cette analyse nous ayant permis de mettre en évidence des phases de réponse, nous avons développé dans un second temps une analyse statistique, afin d'identifier les gènes mettant en évidence un comportement différent au cours des phases de réponse. Finalement, nous avons voulu comprendre les mécanismes biologiques sous-jacents.

6.1. Normalisation des données

Une des difficultés de la technologie des puces à ADN est la quantité d'informations générée par les logiciels d'analyse d'images pour chaque spot. Choisir les données les plus informatives sur les mesures d'expression et leur qualité est donc indispensable. L'autre point critique est la suite des étapes mise en œuvre pour obtenir des données de qualité. A chaque étape des biais expérimentaux peuvent entacher d'erreur la mesure finale. Les variations biologiques d'intérêt peuvent donc être masquées par des bruits techniques et biologiques. Aussi, outre la définition d'un plan expérimental adéquat (replicates, randomisation des facteurs de « nuisances »), il est nécessaire d'appliquer une procédure systématique de traitement et de transformation à ces données afin de minimiser (voire corriger) ces variations indésirables.

- **Pour les puces Affymetrix**

Les données étaient normalisées par la méthode RMA fournie par le package affy de la librairie R Bioconductor. Les sondes étaient convertis en identifiants Ensemble grâce aux fichiers de définition des puces personnalisés fournis par le site internet de Brain Array. Après normalisation, des matrices d'expressions étaient obtenues avec chaque colonne correspondant à un échantillon et chaque ligne correspondant à un gène. Chacun des identifiants Ensembl était ensuite annoté en se connectant à la base de données biomaRt par le package R correspondant, afin de récupérer les symboles et les données de localisation dans le génome. Par la suite, les données de quelques temps de stimulation manquaient pour certains. Afin de remplir les données manquantes et d'assurer une analyse homogène, nous avons ajouté la différence moyenne de tous les échantillons entre le temps initial et celui manquant, aux données d'expressions du temps non manquant.

- **Pour les puces Agilent**

Les puces sont scannées par un laser en utilisant différentes longueurs d'onde de manière à obtenir les intensités numériques de chaque spot. Ainsi, une mesure relative à l'intensité globale d'hybridation est obtenue pour chaque élément sur la puce. Dans notre cas, les puces à ADN ont été scannées avec le logiciel GenePix (GeneP ix T M Pro 4.0). Or, ces données de transcriptome sont bruitées et biaisées. Dans un premier temps, il était donc nécessaire de normaliser les données brutes. Nous avons alors choisi d'appliquer la méthode de normalisation la plus classiquement utilisée, à savoir le traitement du bruit de fond, puis le traitement du biais par rapport à l'intensité totale. Pour cela, nous avons tout d'abord soustrait le bruit de fond pour obtenir l'intensité du signal pour chaque spot. Nous avons encore utilisé le logiciel R pour appliquer la normalisation lowess (Cleveland et Devlin, 1988) en nous basant sur le Bioconductor marray du logiciel R. Nous avons réglé le paramètre de lissage (smooth) à 0,33 comme cela est recommandé. Les valeurs normalisées de chacun des deux canaux (Cy3et Cy5) ont ensuite été combinées pour obtenir un ratio (voir Équation 1.1) ou un log-ratio (voir Équation 1.2).

$$ratio = \frac{Cy_3}{Cy_5} \tag{1.1}$$

$$ratio = log_2(ratio) = log_2(\frac{Cy_3}{Cy_5}) \tag{1.2}$$

Lorsque cela était nécessaire, nous avons regroupé les valeurs des réplicats, par exemple pour comparer les différents points de mesure. Ainsi, pour chaque point de mesure, les différentes valeurs ont été moyennées de manière à obtenir un unique ratio par gène pour un temps donné. Pour cela, nous avons calculé la moyenne arithmétique des log-ratios, ce qui correspond à la moyenne géométrique des ratios. Après que les données brutes ont été normalisées de cette façon, nous avons mené une première analyse globale.

6.2. Identification des gènes différentiellement exprimés

Le premier pas vers l'analyse de transcriptome par la technologie des puces à ADN est la mise en évidence des gènes différentiellement exprimés, entre deux conditions différentes. Pour mesurer cette différence, il faut pouvoir distinguer les variations biologiques qui sont le reflet du fonctionnement de la cellule, des variations expérimentales qui viennent gêner l'interprétation. Nous avons utilisé Les tests non paramétriques car ils sont plus adaptés aux données bruitées telles que celles engendrées par nos puces à ADN affymetrix . Parmi ces tests non paramétriques, l'analyse SAM (Significance Analysis of Microarrays) avec un P-valeur et Fold respectivement fixés à 1,5 et 0,002.

Cependant pour les puces Agilent Nous avons choisi d'utiliser la construction de modèles linéaires implémentée dans le package Limma du logiciel Bioconductor car cette méthode possède les deux principaux avantages de pouvoir prendre en compte un faible nombre de réplicats et d'être applicable rapidement sur des données réelles. En effet, nous ne disposions que de deux réplicats par point de mesure dans la plupart des cas. Pour identifier les gènes dont les niveaux d'expression étaient globalement régulés, nous avons donc utilisé une analyse bayésienne empirique pour estimer les paramètres du modèle. Ensuite nous avons effectué un test statistique (t-test) pour évaluer dans quelle mesure chaque gène suivait le modèle. Enfin, nous avons corrigé les p-values (1,5) obtenues par une correction pour les tests multiples qui permettait de contrôler le taux de faux positifs (FDR). Pour limiter le taux de faux positifs à 5%, nous avons pu fixer un seuil à 10−3sur lap-value en suivant la méthode de Benjamini-Hochberg. En utilisant cette méthode.

En utilisant la technologie Affymetrix, nous avons pu identifier 39 entre les gènes CD36-CD36- KO et exprimé différentiellement WT et l'utilisation de la technologie d'Agilent avec les mêmes paramètres, nous avons identifié 35 gènes exprimés de manière différentielle. La comparaison des deux listes de gènes identifiés entre autre que 30 d'entre eux étaient communs aux deux technologies (Cf. figure 36) .

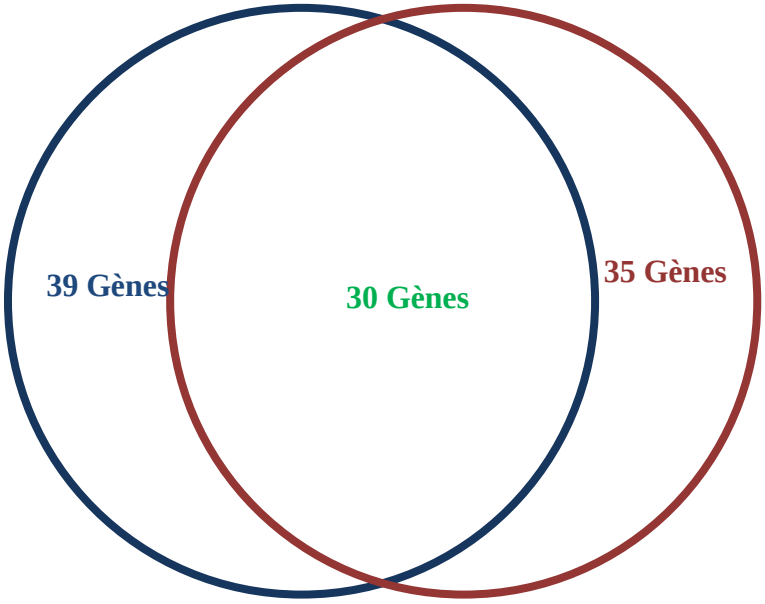


Figure 36: Diagramme de Venn pour comparer les deux technologie de puce à ADN (Agilent technology et affymetrix technology).

6.3.Identification des classes de gènes co-exprimés

Pour regrouper les gènes co-exprimés, nous avons choisi d'étudier les profils d'expression des gènes, et en particulier leur similarité. L'une des approches les plus couramment utilisées pour identifier des classes de gènes ayant des profils d'expression similaires est la classification hiérarchique Cette méthode permet d'obtenir plusieurs partitions possibles de l'ensemble des gènes. Cela implique entre autres que chaque gène appartient à une et une seule classe.

D'abord utilisées en phylogénie, les méthodes de classification hiérarchique sont aujourd'hui les techniques de classification non supervisée les plus utilisées pour étudier les profils d'expression de gènes ou d'échantillons. Elles génèrent des suites de classes emboîtées qui définissent unehiérarchie de partitions encore appelée classification hiérarchique. Les algorithmes de classification travaillent à partir des matrices de distances issues des matrices de données d'expression. Actuellement, il existe trois principales modalités de calcul de distances entre les classes (distance inter-groupes) qui permettent de générer deux grands types d'algorithmes de classifications hiérarchiques : les algorithmes ascendants et les algorithmes descendants.

Des règles de calcul, encore appelées règles d'agglomération, sont nécessaires pour estimer les liaisons entre les groupes disjoints. Les principales distances inter-groupes sont actuellement le lien simple, le lien complet et le lien moyen. Le lien moyen (average linkage) ou UPGMA (Unweighted Pair Group Method of Agregation) est l'approche la plus utilisée. La distance entre deux groupes est la moyenne des distances entre toutes les paires d'objets (gènes ou échantillons biologiques) de ces deux groupes (Cf. figure 37).

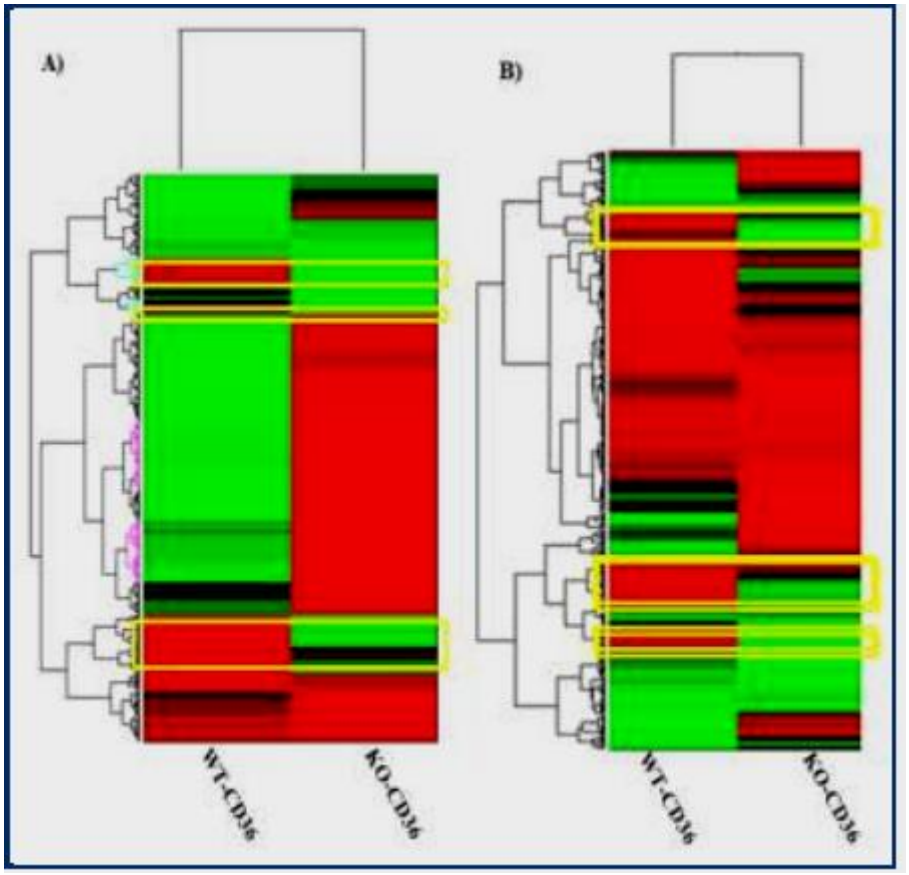


Figure 37 : Classification hiérarchique des gènes sélectionnés: A) classification hiérarchique des gènes sélectionnés en utilisant la technologie Affymetrix; B) classification hiérarchique des gènes sélectionnés en utilisant la technologie Agilent.

6.4.Annotation fonctionnelle

Après avoir identifié des groupes de gènes différentiellement exprimés et afin de pouvoir interpréter les données, il est nécessaire de procéder à des tests d'enrichissement fonctionnel. En effet, les gènes co-exprimés sont généralement impliqués dans des processus ou voies de signalisation similaires (Eisen et al., 1998).

6.4.1. Les différentes sources d'information

Il existe diverses sources d'information utiles pour l'annotation et donc pour l'interprétation des données de puces à ADN. En effet, de très nombreuses bases de données stockent des informations sur la fonction, la localisation, l'expression tissulaire, la régulation et les interactions des gènes ou de leurs produits **(Tableau 5).** En effet, on considère ici que les transcrits identifiés précédemment sont traduits de manière équivalente en quantité de protéines fonctionnelles. Cela ne tient donc pas compte des mécanismes de régulation post-transcriptionnelle et post-traductionnelle.

Parfois, les données sont organisées en un ensemble structuré de termes et concepts au vocabulaire contrôlé, appelé ontologie. Les concepts sont organisés dans un graphe dont les relations peuvent être des relations sémantiques ou des relations d'inclusion. L'objectif

premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné. L'ontologie la plus connue pour l'annotation de données provenant de puces à ADN est Gene Ontology (GO ; (Ashburner et al., 2000)). Celle-ci propose un vocabulaire contrôlé de termes décrivant les propriétés des produits des gènes. Elle est composée de 3 domaines : – compartiment cellulaire, ou cellular component, décrivant la localisation des protéines au sein de la cellule (comme par exemple : noyau, cytoplasme, membrane) ; – fonction moléculaire ou molecular function, décrivant les activités au niveau moléculaire, telles que la liaison (par exemple le terme GO « transcription factor binding », GO :0008134) ou la catalyse ; – processus biologique ou biological process, représentant l'ontologie la plus intéressante pour connaître la fonction des protéines. Elle nous renseigne sur les processus dans lesquels des protéines sont impliquées, comme par exemple la transcription (terme « transcription, DNA-dependent », GO :0006351, Figure 3.6).

6.4.2. Quelques outils d'annotation

Plusieurs outils utilisant cette ontologie ont été créés comme AmiGO, GOToolsBox ((Martin et al., 2004)), FATIGO (Al-Shahrour et al., 2007). Les autres bases de données ont également mis en place un système permettant des recherches en fonction d'un gène, d'une protéine, d'un processus biologique ou d'une voie de signalisation. D'autres approches sont également utilisées pour obtenir des informations sur des gènes telles que des outils de fouille de texte comme Chilibot (Chen & Sharp, 2004), iHOP (Good et al., 2006). Enfin, des logiciels proposent également l'accès à différentes sources de données précédemment citées. Parmi les outils gratuits, ceux principalement utilisés par les biologistes et les bioinformaticiens sont « The Database for Annotation, Visualization and Integrated Discovery » DAVID knowledgebase (Huang et al., 2009) et « Gene Set Enrichment Analysis » GSEA (Subramanian et al., 2005). La base de données DAVID propose ainsi un outil de regroupement d'annotations fonctionnelles permettant l'identification de groupes d'annotations significativement surreprésentées dans une sélection de gènes (Huang da et al. 2007 ; Sherman et al. 2007). Alors que GSEA est une méthode non paramétrique qui détermine si un jeu de gènes défini a priori possède des différences statistiquement significatives entre deux états biologiques ; cette méthode permet de calculer des scores d'enrichissement fonctionnel en utilisant la base de données moléculaire Molecular Signature DataBase (MSigDB) (Subramanian et al., 2005).

Disease	GENETIC ASSOCIATION OMIM DISEASE
Gene Ontology (GO)	GO Biological Process (BP) GO Cellular Component (CC) GO Molecular Function (MF)
General annotations	CHROMOSOME CYTOBAND
Litterature	GENERIF SUMMARY PUBMED ID
Pathways	BIOCARTA KEGG PANTHER REACTOME
Protein domains	BLOCKS COG INTERPRO PFAM SCOP SMART SSF TIGRFAMS
Protein interactions	BIND NCICB CAPATHWAY REACTOME TFBS conserved
Tissues espressions	CGAP EST QUARTILE CGAP SAGE QUARTILE GNF U133A QUARTILE PIR TISSUE SPECIFICITY UNIGENE EST QUARTILE UP TISSUE

Tableau 5 : Liste des principales annotations contenues dans l'outil DAVID knowledgebase, regroupées par domaine

6.4.3. **Représentation des résultats selon la fonction**

Afin d'être capable de mettre en relation ces différents critères, cette méthode s'appuie sur une représentation des données sous forme de groupes de gènes ou de protéines ayant une valeur similaire pour un critère biologique donné. Ainsi, n'importe quelle donnée biologique doit pouvoir être convertie sous forme de collection de groupes et ainsi être comparée à une autre collection (représentant un autre critère). Cette conversion est en fait une étape essentielle pour la comparaison de données, et elle est plus ou moins aisée selon le critère biologique.

Une cardiomyopathie ou myocardiopathie (maladie du muscle cardiaque) correspond a un groupe hétérogène de maladie touchant le myocarde, et responsable d'un

disfonctionnement du muscle. Cependant, il semble de plus en plus que l'insulinoresistance, le diabète de type 2 et la cardiomyopathie ne soient pas des variables indépendantes mais des variables reliées à des modifications du métabolisme. Plus spécifiquement, de fortes concentrations intracellulaires de métabolites d'acides gras à longue chaine, constituent un important facteur du développement de l'insulinoresistance cardiaque.

Cependant, la biologie est un domaine qui manque encore de formalisme strict. Malgré les efforts du consortium HGNC (*HUGO Gene Nomenclature Committee*) (http://www.genenames.org) pour standardiser la nomenclature des gènes, des améliorations sont encore nécessaires pour définir les fonctions des gènes et de leurs produits (Ashburner *et al.*, 2000). Ceci a incité la communauté scientifique à développer des ontologies pour annoter les gènes et leurs produits. La science déploie une problématique ontologique lorsque vient à se poser la question du statut de la réalité des entités qui constituent le référent du discours scientifique » (Encyclopedia Universalis, 1991). Par extension, une ontologie est un vocabulaire structuré et contrôlé qui est une base au développement de la connaissance.)

Ainsi que, après la classification des gènes résultant de nos analyses, selon leurs fonctions, en se basant sur la GO (http://www.geneontology.org), il y a l'apparition de trois groupes de gènes : un groupe de gènes interviennent dans le métabolisme, un groupe des gènes interviennent dans l'angiogènes, et un groupe des gènes interviennent dans la forme de la cellule musculaire cardiaque (Cf. tableau 6).

Tableau 6 : *Résultats de la comparaison d'expression des gènes (gènes ont le même profile de CD36) selon la fonction , A : Métabolisme/signalisation de l'insuline, B : Angiogenèse/Apoptose, C : Remodelage de la cellule.*

A :

Gene Name	Description
CD36	Binds long chain fatty acids and may function in the transport and/or as a regulator of fatty acid transport
Dgat1	Catalyzes the terminal and only committed step in triacylglycerol synthesis by using diacylglycerol and fatty acylCoA as substrates
Adh4	Involved in the reduction of benzoquinones
Irs3	Insulin receptor substrate 3
Irs1	Maediate the control of various cellular processes by insulin.
IL2	Negative regulation of heart contraction
Fat4	Function in the regulation of planar cell polarity.
Pdhb	Complex catalyzes the overall conversion of pyruvate to acetyl-CoA and CO2.
Lpl	The apolipoprotein, APOC2, acts as a coactivator of LPL activity in the presence of lipids on the luminal surface of vascular endothelium
Ide	Plays a role in the cellular breakdown of insulin
Uqcrh	This is a component of the ubiquinol-cytochrome c reductase complex (complex III or cytochrome b-c1 complex), which is part of the mitochondrial respiratory chain.
Uqcrc1	This is a component of the ubiquinol-cytochrome c reductase complex (complex III or cytochrome b-c1 complex), which is part of the mitochondrial respiratory chain
Gyk	Key enzyme in the regulation of glycerol uptake and metabolism
Ifng	Regulation of the force of heart contraction
Mb	Heart development and Serves as a reserve supply of oxygen and facilitates the movement of oxygen within muscles.

B :

Gene Name	Description
CD36	Receptor for thrombospondins, THBS1 AND THBS2, mediating their antiangiogenic effects
Thbs1	Adhesive glycoprotein that mediates cell-to-cell and cell-to-matrix interactions. Ligand for CD36 mediating antiangiogenic properties
Sdc4	thrombospondin receptor activity and Cell surface proteoglycan that bears heparan sulfate
CD9	Involved in platelet activation and aggregation.
Pdgfra	vascular endothelial growth factor receptor signaling pathway
Hand1	Transcription factor that plays an essential role in both trophoblast-giant cells differentiation and in cardiac morphogenesis.
Arnt2	Negative regulation of apoptotic process and positive regulation of cell proliferation
Rag1	Negative regulation of cysteine-type endopeptidase activity involved in apoptotic process
Api5	Antiapoptotic factor that have a role in protein assembly. Negatively regulates ACIN1. Also known to efficiently suppress E2F1-induced apoptosis
Map3K2	Component of a protein kinase signal transduction cascade, Plays a role in caveolae kiss-and-run dynamics Mertk Receptor tyrosine kinase that transduces signals from the extracellular matrix into the cytoplasm by binding to several ligands

C:

Gene Name	Description
Myl4	Myosin light chain
Tnnt2	Atrial cardiac muscle tissue morphogenesis
Hand1	Ventricular cardiac muscle tissue morphogenesis and cardiac septum morphogenesis,
Pdlim3	Actin filament organization in heart development
Pdgfra	Cardiac myofibril assembly

7. Gènes candidats étudies

Selon la classification résultant de notre analyse des données de puces à ADN on constate qu'il y a trois classes de gènes : des gènes qui interviennent dans le métabolisme, des gènes qui interviennent dans le phénomène de l'angiogenèse et l'apoptose, et enfin des gènes qui interviennent dans le remodelage de la cellule musculaire cardiaque.

7.1.Métabolisme

La cellule musculaire cardiaque qui est caractérisée par une activité contractile requiert un apport énergétique ou, plus exactement un flux d'énergie élevé. Ce flux élevé et permanent d'énergie (assuré par la synthèse de molécules d'ATP) ne peut être entretenu que par un métabolisme du myocarde est illustré par l'abondance des mitochondries dans les cardiomyocytes. La cellule musculaire cardiaque ne dispose que d'un stock d'ATP réduit, ne permettant que quelques contractions. La permanence de la fonction cardiaque dépend donc de la capacité du cardiomyocyte de produire une quantité d'ATP elle est correspondante à la demande énergétique.

7.1.1. CD36

Le cœur est un organe riche en lipides en particulier en phospholipides (PL) membranaire en raison de sa riche masse membranaire, composée des mitochondries qui représentent prés de 30% du volume cardiaque (L. Opie et al., 1991). Ainsi la fonction principale de CD36 est son rôle comme récepteur/transporteur des acides gras à longue chaine dans les cellules cardiaques.

La capacité du cardiomyocyte à adopter sa production d'énergie à la demande est un facteur déterminant de la fonction myocardiaque, qui repose sur un métabolisme des acides gras (AG) équilibré. Les acides gras constituent la source énergique principale du cœur, mais couteuse en oxygène et susceptible d'entrainer des effets délétères.

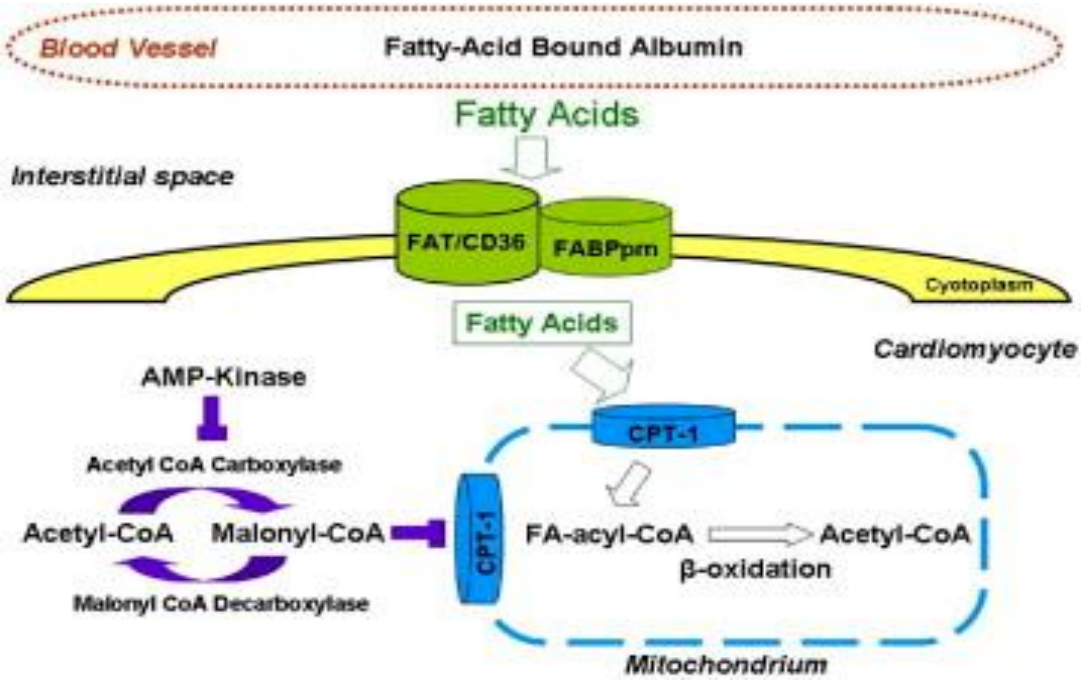


Figure 38 : CD36, transporteur des AGLC (L. Opie et al., 1991).

La principale voie d'utilisation des acides gras dans le myocarde est la β-*oxydation* mitochondriale destinée à la production d'énergie.

L'acide gras transformé en acyl_CoA par l'acyl_CoA synthétase, qui entre dans la mitochondrie grâce à la navette carnitine impliquant les carnitines palmitoyl transférase (CPT1 et CPT2), la β-*oxydation* oxyde l'acyl_CoA en acétyl_CoA qui entre dans le cycle des acides tricarboxyliques (ou cycle de Krebs). Le cycle produit les équivalents réduit (NADH2 et FADH2), qui permettent, via la chaine respiratoire de la membrane mitochondriale, la phosphorylation oxydative qui produit l'ATP à partir d'ADP.

L'absence de CD36 entraine une perte de l'absorption des acides gras à longue chaine, la source principale d'énergie du muscle cardiaque dans les cardiomyocyte , ce qui provoque une diminution de la disponibilité des acyl_CoA qui provoque une inhibition de la β-*oxydation*. Cet arrêt entraine une diminution de la cartinine libre et une accumulation intramitochondriale d'acyl_CoA et d'acyl_cartine a longue chaine. Ces constituants à longue chaine entrainent des altérations membranaire importantes. Ainsi que ces perturbations aboutissent à une baisse importante de la production énergétique sous forme d'ATP , et par conséquence une augmentation de survenues de cardiomyopathie hypertrophique (Guiraud A. , 2006).

7.1.2. IRS

Parmi les gènes on constate le gène IRS, c'est le gène qui code pour la protéine IRS (Insulin Receptor Substrat) impliqué dans la cascade de signalisation de l'insuline. Le signal intracellulaire stimulé par l'insuline passe notamment par les substrats du récepteur de l'insuline, dont les principaux sont IRS1 et IRS2, ces protéines une fois activées par le récepteur de l'insuline jouent le rôle de protéine de liaison, en liant et activant d'autres protéines impliquées dans le processus de signalisation déclenché par l'insuline (Sun et al. , 1991 ; Sun et al., 1995). Ces gènes qui sont impliqué dans la cascade de signalisation de l'insuline semblent être de bons gènes candidats pour jouer un rôle potentiel dans le développement de l'insulinorésistance.

Mais quelle est la relation entre le CD36 et les IRS dans la cellule musculaire cardiaque ?

Il s'avérée qu'un dysfonctionnement des métabolismes des lipides et particulièrement des acides gras à longue chaine associé souvent à une élévation de leur concentration plasmatique ou un stockage ectopique, est un élément essentiel à l'insulinoésistance (JD. McGarry et al., 2001 ; G. Boden et al.,2002).

En 1963, Rondle et ses collaborateurs on mit en évidence à partir d'expériences réalisées in vitro sur le cœur de rat isolé et perfusé, l'existence d'une compétition entre le glucosée et les acides gras à longue chaine, pour être préférentiellement métabolisé. Dans ces expériences une augmentation de la concentration dans le milieu de perfusion entrainant un défaut de captage et d'utilisation du glucose par les myocardes (PJ. Randle et al., 1963).

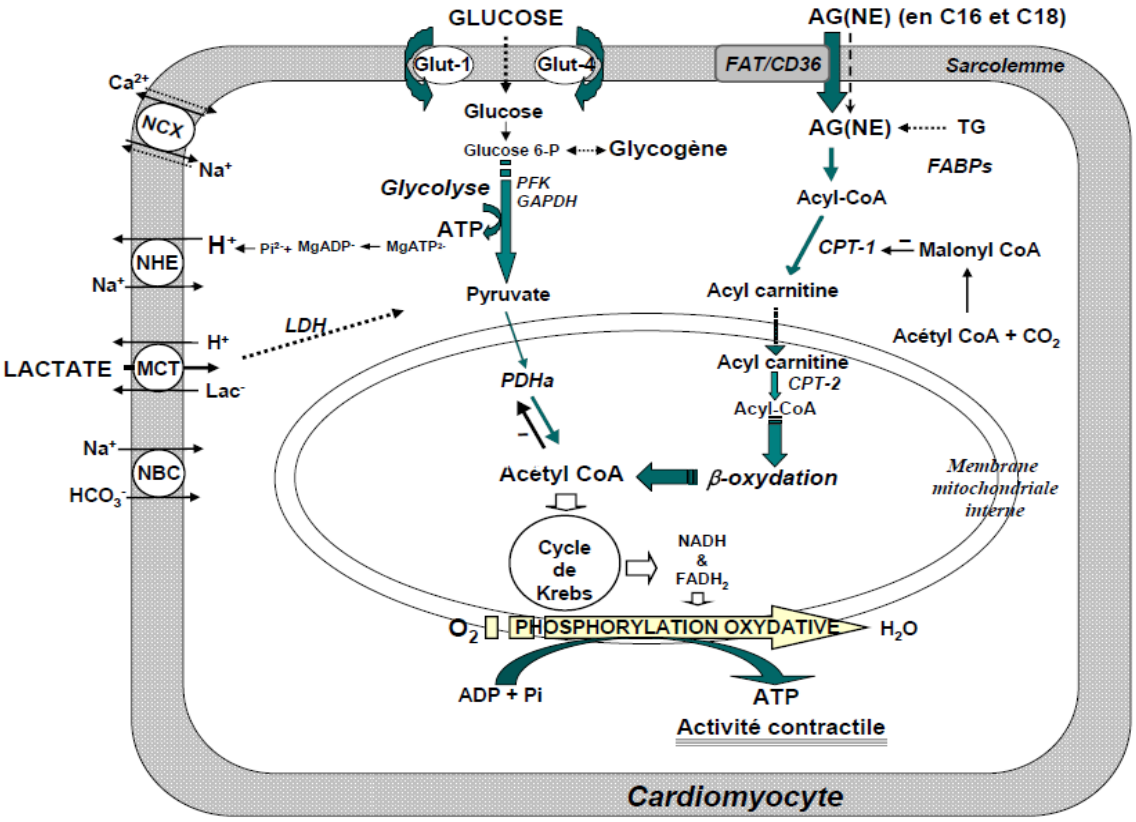


Figure 39: Mécanismes régulateurs du métabolisme cardiaque
(H.Taegtmeyer et al.,2004)

L'absence de CD36, entraine une accumulation intracellulaire importante des métabolites des acides gras à longue chaine : acyl_CoA, céramides diaglycerole, qui forment le DAG_P (acides gras estérifie ou glycérol) donc une augmentation de la concentration des triglycérides et des acides libre dans le plasma. Ces derniers activent toutes une série de protéine kinase C et aussi le IKK.

Parmi les nombreux substrats de ces kinases se trouvent le récepteur à l'insuline, d'importance cible de la voie de signalisation de l'hormone. Un des substrat de PCK est le IRS1 qui se trouve inactive. L'activation de PCK inhibe le IRS1 et entraine son incapacité à secréter et activer la PI3K, cela mène consécutivement à une baisse de la translocation du transporteur de glucose, Glut4 a la surface de la membrane plasmique. Alors une diminution des transporteurs transmembranaires de glucose liée à la carence en insuline produite par l'inhibition des protéines IRS, et dans ce cas les mycètes de cœur présentent un certain nombre de particularités liées au fait que la balance acides gras/glucose est largement déséquilibrée dans le sens d'une utilisation des acides gras. Ce qui provoque une myocardiopathie hypertrophique.

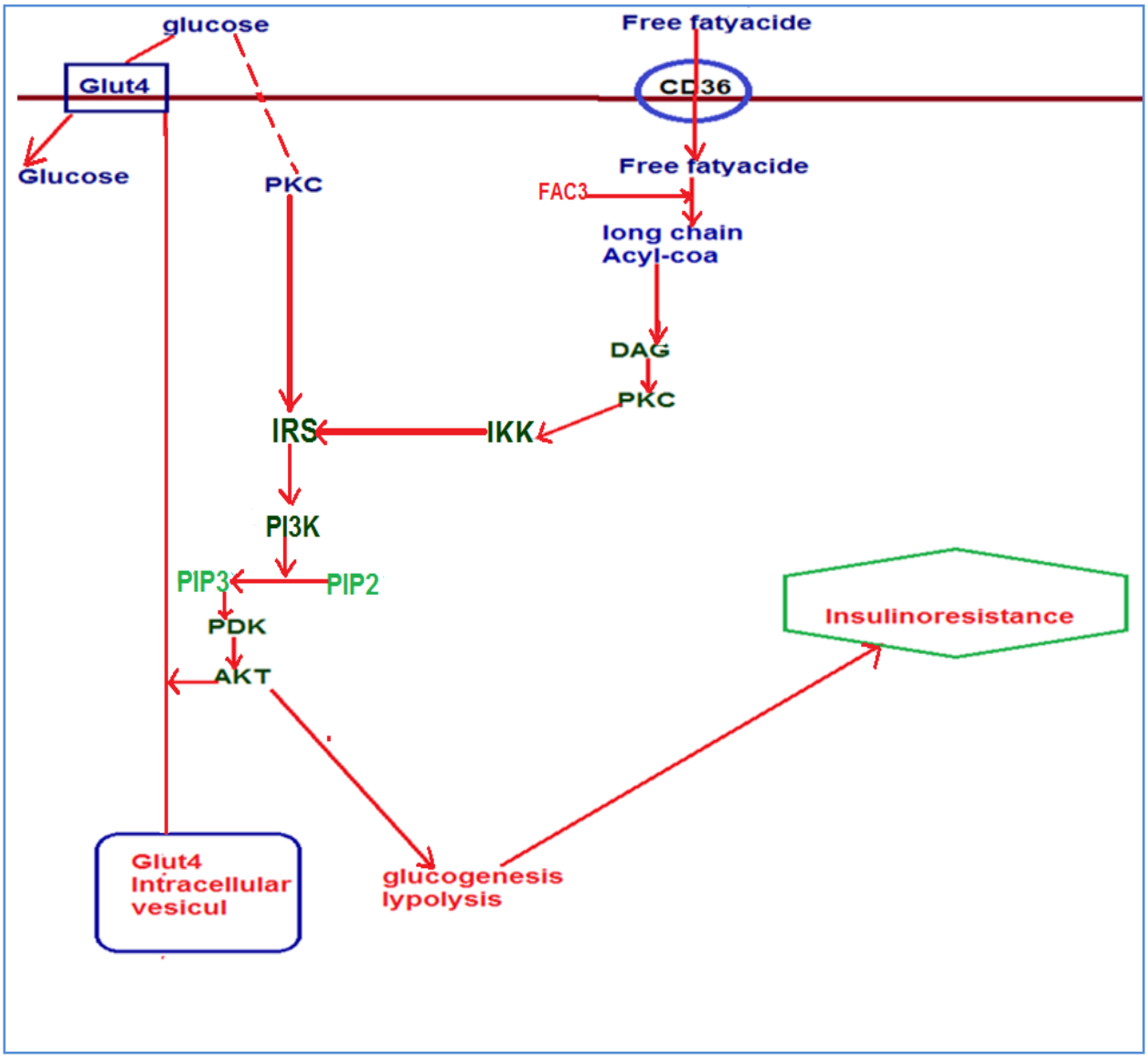


Figure 40 : Voie de fonctionnements de CD36 en relation avec IRS (http://www.sigmaaldrich.com)

7.2.Angiogenèse, apoptose, adhérence des cellules musculaires cardiaques

7.2.1. Thrombospondine-1 : TSP-1

Les cellules de la paroi vasculaire, cellules endothéliales, cellules musculaires lisses et fibroblastes synthétisent et secrètent la thrombospondine-1, qu'elles incorporent également dans la matrice extracellulaire. La thrombospondine-1 est une glycoprotéine de grande taille (450Da).

Bien que le renouvellement des cellules endothéliales soit physiquement d'extrêmement faible, des lésions vasculaires ou l'influence de modulateurs spécifiques induisent une migration et prolifération des cellules endothéliales conduisent à la formation de nouveaux vaisseaux à partir des vaisseaux à partir de vaisseaux préexistants par bourgeonnement et ramification suivant un processus dit d'angiogenèse.

Plusieurs travaux, à la suite des travaux pionnier Restinejad et al. (Restinejad et al., 2006), ont montré que la thrombospondine-1 se comportait comme un inhibiteur de l'angiogenése.

La TSP-1 ou certain de ces fragments exercent une action inhibitrice sur la migration et la prolifération des cellules endothéliales.

Initialement, le CD36 a été décrit comme l'un des récepteurs membranaires de la thrombospondine (Asch et coll, 1987).

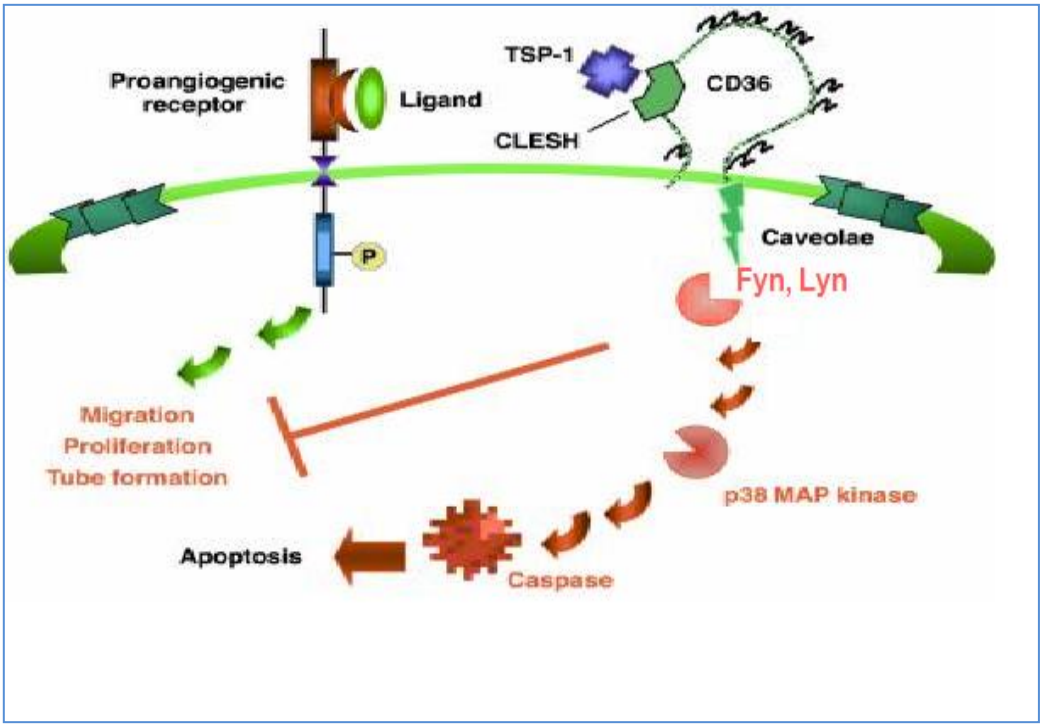


Figure 41 : CD36, TSP-1 molécule inhibitrice de l'angiogenèse (L. Opie et al., 1991).

TSP-1 active le CD36 qui recrute Fyn, Lyn qui active le P38 qui active des caspase qui tuent la cellule endothéliale.

7.2.2. CD9

L'existence de nombreux récepteurs cellulaires capables de fixer la thrombospondine tel que (alpha11belta3, alpha6belta3), a considérablement compliqué l'étude de système ligand-récepteur.

Cependant des observations montrent que CD36 ne contient pas de CxxC qui a été démontré qu'il est nécessaire pour que le LCK s'associe avec CD4 (sachant que le domaine cytoplasmique des CD36 est semblable à celui de CD4 et CD8) ce qui permet de penser qu'il existe d'autre protéines qui peuvent s'associe à CD36 pour faciliter la transduction de signal induit par la liaison CD36_TSP1 (Wei-Min Miao, et al. 2001)

Des études sur les plaquettes, (par solubilisation de la membrane plaquettaire) identifient CD9 et les intégrines alpha11belta3 et alpha6belta1, sont les protéines spécifiquement lié au CD36 (Wei-Min Miao, et al. 2001).

Des études d'immunofluorescence montrent que CD36, CD9, alpha11belta3 et alpha6belta1 colocalisent sur la membrane des plaquettes. Et ils ont constaté que CD9 et CD36 alpha6belt1 sont localisés sur les cellules endothéliales. Ainsi qu'ils ont montré que CD9, CD36, et alpha6belta1 forment un complexe sur les cellules endothéliales..

Le CD36 forme un complexe avec CD9 et alpha6belta1 sur le membrane des cellules endothéliale. Le TSP1, se lie au CD36, et l'effet de cette liaison se fait par l 'activation de P38.

Il a été démontré que CD36 est associé à la Src tyrosine protéine de famille kinase (M. Febbraio et al., 2001): Fyn, Lyn dans les cellules endothéliales. Et l'activation de la liaison TSP1-CD36, permet l'activation de P38 qui à son tour active le caspase8 qui aboutit à la mort cellulaire (apoptose) alors une inhibition de l'angiogenése. Le schéma suivant illustre le cas susmentionné

7.2.3. MAP3K2

La protéine codée par le gène MAP3K2 ou MEKK2 est un membre de la famille de la serine/thréonine kinase.

Les MAP Kinase sont des serines/thréonines kinases de 38 à 54 KDa, dont l'activation par phosphorylation de résidus thréonine (T) et tyrosine (Y) en une sequence T-X-Y, est l'aboutissement d'une cascade de protéine Kinase (Force T., Pombo CM., Avruch et al., 1996). L'expression de ces gènes est induite au cours de l'hypertrophie cardiaque mais ses caractéristiques demeurent inconus (Force T., Bonventre JV., 1998).

Plusieurs études récentes démontrent l'implication des protéines Kinase activées par le stress dans le développement de la réponse hypertrophique.

L'activation constitutive de P38 par l'expression de MAP3K2 (MEKK), la protéine kinase située en amont de P38 dans la cascade d'activation entraine une augmentation de la taille des cellules (Zechner D., Thuerauf DJ., Hanford DS. Et al., 1997). Ces chercheurs ont montré qu'il y a une association physique dans les cellules monocytaires intra-cytoplasmique entre le

domaine carboxyle terminale de CD36 et le complexe de signalisation contenant Lyn et MEKK2, en amont de P38 dans la cascade de MAP Kinase (S.O. Rahman et al., 2006).

8. Conclusion

En conclusion, cette étude donne une aperçu sur les mécanismes par lesquels la protéine CD36 intervienne dans la myocardiopathie observée chez les souris en absence de CD36, et de concevoir des stratégies thérapeutiques basées sur la régulation de l'expression des gènes pour inverser les principaux anomalies métaboliques liées a une carence en CD36

BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(17)

The Whole Ge nome Expression Analysis using T wo Microarray Technologies to Identify Gene Networks That Mediate the Myocardial Phenotype of CD36 Deficiency

Imane Sab aouni [1], Ahmed Moussa [2*], Brigitte Vannier [3], Oussama Semlali [1], Terri A Pietka [4], Nada A Abumrad [4] & Azeddine Ibrahimi [1]

[1]Medical Biotechnology Lab (MedBiotech), Rabat Medical & Pharmacy School, Mohammed V [th] Souissi University, Rabat, Morocco ; [2]LabTIC Laboratory, ENSA, Abdelmalek Essaa di University, Tangier, Morocco; [3]Receptors, R egulation and Tumor Cells (2RTC) Laboratory, University of Poitiers, France ; [4]Division of Biology and Biomedical Sciences, Washington University, USA ; Ahmed Moussa – Email: amoussa@uae.ac.ma ; *Corresponding author

Received January 27, 2013; Accepted September 30, 2013 ; Published October 16, 2013

Abstract :
We have previously shown that CD36 is a membrane protein that facilitates long chain fatty acid (FA) transport by muscle tiss ues. We also documented the significant impact of muscle CD36 expression on heart function, skeletal muscle insulin sensitivity as well as on overall metabolism. To identify a comprehensive set of genes that are differentially regulated by CD36 expression in th e heart, we used two microarray technologies (Affymetrix and Agilent) to compare gene expression in heart tissues from CD36 Kn ocK -Out (KO -CD36) versus wild type (WT -CD36) mice. The obtained results using the two technologies were similar with around 35 genes diff erentially expressed using both technologies. Absence of CD36 led to down -regulation of the expression of three groups of genes involved in pathways of FA metabolism, angiogenesis/apoptosis and structure. These data are consistent with the fact that the CD36 protein binds FA and thrombospondin 1 invoved respectively in lipid metabolism and anti -angiogenic activities. In co nclusion, our findings led to validate our data analysis workflow and identify specific pathways, possibly underlying the phenotypic abnormalities in CD36 Knock -Out hearts .

Keywords: CD36, Fatty Acid, Microarrays, Metabolism, angiogenesis/a poptosis, Pr otein interaction, G ene expression.

Background :
CD36 gene encodes a membrane glycoprotein, and has been identified in wide variety cells types, including platelets, monocytes, and erythroblast , capillary endothelial and mammary epithelial cells [1-6]. CD36 (also known as platelet glycoprotein IV or IIIb) is also a membrane glycoprotein highly expressed in heart tissue. It was shown that CD36 works as receptor/transporter of long chain fatty acids (FA) in muscle tissue and is proposed as one of thrombospondine receptor in endothelial cells [5]. CD36 -KO mi ce (with no expression of CD36 gene) exhibits defective FA uptake by the heart, which is paralleled by an increase in the heart/body index and by an enlargement of left ventricular space [1]. Two sets of studies were done to identify a comprehensive set of genes that are differentially regulated by CD36 expression in the heart. In 2002 and 2007, we used respectively the Affymetrix and the Agilent technologies to analyze CD36 involvement in the fatty acids uptake and heart hypertrophy [7, 3-5, 8]. We propose to compare results obtained from the two microarrays technologies and investigate the consequences of CD36 absence

on CD36 -KO hearts. In this paper, we will describe the methodology used to identify the differentially expressed genes using the Affymetrix and the Agilent technologies. In the second section, we will use the classification results to determine the gene clusters. Finally, these classifications will be used to annotate the functional class of each cluster and characterize the molecular pathway s involved in the myocardial phenotype of KO -CD36 mice.

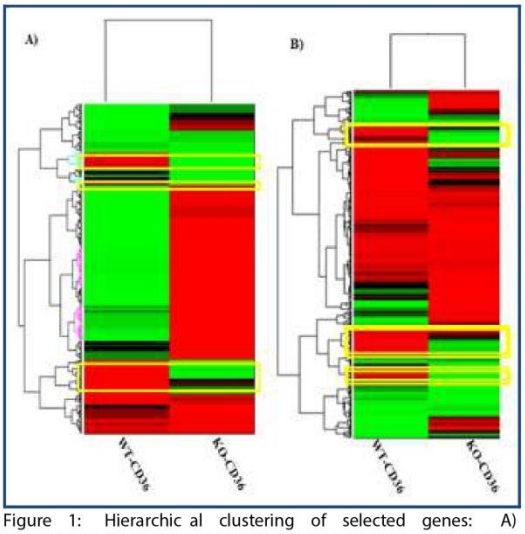


Figure 1: Hierarchic al clustering of selected genes: A) Hierarchical cluste ring using selected genes from Affymetrix Technology ; B) Hierarchical cluste ring using selected genes from Agilent Technology

Methodology :

Data analysis

In order To search for differentially regulated gene networks in the absence of CD36 gene, we performed a comprehensive gene analysis by hybridizing microarray chips with RNA probes prepared from mouse heart CD36 -KO and CD36 -WT. Two technologies were used. T he Affyme trix GeneChip Murine Genome U74 which contains 36.000 probes (Affymetrix, Santa Clara, CA). Once the probe array had been hybridized, stained, and washed, it was scanned using a GeneArray scanner. A GeneChip Operating System, running on a PC workstation wa s used to control the functions of the scanner and collect fluorescent intensity data. The second approach used was the Whole Mouse Genome Microarray 4x44K (Agilent Technologies, Santa Clara, CA). Arrays were washed and dried out using a centrifuge accordi ng to manufacturer's instructions (One -Color Microarray -Based Gene Expression Analysis, Agilent Technologies). Arrays were scanned at 5 mm resolution on an Agilent DNA Microarray Scanner (GenePix 4000B, Agilent Technologies) using the default settings for 4x44k format one -color arrays. Images provided by the scanner were analyzed using Feature Extraction software v10.1.1.1 (Agilent Technologies). Raw data files were analyzed using the software R associated to packages of "Bioconductor" project [8]. Develope d affymetrix workflows begin with data normalization using Robust Multichip Average Method (RMA) [9], which allows reduction of block effect done at the probset level. The second step consisted in selecting differentially expressed genes between CD36 -WT' and CD36 -KO using the Significance Analysis of Microarrays (SAM) algorithm [10] with the Fold Change a nd P -value Cutoffs respectively fixed to at 1.5 and 0.002. Agilent workflow begins with data normalization by using Lowess normalization [11] that applied to a two -color array expression dataset. The second step, as in the case of Affymetrix, SAM algorithm was used to identify differentially expressed genes with the same FC and P -value cuttoffs. Class discovery analyzed a given set of genes to produce subgroups that share common features. An analysis method often used for class discovery is "cluster analy sis" or clustering. It is aimed at dividing the data points (genes or samples) into groups (clusters) using measures of similarity [12, 13] creating hierarchical clustering of co -regulated genes

Identification of differentially expressed genes

Using Aff ymetrix technology, we were able to identify 39 differentially expressed genes between CD36 -KO and CD36 -WT and when using the Agilent technology with the same parameters, we identified 35 differentially expressed genes. The comparison of the two lists of i dentified genes showed that 30 of them were common to the two technologies. Differentially expressed genes were clustered by hierarchical clustering (Figure 1). This type of classification and class discovery involves analyzing a given set of gene expressi on profiles with the goal of discovering subgroups that share common features.

Functional Analysis

After establishing a differentially expressed gene list common to the two technologies (Affymetrix and Agilent), functional annotation allowed the determination of each gene function in the list. Functional classes were extracted using Gene Ontology tools. Associations, biochemical pathway data were retrieved from the Gene Ontology consortium (GO) [14, 15], Kyoto Encyclopedia of Genes and Genomes (KEGG) [16]. Gene sets were arranged and associated with different functional categories. Integrated expression data and evolutionary conservation of proteins were used to predict interacting proteins, protein complexes and proteins functions [17, 18, 19]. Table 1 (See supplementary material) showed the first classification leading to the identification of three gen e categories: a gene group involved in metabolism, another group involved in the Angiogenesis Table 2 (see supplementary material) and a third one implicated in cellular remodeling Table 3 (see supplementary material) (structural genes).

The results appear to be consistent with the role of the CD36 protein in cardiac muscle cells. The identification of the metabolism genes could be explained by the role of CD36 as a receptor / transporter for long -chain fatty acids in heart cells [3]. Inde ed, Randle et al. [20] showed that fatty acids are the main energy source for the heart and have brought to light from in vitro experiments that long chain fatty acids is preferentially metabolized. He also demonstrated the existence of competition between glucose and fatty acids as heart fuel [20]. Moreover, Dyck et al. showed that CD36 plays an important role in the choice of substrate in the heart [21]. In the CD36 -KO hearts, the shift to glucose substrate since less FA is available led to the expression and down regulation of genes

involved in both metabolisms. Secondly, in endothelial heart cells; CD36 has been described as a thrombospondin membrane receptor (TSP -1) and plays a role in the inhibition of endothelial cell migration and apoptosis induction . The absence of CD36 led to down regulation of the expression of its ligand (Tsp1) and the expression of new signaling genes. Finally and in order for CD36 -KO to go through heart hypertrophy, a set of remodeling genes is expressed and can be grouped into the category of structural genes as shown in Table 1 .

Discussion :

In this study, we compared results obtained from two technologies using the same analysis workflow. We first evaluated variance among replicates within each of the platforms and found low levels of variance and high correlation among the two platforms even though the two technologies were used at different labs and time. Agilent oligonucleotide technology was used in 2007 at Georges Washington University (SL, MO) and the Affymetrix U74Av2 technology was used in 2002 at Stony Brook University (Stony Brook, NY). Using SAM, we were not able to find any significant differences among the two platforms looking at their ability to detect differential gene expression between WT -CD36 and KO -CD36. Technological differences may influence the results of transcriptional p rofiling and are important to consider while using published results. However, and based on our study and given high -quality arrays and the appropriate normalization, the primary factor determining variance is biological rather than technological. The biol ogical conditions of the two experiments could explain the small diffrences in the obtained results. Questions remain regarding the importance of technology choice in evaluating the data generated and comparing among experiments from different laboratori es. One of the objective of this comparative study was to elucidate whether gene expression profiles are more influenced by biology or by technological artifacts. Eventhough, the two platforms are based on two distinct manufacturing technologies; a two -color cDNA spotted arrays (Agilent) and in situ synthesized oligonucleotide chips (Affymetrix), our results showed comparable results.

Conclusion :

Our comparative study led to the validation of a data analysis workflow and the identification of, at least, 30 genes involved in the phenotype of CD36 heart hypertrophy. More biological studies are needed to validate the expression of the identified genes using the qRT -PCR. These studies will be complemented with a modeling project based on constructing a bioinformatic platform. It will be reproducing the behavior dynamic of the system under normal conditions and automatically predicting the involvement of different gene networks in the development of pathologies such as cardiomyopathy related to the absence of CD36 protein.

Acknowledgment :

This work was carried out under intramural funding from the Stony Brook University and Mohammed the V th Souissi University to AI. We acknowledge support from Volubilis (French -Moroccan Grant) to AI, BV & AM. This work was also supported by a grant from the NIH for H3Africa BioNet to AI & AM. We would like to thank Microarrays facilities in Stony Brook and G eorge Washington Universities for conducting the arrays hybridization experiments.

References:

[1] Abumrad NA et al. J Biol Chem. 1993 268: 17665 [PMID : 7688729]
[2] Kashiwagi H et al. Blood. 1994 83: 3545 [PMID: 7515716]
[3] Abumrad N et al. J Lipid Res. 1998 39: 2309 [PMID : 9831619]
[4] Ibrahimi A & Abumrad NA, Curr Opin Clin Nutr Metab Care. 2002 5: 139 [PMID : 11844979]
[5] Ibrahimi A et al. Proc Nat Acad Sci U S A . 1996 93: 2646 [PMID : 8610095]
[6] Asch As et al. J Clin Invest. 1987 79: 1054 [PMID: 2435757]
[7] Hajri T et al. J Biol Chem. 2001 276: 23661 [PMID: 11323420]
[8] Bioconductor Project [http://www.bioconductor.org/]
[9] Irizarry RA et al. Nucleic Acids Res. 2003 31: e15 [PMID : 12582260]
[10] Tusher VG et al. Proc Nat Acad Sci U S A . 2001 98: 5116 [PMID : 11309499]
[11] Grosu P et al. Genome Res 2002 12: 1121 [PMID : 12097350]
[12] Aach J et al. Genome Res 2000 10: 431 [PMID : 10779484]
[13] Bethin KE et al. Mol Endocrinol. 2003 17 :1454 [PMID : 12775764]
[14] Doniger SW et al. Genome Biol 2003 4: R7 [PMID: 12540299]
[15] http://www.genome.jp/kegg/
[16] Voit E O & Riley M, Genome Biol 2003 4: 235 [PMID : 14611652]
[17] Ge H et al. Nat Genet. 2001 29: 482 [PMID:11694880]
[18] Jansen R & Gerstein M, Nucleic Acids Res. 2000 28: 1481 [PMID : 10684945]
[19] Teichmann SA & Babu MM, Trends Biotechnol 2002 20: 407 [PMID : 12220896]
[20] Randle PJ, Lancet. 1963 1: 785 [PMID : 13990765]
[21] Van Noort V et al. Trends Genet. 2003 19: 238 [PMID : 12711213]

Edited by P Kangueane
Citation : Sabouni et al . Bioinformation 9 (17): 849-852 (2013)

BIOINFORMATION open access

Supplementry material:

Table 1: Functional classification of differentially expressed genes : A: Metabolism genes

Gene name	Description
CD36	Binds long chain fatty acids and may function in the transport and/or as a regulator of fatty acid transport
Dgat1	Catalyzes the terminal and only committed step in triacylglycerol synthesis by using diacylglycerol and fatty acyl CoA as substrates
Adh4	Involved in the reduction of benzoquinones
Irs3	Insulin receptor substrate 3
Irs1	Maediate the control of various cellular processes by insuli n.
IL2	Negative regulation of heart contraction
Fat4	Function in the regulation of planar cell polarity.
Pdhb	Complex catalyzes the overall conversion of pyruvate to acetyl -CoA and CO_2.
Lpl	The apolipoprotein, APOC2, acts as a coactivator of LPL activity in the presence of lipids on the luminal surface of vascular endothelium
Ide	Plays a role in the cellular breakdown of insulin
Uqcrh	This is a component of the ubiquinol -cytochrome c reductase complex (complex III or cytochrome b -c1 complex), wh ich is part of the mitochondrial respiratory chain.
Uqcrc1	This is a component of the ubiquinol -cytochrome c reductase complex (complex III or cytochrome b -c1 complex), which is part of the mitochondrial respiratory chain
Gyk	Key enzyme in the regulation of glycerol uptake and metabolism
Ifng	Regulation of the force of heart contraction
Mb	Heart development and Serves as a reserve supply of oxygen and facilitates the movement of oxygen within muscles.

Table 2 : Angiogenesis / Apoptosis genes

Gene name	Description
CD36	Receptor for thrombospondins, THBS1 AND THBS2, mediating their antiangiogenic effects
Thbs1	Adhesive glycoprotein that mediates cell -to-cell and cell -to-matrix interactions. Ligand for CD36 mediating antiangiogenic properties
Sdc4	thrombospondin receptor activity and Cell surface proteoglycan that bears heparan sulfate
CD9	Involved in platelet activation and aggregation.
Pdgfra	vascular endothelial growth factor receptor signaling pathway
Hand1	Transcription factor that plays an essential role in both trophoblast -giant cells differentiation and in cardiac morphogenesis .
Arnt2	negative regulation of apoptotic process and positive regulation of cell proliferation
Rag1	negative regulation of cysteine -type endopeptidase activity involved in apoptotic process
Api5	Antiapoptotic factor that have a role in protein assembly. Negatively regulates ACIN1. Also known to efficiently suppress E2F1 -induced apoptosis
Map3K2	Component of a protein kinase signal transduction cascade, Plays a role in caveolae kiss -and-run dynamics
Mertk	Receptor tyrosine kinase that transduces signals from the extracellular matrix into the cytoplasm by binding to several ligands

Table 3 : Cellular remodeling genes (cell structure) .

Gene name	Description
Myl4	Myosin light chain
Tnnt2	atrial cardiac muscle tissue morphogenesis
Hand1	ventricular cardiac muscle tissue morphogenesis and cardiac septum morphogenesis ,
Pdlim3	actin filament organization in heart development
Pdgfra	cardiac myofibril assembly

III. Des protéines aux réseaux biologiques

Dans le but de soutenir certaines interactions prédites, nous avons exploré différentes approches basées sur les domaines d'interaction (article 3), les annotations fonctionnelles, la conservation à travers les organismes, les techniques de détection utilisées pour identifier les interactions sources, ou encore la mise en évidence expérimentale de certaines interactions.

1. Interactions protéine-protéine (IPP)

Les protéines sont l'un des principaux composants de la matière vivante. En effet, elles constituent la majeure partie de la masse sèche des cellules (Alberts, 1998) et sont impliquées dans de très nombreux processus allant de la protection de l'organisme à la réplication de l'information génétique, en passant par la transduction de signaux cellulaires.

Les protéines ne travaillent pas seules. En effet, la majorité des processus biologiques font intervenir plus d'une dizaine d'entre elles, chaque protéine interagissant avec une ou plusieurs autres protéines et formant ainsi des complexes protéiques transitoires ou permanents. Au sein de ce réseau d'interactions, toutes les protéines ne sont pas également connectées. En effets certaines n'interagissent qu'avec une protéine, alors que d'autre interagissent avec plusieurs centaines de protéines. Par analogie avec les réseaux de télécommunications, ces protéines centrales sont dénommées "hub" et sont particulièrement importantes pour le fonctionnement des cellules de par leur rôle central dans la formation de complexes (Jeong et al., 2001, Pang et al., 2010).

1.1. Mise en évidence d'interactions physiques entre protéines

Il existe de nombreuses techniques expérimentales pour mettre en évidence les interactions physiques entre protéines. L'une des premières méthodes haut débit développée est la méthode du "double hybride". Les paires des protéines dont on veut tester l'interaction sont exprimées sous forme de protéines chimériques. Sur l'une des deux protéines on ajoute un domaine de fixation à l'ADN et sur la seconde protéine un domaine activateur de la transcription. Si les deux protéines interagissent, la présence de ces deux domaines entrainera la transcription d'un gène rapporteur et donc la détection de la transcription (Ito et al., 2001). Cette technique est puissante car elle se déroule in vivo et permet de détecter des interactions même transitoires.

1.2. Mise en évidence d'interactions fonctionnelles

Il existe également des méthodes dites indirectes pour détecter des interactions entre protéines. On parle alors plutôt d'interactions fonctionnelles au lieu d'interactions physiques.

Une méthode indirecte utilisée pour les organismes procaryotes est la notion de voisinage génomique. Cette méthodologie est possible grâce à l'organisation en opérons des génomes procaryotes. Les opérons sont des ensembles de gènes voisins qui sont régulés par le même facteur de transcription et impliqués dans les mêmes voies biologiques. Ainsi, en observant que deux gènes sont très fréquemment voisins dans le génome de plusieurs organismes, il est probable que les protéines issues de ces deux gènes aient une interaction fonctionnelle (Overbeek et al., 1999).

Pour les organismes dont les gènes ne sont pas organisés en opérons, il est possible d'étudier les co-expressions de gènes. En effet, une conservation de la co-expression de 2 gènes dans de multiples organismes indique un avantage sélectif lors de l'évolution et donc que les protéines codées par ces gènes interagissent (Stuart et al., 2003) est une grande collection d'interactions de plusieurs sources de données qui identifient des gènes et des réseaux qui sont fonctionnellement associés (protéines et interactions, voies, coexpression, colocalisation et protéine Domaine) avec les ensembles de gènes de requête. Un autre avantage est que l'utilisateur peut exécuter ce comme un plugin avec l'outil Cytoscape permettant à l'utilisateur d'appliquer d'autres outils pour analyser les réseaux (Cf. Figure 42).

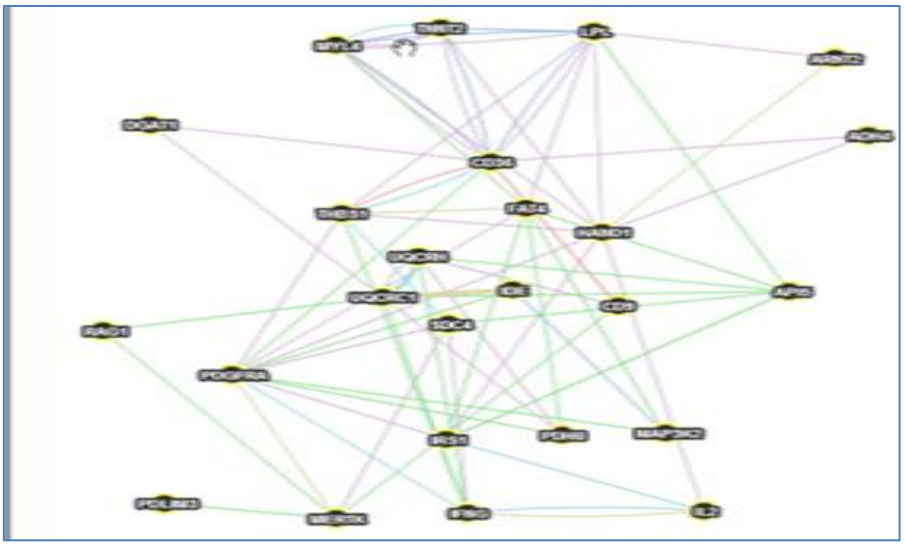


Figure 42 : Réseau d'interaction de la protéine CD36, extrait de la base de données GeneMANIA

Les interactions protéine-protéine peuvent être conservées entre les organismes proches (Walhout et al., 2000), on parle alors d'interologues (issue de la combinaison d'interaction et d'orthologue). En utilisant cette notion, on peut alors prédire des interactions en recherchant les interactions existantes dans des organismes proches.

1.3.Bases de données d'interactions protéine-protéine

Les différentes méthodes mentionnées ci-dessus permettent de mettre en évidence de plus en plus d'interactions protéine-protéine. Devant ce nombre croissant de données, de nombreuses

sources de données se développent afin de mettre les informations sur les interactions protéineprotéine à la disposition des biologistes. L'une des plus importantes est certainement la base de données IntAct[2] qui recense les interactions protéine-protéine décrites dans la littérature (Kerrien et al., 2012). Une autre source de données est la base STRING[3] qui reporte les interactions protéine-protéine élucidées expérimentalement ou prédites. La base de données STRING met à disposition des prédictions de relations fonctionnelles entre protéines (von Mering et al., 2003), (von Mering et al., 2005), (von Mering et al., 2007), (Jensen et al., 2008). Ces relations ne sont pas nécessairement des interactions physiques. En effet, elles proviennent de différentes méthodes de prédiction comme la co-expression, la co-localisation, la co-citation dans la littérature et l'identification expérimentale. De plus, toutes les relations fonctionnelles prédites sont également transférées, selon le principe des interologues, d'une espèce vers une autre. Ainsi, le transfert des relations identifiées expérimentalement correspond au principe de prédiction que nous avons utilisé. (Cf. Figure 43).

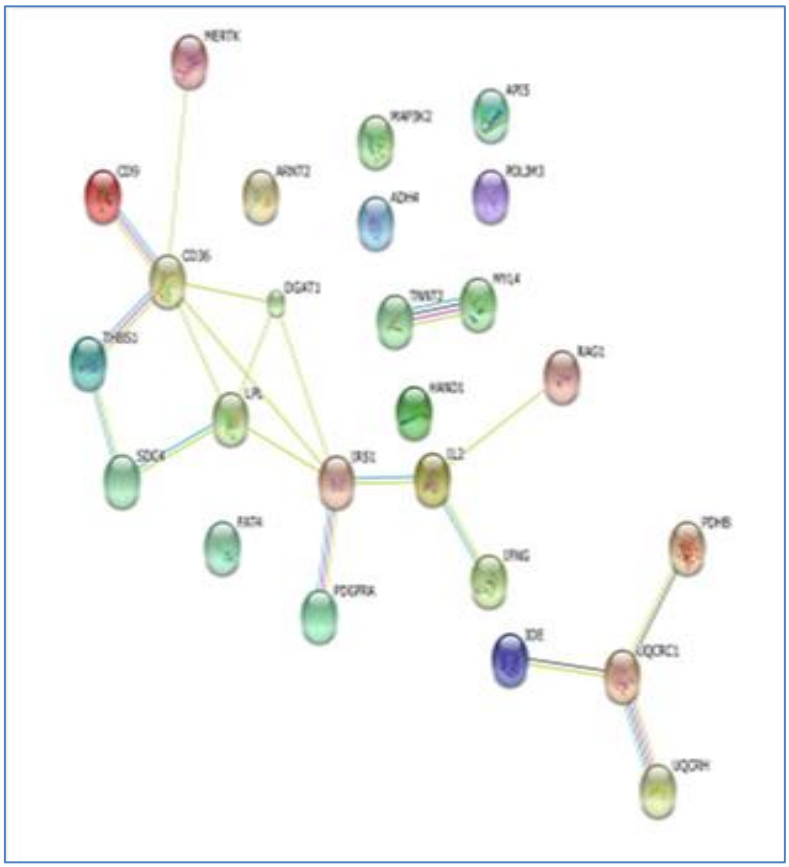
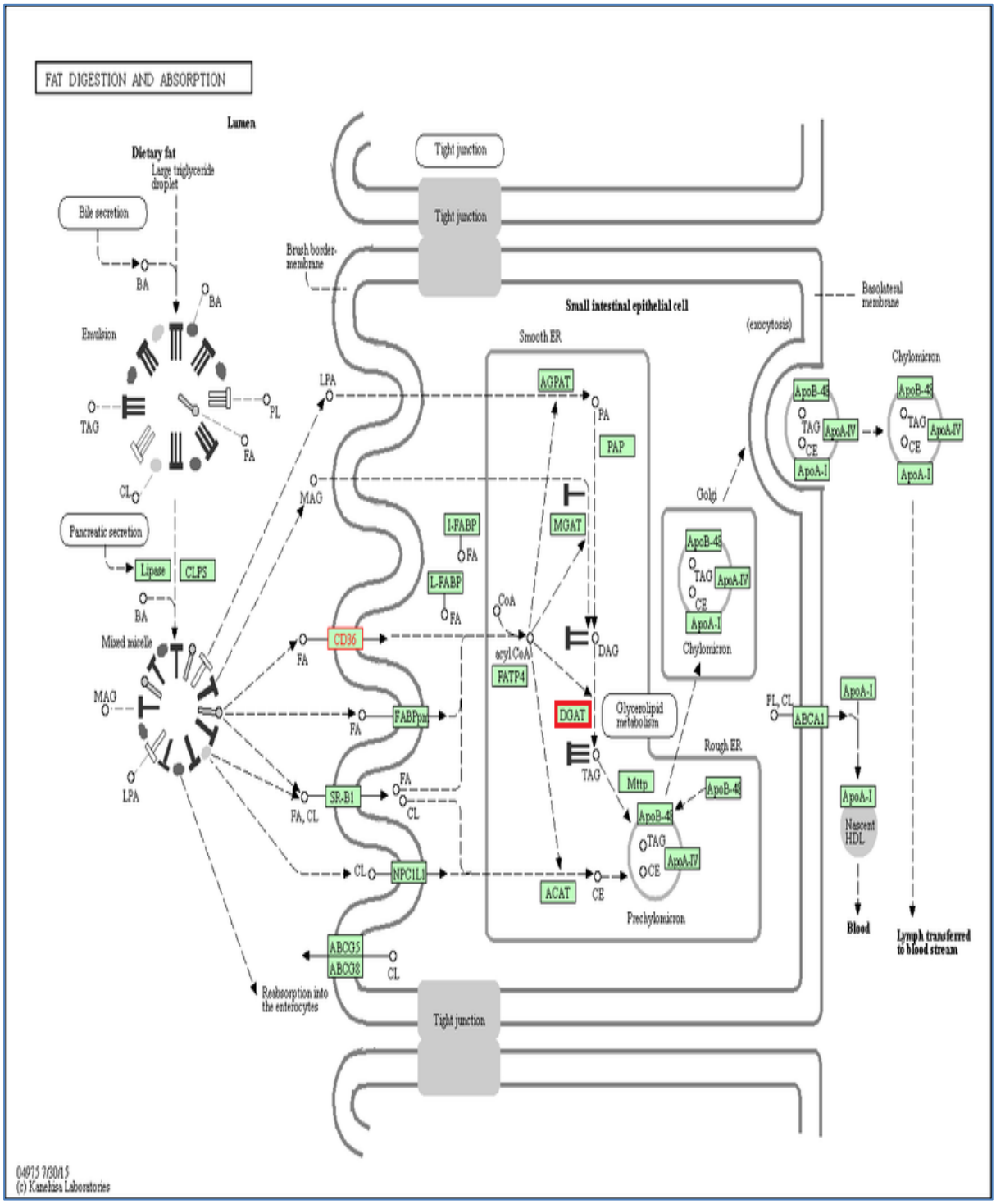


Figure 43 : Réseau d'interaction de la protéine CD36, extrait de la base de données String. Les relations en rose ont été élucidées expérimentalement, les relations bleues sont extraites d'autres bases de données et les relations vertes ont été prédites par fouille de texte

L'interactome peut être divisé en modules. Ainsi, Alberts (1998) compare ces sous-réseaux à des machines, dans lesquelles les protéines sont organisées en modules de manière à réaliser des fonctions précises. Au sein d'un module, les protéines sont fortement connectées entre

elles tandis que les interactions avec des membres extérieurs au module sont plus rares. Par exemple, on peut remarquer sur la figure 44 qu'au sein du module correspondant au métabolisme du FA il existe seulement des connections avec d'autres modules (Cf. Figure 44. a) et la même chose pour la résistance à l'insuline (Cf. Figure 44.b) .

a)



b)

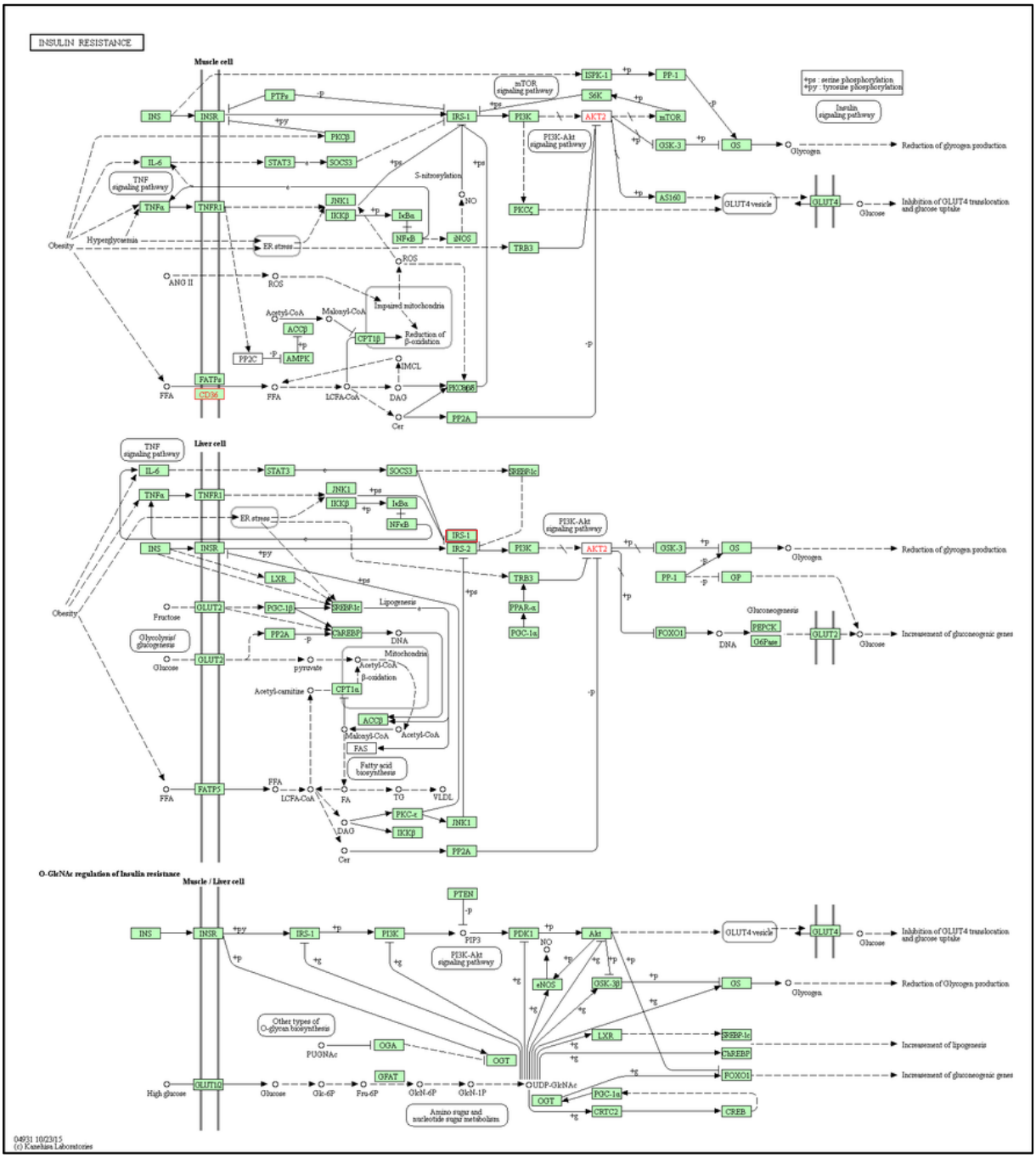


Figure 44 : Voie métabolique KEGG de référence correspondant a) au métabolisme des FA (Acide Gras a long chaine b) à la résistance à l'Insuline

Ainsi, cette organisation peut être utilisée pour définir des réseaux biologiques. Ces réseaux peuvent être définis comme un ensemble d'interactions entre des composants physiques ou génétiques de la cellule, décrivant un processus de cause à effet ou dépendant du temps, et expliquant des phénomènes biologiques observables (Demir et al., 2010). Différents types de processus sont décrits par ces réseaux : la régulation de gènes, le transport de molécules, la transformation de petites molécules ou une interaction entre protéines entrainant la modification de l'une d'entre elle (Schaefer et al., 2009). Les voies métaboliques décrivent le métabolisme de la cellule, comme par exemple le métabolisme des FA (Figure 44), et font

intervenir des petites molécules qui sont modifiées par les réactions chimiques réalisées par des protéines. Les voies de signalisation correspondent à la perception de signaux cellulaires ou extracellulaires, puis à la transduction du signal dans la cellule. Le troisième groupe de voies est associé à la régulation des gènes. Le plus souvent ces 3 voies sont connectées : la cellule perçoit un signal qui sera transmis jusqu'au noyau et entrainera une modification de la régulation de gènes, ce qui pourra provoquer des changements du métabolisme de la cellule.

1.4.Outils de visualisation : Cytoscape

L'ensemble des interactions d'un organisme forme un réseau d'interaction protéine-protéine. Ce réseau est un outil important pour la compréhension des mécanismes cellulaires (Agapito et al., 2013). Étant donné que la visualisation des réseaux peut permettre de mettre en évidence des sousstructures intéressantes, comme des complexes protéiques, la visualisation de réseaux sous forme de graphes est particulièrement répandue (Ciofani et al., 2012, Cheng et al., 2012, Agapito et al., 2013, Campillos et al., 2008). Ainsi, de nombreux outils sont développés afin de filtrer et d'analyser les réseaux biologiques. Ces outils varient notamment sur les formats de données qu'ils utilisent, le mode de visualisation, la possibilité d'interroger des sources de données distantes, la possibilité d'annoter un réseau existant et la capacité de l'utilisateur à développer de nouvelles fonctionnalités au programme via des extensions.
Cytoscape est l'un des outils de visualisation 2D de réseaux les plus populaires (Smoot et al., 2011, Cheng et al., 2012, Agapito et al., 2013). Ce programme permet de visualiser des réseaux allant jusqu'à des centaines de milliers de nœuds et de liens (Figure 45). Le principal objectif de Cytoscape est la visualisation d'interactions moléculaires et leur intégration avec des profils d'expression génique ou d'autres données. Cytoscape permet d'importer des réseaux existant sous différents formats : XML, RDF, OWL, GraphML, XGMML et SBML. Cytoscape permet également d'importer des réseaux à partir de bases de données distantes notamment via IntAct et les bases du NCBI. Cytoscape est capable d'importer des données locales ou distantes (notamment GO) pour annoter les réseaux affichés. Il existe de nombreuses extensions développées par les utilisateurs, certaines permettant notamment de calculer l'enrichissement en termes GO du graphe étudié ou d'un sous graphe (Maere et al., 2005). Il existe également d'autres extensions permettant d'interroger des bases de données distantes afin d'enrichir le réseau affiché ou pour simplement créer un nouveau réseau.

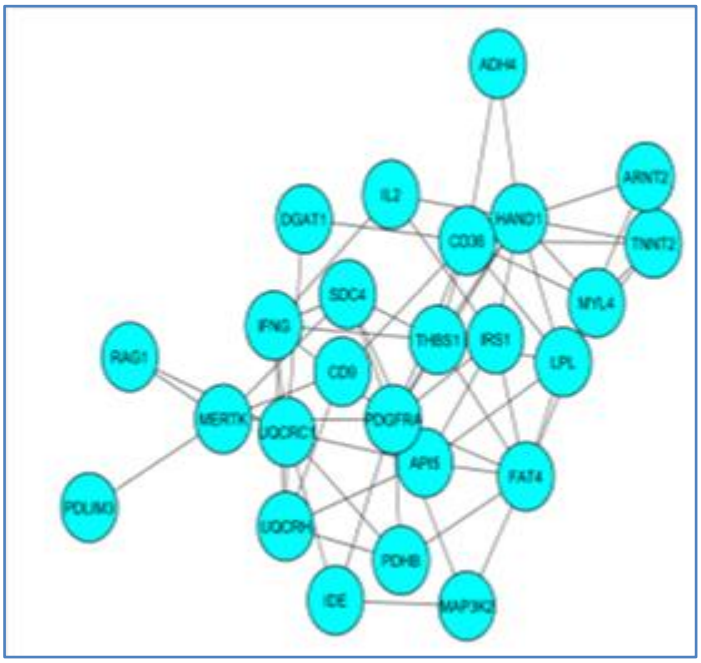


Figure 45 : Réseau d'interaction de la protéine CD36par Cytoscape.

2. Conclusion

Les réseaux biologiques sont de plus en plus utilisés pour comprendre des phénomènes biologiques complexes, comme les maladies génétiques. De nombreuses ressources publiques mettant à disposition les données concernant les réseaux biologiques sont disponibles. De nombreux formats de description de ces réseaux existent et sont tous plus ou moins spécifiques à un problème donné. A cela s'ajoute différents programmes permettant de visualiser et d'analyser ces réseaux. Ainsi, devant cette multitude de ressources et d'outils, l'utilisateur biologiste peut avoir du mal à s'y retrouver. Afin de les aider, il est intéressant d'extraire les informations utiles pour résoudre un problème donné à partir de diverses sources de données. De plus, il est important de pouvoir stocker et tracer l'origine des données ainsi collectées. Cette étape d'intégration et de structuration des données peut se faire en construisant un entrepôt de données. Les types de questions biologiques auxquels les réseaux sont susceptibles de répondre sont assez différents. De fait, cela nécessite de créer un entrepôt de données dont le modèle est suffisamment générique pour pouvoir s'adapter à ces problèmes variés. L'exploitation de ces données peut ensuite se faire via des outils de visualisation existants. Cependant, des méthodes d'analyse plus poussées comme les méthodes de fouille de données, peuvent être utiles afin d'extraire de nouvelles connaissances qui nous permettons de mieux comprendre le mécanisme de la cardiomyopathie observée chez les souris en absence de l'expression du CD36.

BIOINFORMATION
Discovery at the interface of physical and biological sciences
Open access
BIOMEDICAL INFORMATICS
www.bioinformation.net
Volume 12(6)
Hypothesis

Microarray Integrated Analysis of a Gene Network for the CD36 Myocardial Phenotype

Imane Sabaouni [1,*], Brigitte Vannier [2], Ahmed Moussa [3] & Azeddine Ibrahimi [1]

[1]Medical Biotechnology Lab (MedBiotech), Rabat Medical & Pharmacy School, Mohammed Vth University in Rabat, Morocco; [2]Receptors, Regulation and Tumor Cells (2RTC) Laboratory, University of Poitiers, France; [3]LabTIC Laboratory, ENSA, Abdelmalek Essaadi University, Tangier, Morocco; Imane SABAOUNI – E-mail: imanesabaouni@yahoo.com *Corresponding author

Received July 18, 2016; Revised July 29, 2016; Accepted July 30, 2016; Published October 11, 2016

Abstract :
CD36 is a multifunctional membrane -type receptor glycoprotein that reacts with oxidized low -density lipoprotein and long -chain fatty acid (LCFA). However, much remains to be understood about the molecular mechanism of the cardio -myopathy observed in CD36 -KO mice. In this study, we identify different genes pathways involved in response to CD36 cardio -myop athy phenotype by identifying the differences among biological processes, molecular pathways and networks of interactions that emerge from knocking CD3 and using different bioinformatics tools such as STRING, GeneMANIA and Cytoscape. We were able list all the CD36 -regulated genes, their related function and their specific networks. Data analysis showed that CD36 -regulated genes differentially expressed are involved in biological processes such as FA metabolism, angiogenesis/apoptosis and cell structure. The se results provide the first look at mechanisms involved in CD36 deficiency and development of cardio -myopathy and the opportunity to identify new therapeutic targets.

Keywords : CD36, cardio -myopathy, genes networks , genes pathways, m etabolism, angiogenes is/apoptosis.

Background:
Hypertrophic cardio -myopathy ("HCM") is characterized by a myocardium hypertrophy .To date the molecular mechanisms underlying this pathology remain elusive [1-6]. It is most well known as a leading cause of sudden cardiac death in young athletes [7]. The occurrence of hypertrophic cardio -myopathy is a significant cause of sudden unexpected cardiac death in any age group and is a cause of disabling cardiac symptoms. Younger people are likely to have a more severe form of hypertrophic cardio -myopathy. HCM is frequently asymptomatic until sudden cardiac death. As a c onsequence, re -curent screening has been suggested for certain populations [8].

The CD36 gene encodes a membrane glycoprotein (also known as pla telet glycoprotein IV or IIIb), that is expressed in a wide variety cell types, including platelets, monocytes, erythroblasts, capillary endothelial cells and mammary epithelial cells [9-10]. In the heart CD36 is also expressed both in epithelial and muscle cells [11]. In these latter, CD36 has been shown to be a long chain fatty acids (FA) receptor/transporter [12]. In endothelial cells, it has been proposed as one of thrombospondin receptor [13]. CD36 is a multifunctional membrane -type receptor glycoprotein that reacts with thrombospondin, collagen, oxidized low -density lipoprotein and long -chain fatty acid (LCFA) .1,2 LCFA is one of the major cardiac energy substrates; hence, LCFA metabolism may have an important role in cardiac diseases. We present here a patient with hereditary hypertrophic cardiomyopathy (HCM) and type I CD36 deficiency that showed no myocardial LCFA accumulation.

CD36 -KO mice exhibit between 60 and 80% reduction in beta-methyl -p-[123I] -iodophenyl -pentadecanoic acid (BMIPP) uptake by heart tissue [14] which is paralleled by an increase in the heart/body index and the left ventricular size [13]. We propose in this study to define the molecular defects that lead to cardiac hypertrophy and consequent of the fatty acid transporter CD36 deficiency. To define this molecular defect, we used two microarrays technologies (Affymetrix, Agilent) in order to identify genetic alterations related to the myocardial phenotype of CD36 -Ko mice.

In a previous study [15], heterogeneity of the arrays data was analyzed by splitting them on three levels: genes, genes set, and network/pathway. We were able to identify th e CD36 -regulated candidate genes linked to Hypertrophic cardio -myopathy . At the second level we were able to cluster group of genes (gene sets) that may have similar functions [15]. In this study, we aim to ascertain whether each gene set from each subtype is significantly

ISSN 0973 -2063 (online) 0973 -8894 (print)

!

enriched in a list of selected phenotypes. The third level of analysis was to construct gene networks of the proposed CD36 regulated genes from three selected web -based tools: Ingenuity Pathway Analysis (IPA), Search Tool for The Retrieva l of Interacting Genes (STRING), and Gene Multiple Association Network Integration Algorithm (GeneMANIA).

Materials and Methods :

Dataset :

To identify a comprehensive set of genes that are differentially regulated by CD36 expression in the heart, we used t wo microarray technologies (Affymetrix and Agilent) to compare gene expression in heart tissues from CD36 KnocK -Out (KO -CD36) versus wild type (WT -CD36) mice. The obtained results using the two technologies were similar with around 35 genes differentially expressed using both technologies [15]. Absence of CD36 led to down -regulation of the expression of three groups of genes involved in pathways of FA metabolism, angiogenesis/apoptosis and structure. These data are consistent with the fact that the CD36 protein binds FA and thrombospondin 1 involved respectively in lipid metabolism and anti-angiogenic activities [15]. Summary of dataset analysis methodology used in this study is shown in Figure 1 .

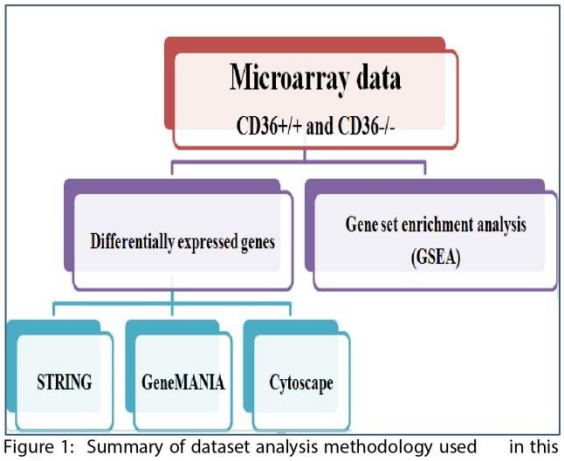


Figure 1: Summary of dataset analysis methodology used in this study

Gene Set Enrichment Analysis and Enrichment Map

Gene set enrichment analyses (GSEA) [16, 17] are commonly used to determine the biological characterization, statistical significance, and concordant differences between an experimental gene set and a selected gene list from annotated gene sets knowledge bases to red on Molecular Signatures Database (MSigDB). GSEA can be downloaded from http://www .broadinstitute.org/gsea/downloads.jsp. The Jaccard coefficient is used to compare the similarity between two sample gene sets A and B and defined as the intersection between group A and B divided by their union. The results from GSEA are then visualized through the enrichment map [18], a Cytoscape plugin for network visualization. The ranked experimental gene list along with the enriched gene sets from GSEA is used to build the network of gene sets (nodes) where edges represent their similarity. The size of a node varies by gene set size and the thickness of the edge represents the degree of correlation between two gene sets.

Networks and Pathway Analysis

Analysis of Network Invoked by CD36 -Regulated Gene s

The biological knowledge of gene and protein interactions is growing rapidly and there are many tools and curated databases available on a large scale. Insightful knowledge gained from studying gene sets rather than individual genes using network -based approaches can reveal network patterns and relevant molecular pathways from the experiment gene sets. In this study, we utilized two different freely accessible and user -friendly web tools as follows. Gene Multiple Association Network Integration Algorithm (GeneMANIA) [19, 20] (http://www.genemania.org/) is a web-based tool for prediction of gene function or implemented as a Cytoscape plugin tool. Based on single gene or gene set query from 7 organisms, it shows results for interactive functional associative network according to their co -expression data from Gene Expression Omnibus (GEO), physical and genetic interaction data derived from BioGRID, predicted protein interaction data based on orthology from I2D, co -localization, shared protein domain, and GO fu nction. Search Tool for the Retrieval of Interacting Genes (STRING) version 9.1 [21, 22] (http://string -db.org/) is an online protein -protein interaction database curated from literature and predicted associations from systemic genome comparisons. The user can query using single or multiple name(s) and protein sequence(s). The protein interactions can be displayed according to their confidence, evidence, actions, or interactions.

Results and Discussion :

Identification of CD36 -Regulated Genes through a Refined Analysis of Data

Our initial analysis of this time series gene expression data for differentially expressed gene identification followed the same method used in a previous paper [15]. All files were processed and normalized by Robust Multiarray Ave rage (RMA) in R as in the original study. The selected normalization method may have an effect on downstream analysis, for example, reverse engineering analyses [15]; however, investigating this effect is beyond the scope of this study. We found that combi ning CD36+/+ and CD36 -/- data compromises the accuracy of selection of d ifferentially expressed. Table 1 summarizes the 30 differentially expressed genes from the combined dataset.

Networks and Pathways Analysis

Network analysis can help understand the mo lecular and cellular interactions [23]. It can be visualized to represent entities (nodes) and their relationships (arcs). The advent of high -throughput technology has led to a large increase in publicly available information. Each data type can capture different aspect of functional roles of i nterested genes. In this section, we investigated functional interaction among genes a nd proteins in the cell using a vailable data and knowledge bases.

ISSN 0973 -2063 (online) 0973 -8894 (print)

!

BIOINFORMATION
Discovery at the interface of physical and biological sciences
Open access

We selected two different web -based network tools: GeneMANIA and STRING using the differentially expre ssed genes as a query gene sets. These toolsets integrates computational methods to predict the gene functions based on a collection of interaction networks. GeneMANIA is a large collection of interaction networks from several data sources which identify g enes and networks that are functionally associated (protein and genetic interactions, pathways, coexpression, colocalization, and protein domain similarity) with the query gene sets. Another advantage is that the user can run this as a plugin with the Cyto scape tool allowing the user to apply other tools to analyze the networks (Figure 2). STRING relies on the phylogeny to infer the functional interaction (protein networks) with direct interaction to score nodes, while GeneMANIA uses functional genomic data with label propagation to score nod es and generate gene networks (Figure 3). STRING uses precomputed networks, while those of GeneMANIA are not precomputed, and user can upload their own networks. STRING covers a large number of organisms, while GeneMANIA only covers 7 organisms but allows the user to upload or add more networks t hrough the plugin. In addition, users can run enrichment analysis (GO, KEGG, PFAM, INTERPRO, and protein -protein interaction) on STRING (Figure 4).

The results of genes and network that are functionally associated with the gene set from early response of CD36 are shown in Figures 3 and 4. We compared the two networks from STRING and GeneMANIA based on the interactions they revealed; here we used CD36 as the centre gene in the comparison. All interactions found in STRING were found in GeneMANIA as described in more detail in Table 2 (Cf. Annex). Comparison between the results from both tools can be used to confirm the functional associations of the interested gene sets. There is evidence of overlap and uniqueness in the interactions revealed by the two web -based tools.

Discussion

In this report, we were able to show that our data is consistent with the role CD36 protein in the FA metabolism, angiogenesis/apoptosis and structure. In cardiac muscle cells, CD36 is known for its lead role as a receptor / transporter of long-chain fatty acids in heart cells and involvement in metabolism in general. Indeed, the ability of cardiomyocytes to adapt its energy demand is a determinant for myocardial function and the Fatty acids are the main energy source of the heart with oxygen. Thus, the CD3 6 absence results in a loss of absorption of long chain fatty acids and cause deleterious effects on the heart muscle, while a decrease in the availability of acyl -CoA causes inhibition of B -oxidation. This leads to a decrease in free carnitine and accumul ation of intra -mitochondrial acyl -carnitine long chain and an important membrane alteration. These disturbances result in a significant decrease in energy production as ATP, which causes the use of an alternative energy source such as glucose. In 1963, Ran dle and his colleagues [24], have brought to light, using in vitro experiments on rat isolated and perfused hearts, the existence of competition between glucose and fatty acids with long chain fatty acids being preferentially metabolized. Moreover, accordi ng to Dyck et al. [25], CD36 plays an important role in the choice of substrate in the heart [25]. WT -CD36 which normally draws more of its energy from fatty acids is forced to uses more glucose in the absence of CD36 (The heart of the CD36 -KO). Our work c onfirms these results as we see the over -expression of genes that stimulate glucose metabolism such as IRS1 , IRS2, IRS3, IDE etc.

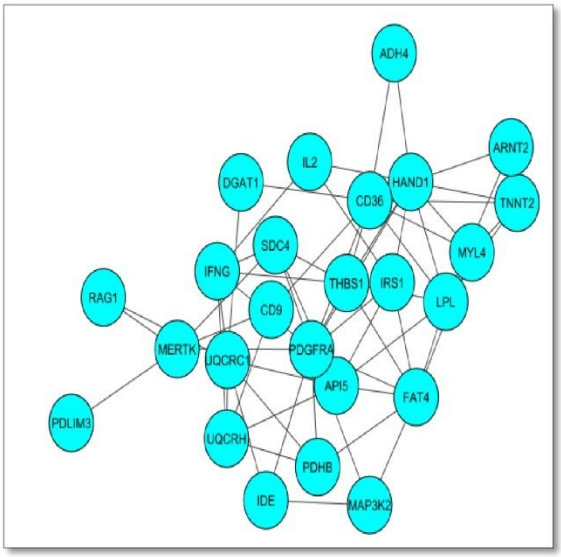


Figure 2: Interaction network Visualization using Cytoscape [2]. Cytoscape and enrichment map were used for visualization of the results; only gene sets from Gene Ontology were used. Nodes represent enriched GO gene sets, whose size reflects the total number of genes in that gene set. Edge thickness represents the num ber of overlapping genes between gene sets calculated using Jaccard coefficient.

On the other hand, in endothelial cells where angiogenesis/apoptosis are critical mechanisms for cell survival, CD36 has been described as a thrombospondin membrane receptor (TSP -1) [26]. TSP -1 plays a role in the inhibition of endothelial cell migration a nd apoptosis induction. Observations indicate that CD36 does not contain the motif CxxC shown to be necessary for the LCK -CD4 association, which allows to think that other proteins can be associated to CD36 to facilitate signal transduction induced by the CD36_TSP1 binding [26].

Our results confirmed that showing that CD9 that encodes the CD9 protein and CD36 share the same expression profiles. Studies on platelets (for solubilization of the platelet membrane) identify CD9; a11b3 and a6b11 integrins are CD36 partners [27]. Immunofluorescence studies show that CD9 and CD36 alpha6belt1 are co -located in endothelial cells. Thus it suggests that CD9, CD36, and alpha 6 beta 1 might form a complex in endothelial cells upon the binding of TSP1. Ou r data also a showed a differential expression of MAP3K2 (also known as MEKK2), a member of the serine /threonine kinase family, and

ISSN 0973 -2063 (online) 0973 -8894 (print)

BIOMEDICAL INFORMATICS

!

an activator of P38 which in turns activates a cascade that results in increased cell size [28]. In monocytic cells, the same results were observed since it was determined the existence of a physical association between the intra -cytoplasmic carboxyl terminus domain of CD36 and a signaling complex containing Lyn and MEKK2, upstream of the P38 MAP kinase cascade was de [29].

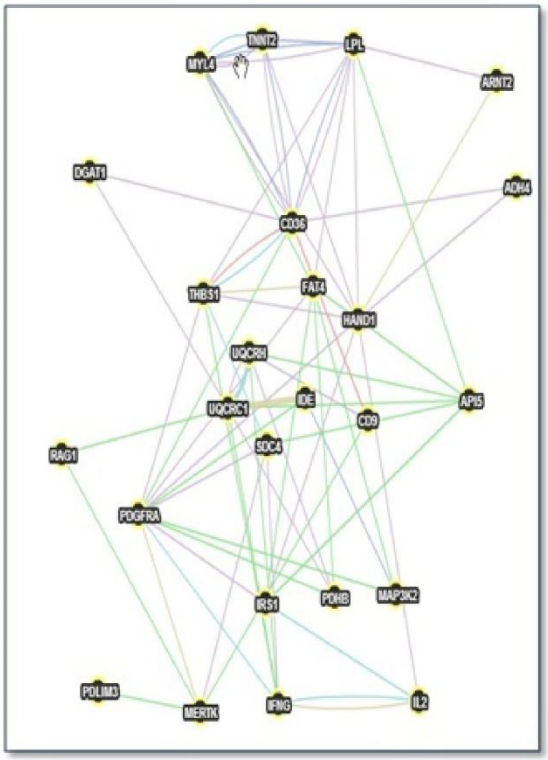


Figure 3: Gene network of CD36 derived from GeneMANIA . A gene network from GeneMANIA shows the relationships for genes from the list (nodes) connected (with edges) according to the functional association networks from the databases.

Data analysis showed that many Structural genes were altered in CD36 -KO. Indeed, failing heart differs from the normal heart in function as well as in structure as failing heart is most often remodeled with hypertrophy. Cardiac imaging with the increases of the ventricular wall thickness and smaller ventricular chambers can clinically recognize hypertrophy . Our work confirms cardiac such remodeling as we see the expression of genes that stimulate cardiac cell structure such as: MYL4, TNNT2, HAND1, PDLIM3 AN D PDGFRA.

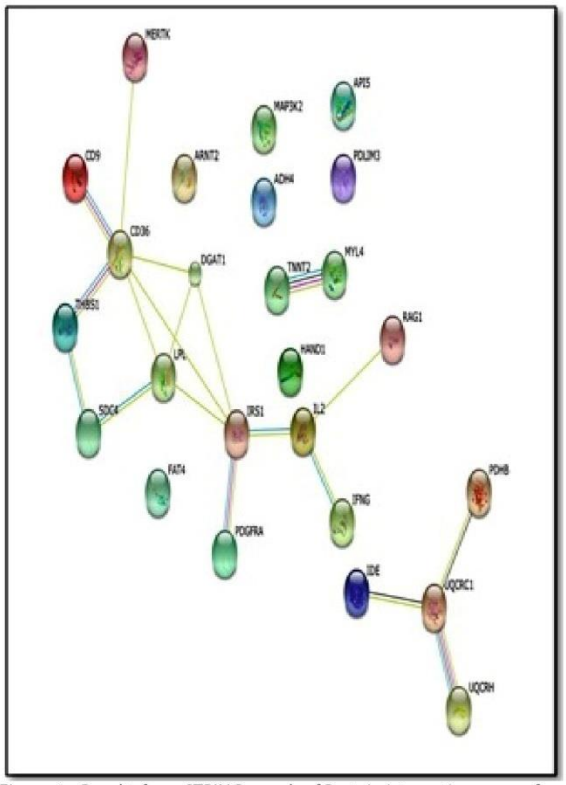


Figure 4: Results from STRING search of Protein interaction for CD36. The figure illustrates the protein interaction upon querying STRING protein network (evidence view) in Mus Musculus with 25 proteins. Additional information from other resources can be retrieved for each protein and interaction. Nodes represent proteins and different line colours denote the type of evidence for the interaction.

Conclusion

In summary, we utilized the strengths of existing network/pathway tools and datab ases to gain insight into processes related to cardio -myopathy have distinct molecular interaction patterns visible from various systems levels, including gene (microarray analysis), gene set, molecular pathway, and gene networks have shown that cardio -myopathy could involve three pathways; FA metabolism, angiogenesis/apoptosis and structure. Discriminating between the three pathways can help to improve the understanding of a drug's mechanism and further improve targeting in therapeutics drug research.

References

[1] Richardson P, et al. 1996 Circulation 93(5): 841 [PMID 8598070]
[2] Maron BJ, 2002 JAMA 287(10): 1308 [PMID 11886323]
[3] Sherrid MV et al. 2003 Ann Thorac Surg. 75(2): 620 [PMID 12607696]
[4] Wigle ED et al. (1985). 28(1): 1 [PMID 3160067]
[5] Wigle ED et al. 1995 92(7): 1680 [PMID 7671349]

ISSN 0973 -2063 (online) 0973 -8894 (print)

[6] Maron BJ et al. 2003 J Am Coll Cardiol 42(9): 1687 [PMID 14607462]
[7] Maron BJ et al. 1996 Circulation 94(4): 850 [PMID 8772711]
[8] Farzin Halabchi , Tohid Seif -Barghi , and Reza Mazaheri , 2011 -Sudden Cardiac Death in Young Athletes; a Literature Review and Special Considerations in Asia - Asian J Sports Med. 2011 Mar;2(1):1 -15. [PMID: 22375212]
[9] Abumrad N et al. 1998 J Lipid Res. 39(12): 2309 [PMID 9831619]
[10] Ibrahimi A & Abumrad NA, 2002 Proc Nat Acad Sci USA 93: 2646 [PMID 8610095]
[11] Ibrahimi A., Sfeir Z., maghaais H. 1993 -Expression of the CD36 homolog (FAT) in fibroblast cells: effects on fatty acid tr ansport. Proc Nat Acad Sci USA 93:2646-2651 [PMID 8610095]
[12] Asch AS et al. 1984 J. C lin. Invest 79: 1054 [PMCID: PMC424283]
[13] Hajri TR et al. (2001) J Biol Chem 276: 23661 [PMID 11323420]
[14] Fukuchi K et al. 1999 J Nucl Med 40: 239 [PMID 10025829]
[15] Sabaouni I et al. 2013 Bioinformation 9(17): 849 [PMID 24250110]
[16] Subramanian A et al. 2005 Proc Nat Acad Sci USA 102(43): 15545 [PMID 16199517]
[17] Subramanian A et al. 2007 Bioinformatics 23(23): 3251 [PMID 17644558]
[18] Merico D et al. 2010 PLoS ONE 5(11): e13984 [PMID 21085593]
[19] Zuberi K et al. 2013 Nucleic Acids Research 41(1): W115 [PMID 23794635]
[20] Warde -Farley D et al. 2010″Nucleic Acids Research 38(supplement 2): W214 [PMID 20576703]
[21] Szklarczyk D et al. 2011 Nucleic Acids Research 39(supplement 1) D561 [PMID 21045058]
[22] Franceschini A et al. 2013 Nucleic Acids Rese arch 41(1) D808 [PMID 23203871]
[23] Ma'ayan A. Sci Signal. 2011 4:190 [PMID 21917719]
[24] Randle PJ et al. 1993 (1): 785 [PMID 13990765].
[25] Kuang M et al. 2004 Circulation 109: 1550 [PMID 15023869].
[26] Asch AS et al. 1991 J. Biol. Chem. 266: 1740 [PMID 1703153].
[27] Miao WM et al. 2001 Blood 97(6): 1689. [PMID: 11238109]
[28] Khatri P & Dr aghici S, 2005 Bioinformatics 21: 3587 [PMID 15994189]
[29] Ge H et al. 2001 Nat. Genet. 29: 482 [PMID 11694880]

Table 1: Gene description with corresponding reference

Gene Name	Synonyms / Identifier	Description	R
CD36	ENSG00000135218, ENSP00000308165, ENSP00000378628, ENSP00000392298, ENSP00000396258, ENSP00000398760, ENSP00000399421, ENSP00000401863, ENSP00000407690, ENSP00000409762, ENSP00000410371, ENSP00000411411, ENSP00000415743, ENSP00000416388, ENSP00000431296, ENSP00000433659, ENSP00000435698, ENSP00000439543, ENSP00000441956, 948, CD36, NP_0000 63, NP_001001547, NP_001001548, NP_001120915, NP_001120916, NM_000072, NM_001001547, NM_001001548, NM_001127443, NM_001127444, GP3B, GP4, GPIV, SCARB3, CD36_HUMAN, P16671,	CD36 molecule (thrombospondin receptor) (472 aa) updated on 21 -Jun-2015	[a]
Dgat1	ENSG00000185000, ENSG00000261698, ENSP00000332258, ENSP00000432795, ENSP00000454624, ENSP00000457814, 8694, DGAT1, NP_036211, NM_012079, ARGP1, DGAT, DGAT1_HUMAN, O75907,	diacylglycerol O -acyltransferase 1; Catalyzes the terminal and only committed step in triacylglycerol synthesis by using diacylglycerol and fatty acyl CoA as substrates. In contrast to DGAT2 it is not essential for survival. May be involved in VLDL (very l ow density lipoprotein) assembly. In liver, plays a role in esterifying exogenous fatty acids to glycerol. Functions as the major acyl -CoA retinol acyltransferase (ARAT) in the skin, where it acts to maintain retinoid homeostasis and prevent retinoid toxic ity leading to skin and hair disorders (488 aa)	[b]
Adh4	ENSG00000198099, ENSP00000265512, ENSP00000397939, ENSP00000423571, ENSP00000424583, ENSP00000424630, ENSP00000425416, ENSP00000426667, ENSP00000427261, ENSP00000427525, 127, ADH4, NP_000661, NM_0 00670, ADH -2, ADH4_HUMAN, P08319,	alcohol dehydrogenase 4 (class II), pi polypeptide (380 aa)	[c]
Irs3	ENSMUSP00000060844	insulin receptor substrate 3	[d]
Irs1	ENSG00000169047, ENSP00000304895, 3667, IRS1, NP_005535, NM_005544, HIRS -1, IRS1_HUMAN, P35568,	insulin receptor substrate 1, May mediate the control of various cellular processes by insulin. When phosphorylated by the insulin receptor binds specifically to various cellular proteins containing SH2 domains such as phosphatidylinositol 3 -kinase p85 subunit or GRB2. Activates phosphatidylinositol 3 -kinase when bound to the regulatory p85 subunit (By similarity)	[e]
IL2	ENSG00000109471, ENSP00000226730, 3558, IL2, NP_000577, NM_000586, IL -2, TCGF, IL2_HUMAN, P60568,	interleukin 2, Produced by T -cells in response to antigenic or mitogenic stimulation, this protein is required for T -cell proliferation and other activities crucial to regulation of the immune response. Can stimulate B -cells, monocytes, lymphokine - activated killer cells, natural killer cells, and glioma cells	[f]
Fat4	ENSG00000196159, ENSP00000335169, ENSP00000377862, 79633, FAT4, NP_078858, NM_024582, CDH F14, CDHR11, FAT -J, FAT4_HUMAN, Q6V0I7,	FAT tumor suppressor homolog 4 (Drosophila), May function in the regulation of planar cell polarity. Cadherins are cell -cell interaction molecules (By similarity)	[g]
Pdhb	ENSG00000168291, ENSP00000307241, ENSP00000373220, ENSP00000417267, ENSP00000418448,	pyruvate dehydrogenase (lipoamide) beta, The pyruvate dehydrogenase complex catalyzes the overall conversion of pyruvate to acetyl -CoA and CO(2), and thereby links the glycolytic pathway to the tricarboxylic cycle	[h]
Lpl	ENSG00000175445, ENSP00000309757, ENSP00000428237, ENSP00000428496, ENSP00000428557,	lipoprotein lipase, The primary function of this lipase is the hydrolysis of triglycerides of circulating chylomicr ons and very low density lipoproteins (VLDL). Binding to heparin sulfate proteogylcans at the cell surface is vital to the function. The apolipoprotein, APOC2, acts as a coactivator of LPL activity in the presence of lipids on the luminal surface of vascul ar endothelium (By similarity)	[i]
Ide	ENSG00000175445, ENSP00000309757, ENSP00000428237, ENSP00000428496, ENSP00000428557,	lipoprotein lipase, The primary function of this lipase is the hydrolysis of triglycerides of circulating chylomicrons and very low density lipoproteins (VLDL). Binding to heparin sulfate proteogylcans at the cell surface is vital to the function. The apol ipoprotein, APOC2, acts as a coactivator of LPL activity in the presence of lipids on the luminal surface of vascular endothelium (By similarity)	[j]
Uqcrh	ENSG00000173660, ENSP00000309565, 440567, 7388, UQCRH, NP_001083060, NP_005995, NM_001089591, NM_ 006004, QCR6, UQCR8, P07919, QCR6_HUMAN,	ubiquinol -cytochrome c reductase hinge protein, This is a component of the ubiquinol - cytochrome c reductase complex (complex III or cytochrome b -c1 complex), which is	[k]

		part of the mitochondrial respiratory chain. Th is protein may mediate formation of the complex between cytochromes c and c1 (91 aa)	
Uqcrc1	ENSG00000010256, ENSP00000203407, ENSP00000388660, ENSP00000393696, 7384, UQCRC1, NP_003356, NM_003365, D3S3191, QCR1, UQCR1, P31930, QCR1_HUMAN,	ubiquinol -cytochrome c reductase core protein I, This is a component of the ubiquinol - cytochrome c reductase complex (complex III or cytochrome b -c1 complex), which is part of the mitochondrial respiratory chain. This protein may mediate formation of the complex bet ween cytochromes c and c1 (480 aa)	[l]
Gyk	ENSMUSP00000119564	glycerol kinase, Key enzyme in the regulation of glycerol uptake and metabolism (By similarity) (553 aa)	[m]
Thbs1	ENSG00000137801, ENSP00000260356, ENSP00000380720, 7057, THBS1, NP_003237, NM_003246, THBS, THBS -1, TSP, TSP -1, TSP1, P07996, TSP1_HUMAN,	thrombospondin 1, Adhesive glycoprotein that mediates cell -to-cell and cell -to-matrix interactions. Binds heparin. May play a role in dentinogenesis an d/or maintenance of dentin and dental pulp (By similarity). Ligand for CD36 mediating antiangiogenic properties (1170 aa)	[n]
Sdc4	ENSP00000361818	Cell surface proteoglycan that bears heparan sulfate	[o]
CD9	ENSG00000010278, ENSP00000009180, ENSP00000371955, ENSP00000371958, ENSP00000371959, ENSP00000440985, 928, CD9, NP_001760, NM_001769, BA2, MIC3, MRP -1, TSPAN29, CD9_HUMAN, P21926,	molecule, Involved in platelet activation and aggregation. Regulates parano dal junction formation. Involved in cell adhesion, cell motility and tumor metastasis. Required for sperm -egg fusion	[p]
Pdgfra	ENSG00000134853, ENSP00000257290, ENSP00000424218, ENSP00000425232, ENSP00000425626, ENSP00000425648, ENSP00000425902, ENSP00 000426472, 5156, PDGFRA, NP_006197, NM_006206, CD140a, PDGFR2, P16234, PGFRA_HUMAN,	platelet-derived growth factor receptor, alpha polypeptide, Tyrosine -protein kinase that acts as a cell -surface receptor for PDGFA, PDGFB and PDGFC and plays an essential role in the regulation of embryonic development, cell proliferation, survival and chemotaxis. Depending on the context, promotes or inhibits cell proliferation and cell migration. Plays an important role in the differentiation of bone marrow -derived mesenchymal stem cells. Required for normal skeleton development and cephalic closure during embryonic development. Required for normal development of the mucosa lining th [...] (1089 aa)	[q]
Hand1	ENSG00000113196, ENSP00000231121, 9421, HAND1, NP_004812, NM _004821, bHLHa27, eHand, Hxt, Thing1, HAND1_HUMAN, O96004,	heart and neural crest derivatives expressed 1, Transcription factor that plays an essential role in both trophoblast -giant cells differentiation and in cardiac morphogenesis. In the adult, could be required for ongoing expression of cardiac - specific genes. Binds the DNA sequence 5' - NRTCTG -3' (non -canonical E -box) (By similarity) (215 aa)	[r]
Arnt2	ENSG00000172379, ENSP00000307479, ENSP00000452961, ENSP00000453651, ENSP00000453792, 9915, ARNT2, NP_055677, NM_014862, bHLHe1, KIAA0307, ARNT2_HUMAN, Q9HBZ2,	aryl -hydrocarbon receptor nuclear translocator 2, Specifically recognizes the xenobiotic response element (XRE)	[s]
Rag1	ENSG00000166349, ENSP00000299440, ENSP00000434610, 5896, RAG1, NP_000439, NM_000448, MGC43321, RNF74, P15918, RAG1_HUMAN,	recombination activating gene 1, Catalytic component of the RAG complex, a multiprotein complex that mediates the DNA cleavage phase during V(D)J recombination. V(D)J recombination assembles a diverse repertoire of immunoglobulin and T -cell receptor genes in developing B and T - lymphocytes through rearrangement of different V (variable), in some cases D (diversity), and J (joining) gene segments. In the RAG complex, RAG1 mediates the DNA -binding to the conserved recombination signal sequences (RSS) and catalyzes the DNA cleavage activities by introducing a double -strand break between t [...]	[t]
Api5	ENSG00000166181, ENSP00000368129, ENSP00000399341, ENSP00000402540, ENSP00000431391, ENSP00000434462, ENSP00000436189, ENSP00000436436, 8539, API5, NP_001136402, NP_001136403, NP_006586, NM_001142930, NM_001142931, NM_006595, AAC -11, AAC11, API5L1, API5_HUMAN, Q9BZZ5,	apoptosis inhibitor 5, Antiapoptotic factor that may have a role in protein assembly. Negatively regulates ACIN1. By binding to ACIN1, it suppresses ACIN1 cleavage from CASP3 and ACIN1 -mediated DNA fragmentation. Also known to efficiently suppress E2F1 -induced apoptosis. Its depletion enhances the cytotoxic action of the chemotherapeutic drugs	[u]
Map3K2	ENSG00000166181, ENSP00000368129, ENSP00000399341, ENSP00000402540, ENSP00000431391, ENSP00000434462, ENSP00000436189, ENSP00000436436, 8539, API5, NP_001136402, NP_001136403, NP_006586, NM_001142930, NM_001142931, NM_006595, AAC -11, AAC11, API5L1, API5_HU MAN, Q9BZZ5,	apoptosis inhibitor 5, Antiapoptotic factor that may have a role in protein assembly. Negatively regulates ACIN1. By binding to ACIN1, it suppresses ACIN1 cleavage from CASP3 and ACIN1 -mediated DNA fragmentation. Also known to efficiently su ppress E2F1 -induced apoptosis. Its depletion enhances the cytotoxic action of the chemotherapeutic drugs	[v]
Mertk	ENSG00000153208, ENSP00000295408, ENSP00000387277, ENSP00000389152, ENSP00000402129, ENSP00000412660, 10461, MERTK, NP_006334, NM_006343, mer, RP38, MERTK_HUMAN, Q12866,	c-mer proto -oncogene tyrosine kinase, Receptor tyrosine kinase that transduces signals from the extracellular matrix into the cytoplasm by binding to several ligands including LGALS3, TUB, TULP1 or GAS6. Regulates many phy siological processes including cell survival, migration, differentiation, and phagocytosis of apoptotic cells (efferocytosis). Ligand binding at the cell surface induces autophosphorylation of MERTK on its intracellular domain that provides docking sites for downstream signaling molecules. Following activation by ligand, interacts with GRB2 or PLCG2 and induces phosph [...] (999 aa)	[w]
Myl4	ENSG00000198336, ENSP00000347055, ENSP00000377096, ENSP00000442375, ENSP00000458194, ENSP00000458907, ENSP00000459035, ENSP00000460734, ENSP00000461121, ENSP00000461570, ENSP00000461747, 4635, MYL4, NP_001002841, NP_002467, NM_001002841, NM_002476, AMLC, G T1, PRO1957, MYL4_HUMAN, P12829,	myosin, light chain 4, alkali, atrial, embryonic, Regulatory light chain of myosin. Does not bind calcium (197 aa)	[x]
Tnnt2	ENSG00000118194, ENSP00000236918, ENSP00000353535, ENSP00000356284, ENSP00000356286, ENSP0000 0356287, ENSP00000356289, ENSP00000356291, ENSP00000387874, ENSP00000395163, ENSP00000402238, ENSP00000404134, ENSP00000408731, ENSP00000414036, ENSP00000422031, 7139, TNNT2, NP_000355, NP_001001430, NP_001001431, NP_001001432, NP_001263274, NP_001263275, NP_001263276, NM_000364, NM_001001430, NM_001001431, NM_001001432, NM_001276345, NM_001276346, NM_001276347, CMH2, P45379, TNNT2_HUMAN,	troponin T type 2 (cardiac), Troponin T is the tropomyosin -binding subunit of troponin, the thin filament regulatory co mplex which confers calcium -sensitivity to striated muscle actomyosin ATPase activity (288 aa)	[y]
Pdlim3	ENSP00000284770	PDZ and LIM domain 3,May play a role in the organization of actin filament arrays within muscle respiratory chain. This protein may mediate formation of the complex between cytochromes c and c1 (480 aa)	[z]
Ifng	ENSG00000111537, ENSP00000229135, 3458, IFNG, NP_000610, NM_000619, IFNG_HUMAN, P01579,	interferon, gamma, Produced by lymphocytes activated by specific antigens or mitogens. IFN -gamma, in addition to having antiviral activity, has important immunoregulatory functions. It is a potent activator of macrophages, it has antiproliferative effects on transformed cells and it can potentiate the antiviral and antitumor effects of the type I interferons (By similarity)	[aa]
Pdlim3	ENSP00000284770	PDZ and LIM domain 3,May play a role in the organization of actin filament arrays within muscle cells (By similarity) (364 aa)	[bb]

R = Reference

Table 3: Reference list to Table 1

Reference [R]	URL
[a]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=948

[b]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=8694
[c]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=127
[d]	http://www.uniprot.org/uniprot/O5
[e]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=3667
[f]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search& term=3558
[g]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=79633
[h]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=5162
[i]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=4023
[j]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=4023
[k]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=440567
[l]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=7384
[m]	http://string -db.org/newst ring_cgi/show_network_section.pl
[n]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=7057
[o]	http://www.uniprot.org/uniprot/B4E1S6
[p]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=928
[q]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=5156
[r]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=9421
[s]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=9915
[t]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=5896
[u]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=8539
[v]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=8539
[w]	http://www.uniprot.org/uniprot/B2RE75
[x]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=4635
[y]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=7139
[z]	http://www.uniprot.org/uniprot/D6RAF1
[aa]	http://www.ncbi .nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=3458
[bb]	http://www.uniprot.org/uniprot/D6RAF1

Table 2: Interactions for CD36 (early response) network using GeneMania and STRING

Interactions	GENMANIA	STRING
CD36 -> Dgat1	Co-localization	Co-Mentioned in PubMed Abstracts
CD36 -> Lpl	Co-expression	Co-Expression and
CD36 -> Uqcrc2	Co-expression	Co-Mentioned in PubMed Abstracts
CD36 -> Scab1	Predicted shared protein domain	-
CD36 -> Scap2	Predicted, shared protein domain co -expression	-
CD36 -> Mertk	Predicted	Experimental/Biochemical Data : putative homologs were found interacting in other species ; Co-Mentioned in PubMed Abstracts
CD36 -> CD9	Predicted	Experimental/Biochemical Data : putative homologs were found interacting in other species Co-Mentioned in PubMed Abstracts Association in Curated Databases
CD36 -> Hand1	Predicted	-
CD36 -> Rag1	Predicted	-
CD36 -> Map3K2	Predicted	-
CD36 -> Il2	Predicted	-
CD36 -> Ifg	Predicted	-
CD36 -> Thbs1	Predicted	Experimental/Biochemical Data : putative homologs were found interacting in other species Co-Mentioned in PubMed Abstracts Association in Curated Databases
CD36 -> Sdc4	Predicted	Experimental/Biochemical Data : putative homologs were found interacting in other species Co-Mentioned in PubMed Abstracts Association in Curated Databases
CD36 -> Irs1	Predicted	-
CD36 -> Pdagfra	Predicted and Co -expression	-
CD36 -> TnnT3	Co-expression	
CD36 -> TnnT1	Co-expression	
CD36 -> Sdc1	Predicted	Experimental/Biochemical Data : putative homologs were found interacting in other species Co-Mentioned in PubMed Abstracts Association in Curated Databases
CD36 -> Sdc3	Predicted	Experimental/Biochemical Data : putative homologs were found interacting in other species Co-Mentioned in PubMed Abstracts Association in Curated Databases
TNNT2 ->MYL4	Co-expression	Co-Expression Association in Curated Databases; Co -Mentioned in PubMed Abstracts
UQCRC1 ->UQCRH	Co-expression	-
CD9 ->UQCRH	Co-expression	-
PDHB ->UQCRH	Co-expression	-
UQCRC1 ->PDHB	Co-expression	-
HAND1 ->TNNT2	Co-expression	-
PDGFRA ->THBS1	Co-expression	-
LPL ->THBS1	Co-expression	-
IRS1 ->PDGFRA	Co-expression	-
LPL ->MYL4	Co-expression	-
SDC4 ->MERTK	Co-expression	-
SDC4 ->PDGFRA	Co-expression	-
SDC4 ->IFNG	Co-expression	-
HAND1 ->MYL4	Co-expression	-
HAND1 ->PDGFRA	Co-expression	-
HAND1 ->THBS1	Co-expression	-
HAND1 ->IRS1	Co-expression	-
HAND1 ->LPL	Genetic interactions	-
UQCRC1 ->DGAT1	Genetic interactions	-
HAND1 ->ADH4	Genetic interactions	-
HAND1 ->IL2	Genetic interactions	-
LPL ->FAT4	Genetic interactions	-
PDGFRA ->FAT4	Genetic interactions	-
LPL ->TNNT2	Genetic interactions	-
MAP3K2 ->IDE	Genetic interactions	-
LPL ->MYL4	Genetic interactions	-
TNNT2 ->MYL4	Genetic interactions	-
UQCRC1 ->RAG1	Genetic interactions	-
IFNG ->THBS1	Genetic interactions	-
SDC4 ->API5	Genetic interactions	-
UQCRC1 ->API5	Genetic interactions	-

LPL ->API5	Genetic interactions	-
MERTK ->RAG1	Genetic interactions	-
IFNG ->UQCRC1	Genetic interactions	-
MYL4 ->FAT4	Genetic interactions	-
UQCRH ->API5	Genetic interactions	-
IFNG ->UQCRH	Genetic interactions	-
IRS1 ->API5	Genetic interactions	-
PDGFRA ->MAP3K2	Genetic interactions	-
MAP3K2 ->FAT4	Genetic interactions	-
MERTK ->PDLIM3	Genetic interactions	-
CD9 ->MERTK	Genetic interactions	-
PDGFRA ->IDE	Genetic interactions	-
FAT4 ->API5	Genetic interactions	-
IRS1 ->FAT4	Genetic interactions	-
PDGFRA ->PDHB	Genetic interactions	-
PDHB ->FAT4	Genetic interactions	-
SDC4 ->THBS1	Pathway	-
TNNT2 ->MYL4	Pathway	-
IFNG ->PDGFRA	Pathway	-
IFNG ->IL2	Pathway	-
IRS1 ->IL2	Pathway	-
UQCRC1 ->IDE	Shared protein domains	-
PDGFRA ->MERTK	Shared protein domains	-
IFNG ->IL2	Shared protein domains	-
THBS1 ->FAT4	Shared protein domains	-

Edited by P Kangueane
Citation : Sabaouni et al. Bioinformation 12(6): 332-339 (2016)

Conclusion générale

Le travail présenté dans cette étude correspond à l'acquisition au sein du laboratoire MedBiotech d'une compétence en analyse du transcriptome. Cette compétence port sur les aspects méthodologique (outils bioinformatiques pour la conception et l'analyse des données des puces à ADN).

Les résultats déjà obtenus démontrent d'une part la faisabilité de ce type d'analyse sur les modèles de souris (Wt, Ko, Gr), et d'autre part donnent une première vision globale du système de fonctionnement de l'expression du gène CD36 dans les cellules myocardiaque chez les modèles de souris suscités.

L'étude réalisée finit sur une liste des résultats issus de nos modèles de souris, suivie d'une question très importante : est ce que vraiment l'absence de l'expression de CD36 dans les cellules cardiaque provoque les dysfonctionnements cardiaques observés chez les souris Ko. A cet effet, l'ensemble de ces travaux a permis de valider la pertinence de la représentation des données interviennent dans les mécanismes moléculaire de la maladie myocardiopathie hypertrophique observé chez le modèles des souris KO (absence d'expression de CD36) tel que la balance glucose/acide gras en tenant compte que le protéine CD36 est connue comme récepteur/ transporteur des Acides gras , apoptose/angiogenèse et ce par le biais de la trampospodine-1 et le protéine CD36 est un récepteur de TSP-1 qui active toutes une cascade de l'angiogenèse. Et cela faite par l'intégration des données issue de la technologie des puces à ADN et cette intégration passe par une réflexion sur la façon de représenter tous ces données très hétérogènes afin qu'on puisse les combinés.

La compréhension des mécanismes moléculaires permettant la propagation et l'intégration des signaux émis par la protéine CD36 dans les cellules cardiaques est un objectif majeur des études en biologie. Si d'énormes progrès ont été accomplis sur la compréhension globale de réseau signalétique, la façon dont ils intègrent les flux d'information et leur capacité à induire une réponse cellulaire. Dans ce contexte la compréhension de réseaux de signalisation de CD36 est un enjeu majeur pour aider à détecter et soigner la myocardiopathie et d'autres maladies lie au CD36.

BIBLIOGRAPHIE

Abdelkarim, M, Caron, S, Duhem, C, Prawitt, J, Dumont, J, Lucas, A, Bouchaert, E, Briand, O, Brozek, J, Kuipers, F, Fievet, C, Cariou, B & Staels, B. (2010). The farnesoid X receptor regulates adipocyte differentiation and function by promoting peroxisome proliferator-activated receptor-gamma and interfering with the Wnt/beta-catenin pathways. *J Biol Chem*, 285, 36759-36767.

Abumrad, NA, el-Maghrabi, MR, Amri, EZ, Lopez, E & Grimaldi, PA. (1993). Cloning of a rat adipocyte membrane protein implicated in binding or transport of long-chain fatty acids that is induced during preadipocyte differentiation. Homology with human CD36. *J Biol Chem*, 268, 17665-17668.

Agapito,G., Guzzi,P.H. et Cannataro,M. (2013) Visualization of protein interaction networks : problems and solutions. BMC Bioinformatics, 14 Suppl 1, S1.

Ahrendt, S.A., et al., (1999). *Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array.* Proc Natl Acad Sci U S A, 96(13): p. 7382-7.

Alberts,B. (1998) The cell as a collection of protein machines : preparing the next generation of molecular biologists. Cell, 92 (3), 291–294.

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

Allison, D.B., et al., (2006). *Microarray data analysis: from disarray to consolidation and consensus.* Nat Rev Genet, 7(1): p. 55-65.

Alter, O., P.O. Brown, and D. Botstein, (2000). *Singular value decomposition for genome-wide expression data processing and modeling.* Proc Natl Acad Sci U S A, 97(18): p. 10101-

6.

Allred, KF, Smart, EJ & Wilson, ME. (2006). Estrogen receptor-alpha mediates gender differences in atherosclerosis induced by HIV protease inhibitors. *J Biol Chem*, 281, 1419-1425.

Al-Shahrour, F., Minguez, P., Tárraga, J., Medina, I., Alloza, E., Montaner, D. & Dopazo, J. (2007) FatiGO + : a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic acids research, 35 (Web Server issue), W91–6.

Alwine, J.C., D.J. (1997). Kemp, and G.R. Stark, *Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.* Proc Natl Acad Sci U S A, 74(12): p. 5350-4.

Andrea Sackmann, Monika Heiner, et Ina Koch, 2006. Application of Petri net based analysis techniques to signal transduction pathways. BMC Bioinformatics, 7 :482.

Anwar, K, Voloshyna, I, Littlefield, MJ, Carsons, SE, Wirkowski, PA, Jaber, NL, Sohn, A, Eapen, S & Reiss, AB. (2011). COX-2 inhibition and inhibition of cytosolic phospholipase A2 increase CD36 expression and foam cell formation in THP-1 cells. *Lipids*, 46, 131-142.

Asch, AS, Barnwell, J, Silverstein, RL & Nachman, RL. (1987). Isolation of the thrombospondin membrane receptor. *J Clin Invest*, 79, 1054-1061.

Asch As. Nachman RL: thrombospondin phenomenology to furiction Prog. Hemost Thromb.9: 157-178, 1989. membranr receptor J. clin. Invest 79:1054-1061, 1984.

Ashburner M., A. Ball C., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K.,. Dwight S. S, Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M. et Sherlock G. (2000): Gene ontology : tool for the unification of biology. the gene ontology consortium. Nat Genet, 25(1) :25–9.
Ashburner,M.et al.(2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.Nat. Genet., 25, 25–29.

Bamberger, ME, Harris, ME, McDonald, DR, Husemann, J & Landreth, GE. (2003). A cell surface receptor complex for fibrillar beta-amyloid mediates microglial activation. *J Neurosci*, 23, 2665-2674.

Barrett, J.C. and E.S. Kawasaki, (2003). *Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression.* Drug Discov Today, 8(3): p. 134-41.

Beardmore, J.A., (1997) Transgenics, autotransgenics, and allotransgenics, Transgenics Research, 6, 107-108

Beer, D.G., et al., (2002). *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nat Med, 8(8): p. 816-24.

Ben Mepham, T., Combes, R.D., Balls, M., Barbieri, O., Blokhuis, H.J., Costa, P., Crilly, R.E., De Cock Buning, T., Delpire, V.C., O'hare, M.J., Houdebine, L.M., Van Kreijl, C.F., Van Der Meer, M., Reinhardt, C.A., Wolfe, E., Van Zeller, A-M., (1998) The use of transgenic animals in the European Union : The Report and Recommendations of ECVAM Workshop 28, ATLA, 26, 21-43

Benjamini, Y., et al., (2001). *Controlling the false discovery rate in behavior genetics research.* Behav Brain Res, 125(1-2): p. 279-84.

Bidasee KR, Dincer ÜD, Besch HR (2001) Ryanodine receptor dysfunction in hearts of streptozotocin-induced diabetic rats. Mol Pharmacol 60(6): 1356-1364.

Bishop, J.O., (1997) Chromosomal Insertion of Foreign DNA, In : HOUDEBINE, L.M., Transgenic Animals. Generation and Use, Amsterdam : Harwood Academic Publishers, 219-224

Bittner, M., et al., (2000). *Molecular classification of cutaneous malignant melanoma by gene expression profiling.* Nature, 406(6795): p. 536-40.

Blanchard, A.P., R.J. Kaiser, and L.E. Hood, (1996). *High-density oligonucleotide arrays.* Biosens. Bioelectron., 6/7: p. 687-690.

Bodart, V, Bouchard, JF, McNicoll, N, Escher, E, Carriere, P, Ghigo, E, Sejlitz, T, Sirois, MG, Lamontagne, D & Ong, H. (1999). Identification and characterization of a new growth hormone-releasing peptide receptor in the heart. Circ Res, 85, 796-802.

Bodart, V, Febbraio, M, Demers, A, McNicoll, N, Pohankova, P, Perreault, A, Sejlitz, T, Escher, E, Silverstein, RL, Lamontagne, D & Ong, H. (2002). CD36 mediates the cardiovascular action of growth hormone-releasing peptides in the heart. *Circ Res*, 90, 844-849.

Boden G, Shulman GI.(2002) Free fatty acids in obesity and type 2 diabetes: defining their role in the development of insulin resistance and beta-cell dysfunction. Eur J Clin Invest;32(3):14–23.

Bohen, S.P., et al., (2003). *Variation in gene expression patterns in follicular lymphoma and the response to rituximab.* Proc Natl Acad Sci U S A, 100(4): p. 1926-30.

Bolin, RB, Medina, F & Cheney, BA. (1981). Glycoprotein changes in fresh vs. room temperature-stored platelets and their buoyant density cohorts. *J Lab Clin Med*, 98, 500-510

Bonen, A, Han, XX, Habets, DD, Febbraio, M, Glatz, JF & Luiken, JJ. (2007). A null mutation in skeletal muscle FAT/CD36 reveals its essential role in insulin- and AICAR-stimulated fatty acid metabolism. *Am J Physiol Endocrinol Metab*, 292, E1740-E1749.

Bonen, A, Parolin, ML, Steinberg, GR, Calles-Escandon, J, Tandon, NN, Glatz, JF, Luiken, JJ, Heigenhauser, GJ & Dyck, DJ. (2004). Triacylglycerol accumulation in human obesity and type 2 diabetes is associated with increased rates of skeletal muscle fatty acid transport and increased sarcolemmal FAT/CD36. *FASEB J*, 18, 1144-1146.

Bowers, CY, Momany, FA, Reynolds, GA & Hong, A. (1984). On the in vitro and in vivo activity of a new synthetic hexapeptide that acts on the pituitary to specifically release growth hormone. *Endocrinology*, 114, 1537-1545.

Bowers, CY. (1998). Growth hormone-releasing peptide (GHRP). *Cell Mol Life Sci*, 54, 1316-1329.

Brinkmann JFF, Abumrad NA, Ibrahimi A, van der Vusse GJ, Glatz JFC. (2002) New insights into long-chain fatty acid uptake by heart muscle: a crucial role for fatty acid translocase/CD36. Biochem J.;367:561–570.

Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E., Palmiter, R.D., (1988) Introns increase transcriptionnal efficiency in transgenic mice, Proc. Natl. Acad. Sci. USA, 85, 836-840

Brinster, R.L., Chen, H.Y., Trumbauer, M.E., Yagle, M.K., (1985) Factors affecting the efficiency of introducing foreign DNA into mice by micro-injecting eggs, Proc. Natl. Acad. Sci. USA,, 82, 4438-4442

Brown, AJ, Dean, RT & Jessup, W. (1996). Free and esterified oxysterol: formation during copper-oxidation of low density lipoprotein and uptake by macrophages. *J Lipid Res*, 37, 320-335.

Brown, P.O. and D. Botstein, (1999). *Exploring the new world of the genome with DNA microarrays.* Nat Genet, 21(1 Suppl): p. 33-7.

Brown, AJ, Mander, EL, Gelissen, IC, Kritharides, L, Dean, RT & Jessup, W. (2000). Cholesterol and oxysterol metabolism and subcellular distribution in macrophage foam cells: accumulation of oxidized esters in lysosomes. *J Lipid Res*, 41, 226-236.

Bruni, F, Pasqui, AL, Pastorelli, M, Bova, G, Cercignani, M, Palazzuoli, A, Sawamura, T, Gioffre, WR, Auteri, A & Puccetti, L. (2005). Different effect of statins on platelet oxidized-LDL receptor (CD36 and LOX-1) expression in hypercholesterolemic subjects. *Clin Appl Thromb Hemost*, 11, 417-428.

Calvo, D, Gomez-Coronado, D, Suarez, Y, Lasuncion, MA & Vega, MA. (1998). Human CD36 is a high affinity receptor for the native lipoproteins HDL, LDL, and VLDL. *Lipid Res*, 39, 777-788.

Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. (2008): Drug target identification using side-effect similarity. Science, 321(5886):263–266.

Campbell, SE, Tandon, NN, Woldegiorgis, G, Luiken, JJ, Glatz, JF & Bonen, A. (2004). A novel function for fatty acid translocase (FAT)/CD36: involvement in long chain fatty acid transfer into the mitochondria. *J Biol Chem*, 279, 36235-36241.

Carvalho, MD, Vendrame, CM, Ketelhuth, DF, Yamashiro-Kanashiro, EH, Goto, H & Gidlund, M. (2010). High-density lipoprotein inhibits the uptake of modified low- density lipoprotein and the expression of CD36 and FcgammaRI. *J Atheroscler Thromb*, 17, 844-857.

Castle, J., et al., (2003). *Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing.* Genome Biol, 4(10): p. R66.

Catimel B, McGregor JL, Hasler T, Greenwalt DE, Howard RJ, Leung LL. Epithelial membrane glycoprotein PAS-IV is related to platelet glycoprotein IIIb binding to thrombospondin but not to malaria-infected erythrocytes. *Blood.* 1991; 77(12): 2649-54.

Cawley, S., et al., (2004). *Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.* Cell, 116(4): p. 499-509.

Chabowski, A, Coort, SL, Calles-Escandon, J, Tandon, NN, Glatz, JF, Luiken, JJ & Bonen, A. (2004). Insulin stimulates fatty acid transport by regulating expression of FAT/CD36 but not FABPpm. *Am J Physiol Endocrinol Metab*, 287, E781-E789.

Chada, K. Magram, J., Raphael, K., Radice, G., Lacy, E., COSTANTINI, F., (1985) Specific expression of a foreign beta-globin gene in erythroid cells of transgenic mice, Nature, 314, 377-380

Chen, M, Yang, Y, Braunstein, E, Georgeson, KE & Harmon, CM. (2001a). Gut expression and regulation of FAT/CD36: possible role in fatty acid transport in rat enterocytes. *Am J Physiol Endocrinol Metab*, 281, E916-E923.

Chen, Z, Ishibashi, S, Perrey, S, Osuga, J, Gotoda, T, Kitamine, T, Tamura, Y, Okasaki, H, Yahagi, N, Iizuka, Y, Shionoiri, F, Ohashi, K, Harada, K, Shimano, H, Nagai, R & Yamada, N. (2001b). Troglitazone inhibits atherosclerosis in apolipoprotein E-knockout mice: pleiotropic effects on CD36 expression and HDL. *Arterioscler Thromb Vasc Biol*, 21, 372-377.

Chen, G, Liang, G, Ou, J, Goldstein, JL & Brown, MS. (2004). Central role for liver X receptor in insulin-mediated activation of Srebp-1c transcription and stimulation of fatty acid synthesis in liver. *Proc Natl Acad Sci U S A*, 101, 11245-11250.

Chen, H. & Sharp, B. M. (2004) Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics, 5, 147.

Chen, M, Yang, YK, Loux, TJ, Georgeson, KE & Harmon, CM. (2006). The role of hyperglycemia in FAT/CD36 expression and function. *Pediatr Surg Int*, 22, 647-654.

Cheng,F., Liu,C., Jiang,J., Lu,W., Li,W., Liu,G., Zhou,W., Huang,J. et Tang,Y. (2012) Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol, 8 (5), e1002503.

Chuang, PC, Lin, YJ, Wu, MH, Wing, LY, Shoji, Y & Tsai, SJ. (2010). Inhibition of CD36-dependent phagocytosis by prostaglandin E2 contributes to the development of endometriosis. *Am J Pathol*, 176, 850-860.

Chinetti, G, Gbaguidi, FG, Griglio, S, Mallat, Z, Antonucci, M, Poulain, P, Chapman, J, Fruchart, JC, Tedgui, A, Najib-Fruchart, J & Staels, B. (2000). CLA-1/SR-BI is expressed in atherosclerotic lesion macrophages and regulated by activators of peroxisome proliferator-activated receptors. *Circulation*, 101, 2411-2417.

Chinetti, G, Lestavel, S, Bocher, V, Remaley, AT, Neve, B, Torra, IP, Teissier, E, Minnich, A, Jaye, M, Duverger, N, Brewer, HB, Fruchart, JC, Clavey, V & Staels, B. (2001). PPAR-alpha and PPAR-gamma activators induce cholesterol removal from human macrophage foam cells through stimulation of the ABCA1 pathway. *Nature Med*, 7, 53-58

Chiurchiu, V, Izzi, V, D'Aquilio, F, Vismara, D, Carotenuto, F, Catanzaro, G & Maccarrone, M. (2011). Endomorphin-1 prevents lipid accumulation via CD36 down-regulation and modulates cytokines release from human lipid-laden macrophages. *Peptides*, 32, 80-85.

Collot-Teixeira, S, Martin, J, McDermott-Roe, C, Poston, R & McGregor, JL. (2007). CD36 and macrophages in atherosclerosis. *Cardiovasc Res*, 75, 468-477.

Ciofani,M., Madar,A., Galan,C., Sellars,M., Mace,K., Pauli,F., Agarwal,A., Huang,W., Parkurst,C.N.,Muratet,M., Newberry,K.M., Meadows,S., Greenfield,A., Yang,Y., Jain,P., Kirigin,F.K., Birchmeier,C., Wagner,E.F., Murphy,K.M., Myers,R.M., Bonneau,R. et Littman,D.R. (2012) A validated regulatory network for Th17 cell specification. Cell, 151 (2), 289–303.

Coort, SL, Luiken, JJ, Van Der Vusse, GJ, Bonen, A & Glatz, JF. (2004). Increased FAT (fatty acid translocase)/CD36-mediated long-chain fatty acid uptake in cardiac myocytes from obese Zucker rats. *Biochem Soc Trans*, 32, 83-85.

Clotilde Claudel-Renard, Claude Chevalet, Thomas Faraut, et Daniel Kahn, 2003. Enzyme-specific profiles for genome annotation : PRIAM. Nucleic Acids Res, 31(22) :6633–6639.

Chiappetta, P., M.C. Roubaud, and B. Torresani, (2004). *Blind source separation and the analysis of microarray data.* J Comput Biol, 11(6): p. 1090-109.

Chen, J.J., et al., (2004). *Analysis of variance components in gene expression data.* Bioinformatics, 20(9): p. 1436-46.

Chudin, E., et al., (2002). *Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays.* Genome Biol, 3(1): p. RESEARCH0005.

Cho, R.J., et al., (2001). *Transcriptional regulation and function during the human cell cycle.* Nat Genet, 27(1): p. 48-54.

Cho, R.J., et al., (1998). *A genome-wide transcriptional analysis of the mitotic cell cycle.* Mol Cell, 2(1): p. 65-73.

Chu, S., et al., (1998). *The transcriptional program of sporulation in budding yeast.* Science,

282(5389): p. 699-705.

Churchill, G.A., (2002). *Fundamentals of experimental design for cDNA microarrays.* Nat Genet, 32 Suppl: p. 490-5.

Clark, E.A., et al., (2000). *Genomic analysis of metastasis reveals an essential role for RhoC.* Nature, 406(6795): p. 532-5.

Davignon, L., et al., (2005). *Use of resequencing oligonucleotide microarrays for identification of Streptococcus pyogenes and associated antibiotic resistance determinants.* J Clin Microbiol, 43(11): p. 5690-5.

David L. Nelson et Michael M. Cox. April 2004 Lehninger Prin-ciples of Biochemistry, Fourth Edition. W. H. Freeman,.

Davis, S.J. and A.J. Millar, (2001). *Watching the hands of the Arabidopsis biological clock.* Genome Biol, 2(3): p. REVIEWS1008.

Dawson, DW, Pearce, SF, Zhong, R, Silverstein, RL, Frazier, WA & Bouck, NP. (1997). CD36 mediates the In vitro inhibitory effects of thrombospondin-1 on endothelial cells. *J Cell Biol*, 138, 707-717.

Demers, A, McNicoll, N, Febbraio, M, Servant, M, Marleau, S, Silverstein, R & Ong, H. (2004). Identification of the growth hormone-releasing peptide binding site in CD36: a photoaffinity cross-linking study . *Biochem J*, 382, 417-424.

Demir,E., Cary,M.P., Paley,S., Fukuda,K., Lemer,C., Vastrik,I., Wu,G., D'Eustachio,P., Schaefer,C., Luciano,J., Schacherer,F., Martinez-Flores,I., Hu,Z., Jimenez-Jacinto,V., Joshi-Tope,G., Kandasamy,K., Lopez-Fuentes,A.C., Mi,H., Pichler,E., Rodchenkov,I., Splendiani,A., Tkachev,S., Zucker,J., Gopinath,G., Rajasimha,H., Ramakrishnan,R., Shah,I., Syed,M., Anwar,N., Babur,O., Blinov,M., Brauner,E., Corwin,D., Donaldson,S., Gibbons,F., Goldberg,R., Hornbeck,P., Luna,A., Murray-Rust,P., Neumann,E., Ruebenacker,O.,

Reubenacker,O., Samwald,M., van Iersel,M., Wimalaratne,S., Allen,K., Braun,B., Whirl-Carrillo,M., Cheung,K.H., Dahlquist,K., inney,A., Gillespie,M., Glass,E., Gong,L., Haw,R., Honig,M., Hubaut,O., Kane,D., Krupa,S., Kutmon,M., Leonard,J., Marks,D., Merberg,D., Petri,V., Pico,A., Ravenscroft,D., Ren,L., Shah,N., Sunshine,M., Tang,R., Whaley,R., Letovksy,S., Buetow,K.H., Rzhetsky,A., Schachter,V., Sobral,B.S., Dogrusoz,U., McWeeney,S., Aladjem,M., Birney,E., Collado-Vides,J., Goto,S., Hucka,M., Le Novere,N., Maltsev,N., Pandey,A., Thomas,P., Wingender,E., Karp,P.D., Sander,C. et Bader,G.D. (2010) The BioPAX community standard for pathway data sharing. Nat. Biotechnol., 28 (9), 935–942.

Depré C, Rider MH, Hue L. (1998); Mechanisms of control of heart glycolysis. Eur J Biochem FEBS.;258:277–290.

Depre C, Vanoverschelde JL, Taegtmeyer H. (1999); Glucose for the heart. Circulation.;99:578–588.

DeRisi, J.L., V.R. Iyer, and P.O. Brown, (1997). *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 278(5338): p. 680-6.

DeRisi, J., et al., (1996). *Use of a cDNA microarray to analyse gene expression patterns in human cancer.* Nat Genet, 14(4): p. 457-60.

Devaraj, S, Hugou, I & Jialal, I. (2001). α-Tocopherol decreases CD36 expression in human monocyte-derived macrophages. *J Lipid Res*, 42, 521-527.

Dressman, J, Kincer, J, Matveev, SV, Guo, L, Greenberg, RN, Guerin, T, Meade, D, Li, XA, Zhu, W, Uittenbogaard, A, Wilson, ME & Smart, EJ. (2003). HIV protease inhibitors promote atherosclerotic lesion formation independent of dyslipidemia by increasing CD36-dependent cholesteryl ester accumulation in macrophages. *J Clin Invest*, 111, 389-397.

Drover, VA, Ajmal, M, Nassir, F, Davidson, NO, Nauli, AM, Sahoo, D, Tso, P & Abumrad, NA. (2005). CD36 deficiency impairs intestinal lipid secretion and clearance of chylomicrons from the blood. *J Clin Invest*, 115, 1290-1297.

Drover, VA, Nguyen, DV, Bastie, CC, Darlington, YF, Abumrad, NA, Pessin, JE, London, E, Sahoo, D & Phillips, MC. (2008). CD36 mediates both cellular uptake of very long chain fatty acids and their intestinal absorption in mice. *J Biol Chem*, 283, 13108-13115.

Eaton S. (2002) Control of mitochondrial beta-oxidation flux. Prog Lipid Res.;41:197–239.

Ehehalt, R, Fullekrug, J, Pohl, J, Ring, A, Herrmann, T & Stremmel, W. (2006). Translocation of long chain fatty acids across the plasma membrane--lipid rafts and fatty acid transport proteins. *Mol Cell Biochem*, 284, 135-140.

Eisen, M.B. , Spellman, P.T., Brown, P.O. and Botstein, D. (1998): Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 95(25), 14863-14868.

Eisen, M.B. and P.O. Brown, (1999). *DNA arrays for analysis of gene expression.* Methods Enzymol, 303: p. 179-205.

Eisen, M.B., et al., (1998). *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 95(25): p. 14863-8.

Ekins, R., F. Chu, and J. Micallef,(1989) *High specific activity chemiluminescent and fluorescent markers: their potential application to high sensitivity and 'multi-analyte' immunoassays.* J Biolumin Chemilumin, 4(1): p. 59-78.

Ekins, R.P., (1989). *Multi-analyte immunoassay.* J Pharm Biomed Anal, 7(2): p. 155-68.

Ekins, R.P., F. Chu, and E. Biggart, (1990). *Multispot, multianalyte, immunoassay.* Ann Biol Clin (Paris), 48(9): p. 655-66.

Encyclopedia Universalis (1991) Ontologie, In Encyclopedia Universalis. 16:902-910. Universalis, Paris.

Endemann, G, Stanton, LW, Madden, KS, Bryant, CM, White, RT & Protter, AA. (1993). CD36 is a receptor for oxidized low density lipoprotein. *J Biol Chem*, 268, 11811-11816.

Eulgem, T., (2005). *Regulation of the Arabidopsis defense transcriptome.* Trends Plant Sci,

10(2): p. 71-8.

Fang ZY, Prins JB, Marwick TH (2004) Diabetic cardiomyopathy: evidence, mechanisms, and therapeutic implications. Endocr Rev 25(4):543-67.

Febbraio, M, Abumrad, NA, Hajjar, DP, Sharma, K, Cheng, W, Pearce, SF & Silverstein, RL. (1999). A null mutation in murine CD36 reveals an important role in fatty acid and lipoprotein metabolism. *J Biol Chem*, 274, 19055-19062.

Febbraio, M, Podrez, EA, Smith, JD, Hajjar, DP, Hazen, SL, Hoff, HF, Sharma, K & Silverstein, RL. (2000). Targeted disruption of the class B scavenger receptor CD36 protects against atherosclerotic lesion development in mice. *J Clin Invest*, 105, 1049-1056.

Febbraio, M, Hajjar, DP & Silverstein, RL. (2001). CD36: a class B scavenger receptor involved in angiogenesis, atherosclerosis, inflammation, and lipid metabolism. *J Clin Invest*, 108, 785-791.

Febbraio, M & Silverstein, RL. (2007). CD36: implications in cardiovascular disease. *Int J Biochem Cell Biol*, 39, 2012-2030.

Feng, J, Han, J, Pearce, SFA, Silverstein, RL, Gotto, AM, Jr., Hajjar, DP & Nicholson, AC. (2000). Induction of CD36 expression by oxidized LDL and IL-4 by a common signaling pathway dependent on protein kinase C and PPAR-γ. *J Lipid Res*, 41, 688-696.

Fodor, S.P., et al., (1991). *Light-directed, spatially addressable parallel chemical synthesis.* Science, 1991. 251(4995): p. 767-73.

Force T, Bonventre JV. (1998) Growth factors and mitogen-activated protein kinases. Hypertension; 31, 152-161.

Force T, Alessandrini A, Bonventre, JV.(1999) Cell signaling. In The Kidney : Physiology and Pathophysiology. D. W. Seldin, and G. Giebisch, editors. Lippincott-Raven, Philadelphia,

Fuhrman, B, Koren, L, Volkova, N, Keidar, S, Hayek, T & Aviram, M. (2002a). Atorvastatin therapy in hypercholesterolemic patients suppresses cellular uptake of oxidized-LDL by differentiating monocytes. *Atherosclerosis*, 164, 179-185.

Fuhrman, B, Volkova, N & Aviram, M. (2002b). Oxidative stress increases the expression of the CD36 scavenger receptor and the cellular uptake of oxdized low-density lipoprotein in macrophages from atherosclerotic mice: protective role of antioxidants and of paraoxonase. *Atherosclerosis*, 161, 307-316.

Gagne, M., Pothier, F., Sirard, M.A., Gene Micro-injection into Bovine Pronuclei, In : HOUDEBINE, L.M., (1997) Transgenic Animals. Generation and Use, Amsterdam : Harwood Academic Publishers, 27-36.

Gao, D, Ashraf, MZ, Kar, NS, Lin, D, Sayre, LM & Podrez, EA. (2010). Structural basis for the recognition of oxidized phospholipids in oxidized low density lipoproteins by class B scavenger receptors CD36 and SR-BI. *J Biol Chem*, 285, 4447-4454.

Garber, M.E., et al., (2001). *Diversity of gene expression in adenocarcinoma of the lung.* Proc Natl Acad Sci U S A, 98(24): p. 13784-9.

Gasch, A.P., et al., (2000). *Genomic expression programs in the response of yeast cells to environmental changes.* Mol Biol Cell, 11(12): p. 4241-57.

Gieseg, SP, Amit, Z, Yang, YT, Shchepetkina, A & Katouah, H. (2010). Oxidant production, oxLDL uptake, and CD36 levels in human monocyte-derived macrophages are downregulated by the macrophage-generated antioxidant 7,8-dihydroneopterin. *Antioxid Redox Signal*, 13, 1525-1534.

Glatz JFC, Luiken JJFP, Bonen A. (2010); Membrane fatty acid transporters as regulators of lipid metabolism:implications for metabolic disease. Physiol Rev.;90:367–417.

Glatz JFC, Luiken JJFP, Bonen A. (2010) Membrane fatty acid transporters as regulators of lipid metabolism:implications for metabolic disease. Physiol Rev.;90:367–417.

Golfman LS, Takeda N and Dhalla NS (1996) Cardiac membrane Ca^{2+}-transport in alloxan-induced diabetes in rats. Diabetes Res Clin Pract 31 suppl.: S73-77.

Golfman L, Dixon IMC, Takeda N, Chapman D, Dhalla NS (1999) Differential changes in cardiac myofibrillar and sarcoplasmic reticular gene expression in alloxaninduced diabetes. Mol Cell Biochem 200: 15-25.

Golub, T.R., et al., (1999). *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 286(5439): p. 531-7.

Good, B. M., Kawas, E. A., Kuo, B. Y.-L. & Wilkinson, M. D. (2006) iHOPerator : user-scripting a personalized bioinformatics Web, starting with the iHOP website. BMC bioinformatics, 7, 534.

Gordon, J.W., Ruddle, F.H., (1981) Integration and stable germ lime transmission of genes injected into mouse pronuclei, Science, , 214, 1244-1246

Gordon, J.W., Scangos G.A., Plotkin, D.J., Barbosa, J.A., (1980) Ruddle, F.H., Genetic transformation of mouse embryos by micro-injection of purified DNA, PNAS USA, 77, 7380-84

Gordon, J.W., (1997) Transgenic Technology and Laboratory Animal Science, ILAR Journal, 38 (1), 32-41

Goudriaan, JR, den Boer, MA, Rensen, PC, Febbraio, M, Kuipers, F, Romijn, JA, Havekes, LM & Voshol, PJ. (2005). CD36 deficiency in mice impairs lipoprotein lipase-mediated triglyceride clearance. *J Lipid Res*, 46, 2175-2181.

Goudriaan, JR, Dahlmans, VE, Teusink, B, Ouwens, DM, Febbraio, M, Maassen, JA, Romijn, JA, Havekes, LM & Voshol, PJ. (2003). CD36 deficiency increases insulin sensitivity in muscle, but induces insulin resistance in the liver in mice. *J Lipid Res*, 44, 2270-2277.

Greenwalt DE, Watt KW, So OY, Jiwani N. PAS IV, an integral membrane protein of mammary epithelial cells, is related to platelet and endothelial cell CD36 (GP IV). Biochemistry. 1990; 29(30): 7054-9.

Greenwalt, DE, Lipsky, RH, Ockenhouse, CF, Ikeda, H, Tandon, NN & Jamieson, GA. (1992). Membrane glycoprotein CD36: a review of its roles in adherence, signal transduction, and transfusion medicine. J Am Soc Hematology, 80, 1105-1115.

Griffin, E, Re, A, Hamel, N, Fu, C, Bush, H, McCaffrey, T & Asch, AS. (2001). A link between diabetes and atherosclerosis: glucose regulates expression of CD36 at the level of translation. *Nature Med*, 7, 840-846.

Gruarin, P, Throne, RF, Dorahy, DJ, Burns, GF, Sitia, R & Alessio, M. (2000). CD36 is a ditopic glycoprotein with the N-terminal domain implicated in intracellular transport. *Biochem Biophys Res Commun*, 275, 446-454.

Hacia, J.G., (1999). *Resequencing and mutational analysis using oligonucleotide microarrays.* Nat Genet, 21(1 Suppl): p. 42-7.

Hamby RI, Zoneraich S, Sherman L (1974) Diabetic cardiomyopathy. JAMA 229(13):1749-54.

Hanlon, S.E. and J.D. Lieb, 2004). *Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays.* Curr Opin Genet Dev, 14(6): p. 697-705.

Han, J, Hajjar, DP, Febbraio, M & Nicholson, AC. (1997). Native and modified low density lipoproteins increase the functional expression of the macrophage class B scavenger receptor, CD36. *J Biol Chem*, 272, 21654-21659.

Han, J, Hajjar, DP, Tauras, JM & Nicholson, AC. (1999). Cellular cholesterol regulates expression of the macrophage type B scavenger receptor, CD36. *J Lipid Res*, 40, 830-838.

Han, J, Hajjar, DP, Tauras, JM, Feng, J, Gotto, AM, Jr. & Nicholson, AC. (2000). Transforming Growth Factor-β1(TGF-β1) and TGF-β2 Decrease Expression of CD36, the

Type B Scavenger Receptor, through Mitogen-activated Protein Kinase phosphorylation of Peroxisome Proliferator-activated Receptor-γ. *J Biol Chem*, 275, 1241-1246.

Han, S & Sidell, N. (2002). Peroxisome-proliferator-activated-receptor gamma (PPARγ) independent induction of CD36 in THP-1 monocytes by retinoic acid. *Immunology*, 106, 53-59.

Han, J, Hajjar, DP, Zhou, X, Gotto, AM, Jr. & Nicholson, AC. (2002). Regulation of peroxisome proliferator-activated receptor-γ-mediated gene expression. A new mechanism of action for high density lipoprotein. *J Biol Chem*, 277, 23582-23586.

Han, J, Parsons, M, Zhou, X, Nicholson, AC, Gotto, AM, Jr. & Hajjar, DP. (2004). Functional interplay between the macrophage scavenger receptor class B type I and pitavastatin (NK-104). *Circulation*, 110, 3472-3479.

Hastie, T., et al., (1999). *Imputing missing data for gene expression arrays*. Department of Statistics Stanford University.

Hedenfalk, I., et al., (2001). *Gene-expression profiles in hereditary breast cancer.* N Engl J Med, 344(8): p. 539-48.

Hess, K.R., et al., (2001). *Microarrays: handling the deluge of data and extracting reliable information.* Trends Biotechnol, 19(11): p. 463-8.

Hirai, M.Y., et al., (2004). *Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana.* Proc Natl Acad Sci U S A, 101(27): p. 10205-10.

Hoen, P.A., et al., (2004). *Intensity-based analysis of two-colour microarrays enables efficient and flexible hybridization designs.* Nucleic Acids Res, 32(4): p. e41.

Hong-Wu Ma et An-Ping Zeng, 2003. The connectivity structure, giant strong component and centrality of metabolic networks. Bio-informatics, 19(11) :1423–1430.

Hoosdally, SJ, Andress, EJ, Wooding, C, Martin, CA & Linton, KJ. (2009). The Human Scavenger Receptor CD36: glycosylation status and its role in trafficking and function. *J Biol Chem*, 284, 16277-16288.

Huh, HY, Lo, SK, Yesner, LM & Silverstein, RL. (1995). CD36 induction on human monocytes upon adhesion to tumor necrosis factor-activated endothelial cells. *J Biol Chem*, 270, 6267-6271.

Houdebine, l.m., (1998) Les animaux transgéniques, 1ère édition. Paris : Tec&Doc Lavoisier, 181 p.

Houdebine, L.M., (1999) Ethical implications of knock-out and transgenic techniques for animal research, In : Handbook of Molecular-Genetic Techniques for Brain and Behavior Research (Techniques in the Behavioral and Neural Sciences, vol. 13), Elsevier Science BV, Ed : CRUSIO, W.E., GERLAI, R.T., 936-948.

Holloway, A.J., et al., (2002). *Options available--from start to finish--for obtaining data from DNA microarrays II.* Nat Genet, 32 Suppl: p. 481-9.

Hong Yue, Martin Brown, Joshua Knowles, Hong Wang, David S Broomhead, et Douglas B Kell, 2006. Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis : a case study of an NF-kappaB signalling pathway. Mol Biosyst, 2(12) :640–649.

Hoyle, D.C., et al., (2002). *Making sense of microarray data distributions.* Bioinformatics, 18(4): p. 576-84

Huang, JT, Welch, JS, Ricote, M, Binder, CJ, Willson, TM, Kelly, C, Witztum, JL, Funk, CD, Conrad, D & Glass, CK. (1999). Interleukin-4-dependent production of PPAR-gamma ligands in macrophages by 12/15-lipoxygenase. *Nature*, 400, 378-382.

Huang, ZH, Lin, C-Y, Oram, JF & Mazzone, T. (2001). Sterol efflux mediated by endogenous macrophage ApoE expression in independent of ABCA1. *Arterioscler Thromb Vasc Biol*, 21, 2019-2025.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57.

Huh, HY, Lo, SK, Yesner, LM & Silverstein, RL. (1995). CD36 induction on human monocytes upon adhesion to tumor necrosis factor-activated endothelial cells. *J Biol Chem*, 270, 6267-6271.

Huh, HY, Pearce, SF, Yesner, LM, Schindler, JL & Silverstein, RL. (1996). Regulated expression of CD36 during monocyte-to-macrophage differentiation: potential role of CD36 in foam cell formation. *Blood*, 87, 2020-2028.

Hughes, T.R., et al., (2001). *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.* Nat Biotechnol, 19(4): p. 342-7.

Hui, DY. (2003). HIV protease inhibitors and atherosclerosis. *J Clin Invest*, 111, 317-318. .

Hwang, D., W.A. Schmitt, and G. Stephanopoulos, (2002). *Determination of minimum sample size and discriminatory expression patterns in microarray data.* Bioinformatics, 18(9): p. 1184-93.

Ibrahimi, A, Bonen, A, Blinn, WD, Hajri, T, Li, X, Zhong, K, Cameron, R & Abumrad, NA. (1999). Muscle-specific overexpression of FAT/CD36 enhances fatty acid oxidation by contracting muscle, reduces plasma triglycerides and fatty acids, and increases plasma glucose and insulin. *J Biol Chem*, 274, 26761-26766.

Inoue, M, Ohtake, T, Motomura, W, Takahashi, N, Hosoki, Y, Miyoshi, S, Suzuki, Y, Saito, H, Kohgo, Y & Okumura, T. (2005). Increased expression of PPARgamma in high fat diet-induced liver steatosis in mice. *Biochem Biophys Res Commun*, 336, 215-222.

Irene Nobeli et Janet M Thornton. , May 2006 A bioinformatician’s view of the metabolome. Bioessays, 28(5) :534–545.

Irwin, M.H., Moffatt, R.J., Pinkert, C.A., (1996) Identification of transgenic mice by PCR analysis of saliva, Nat Biotechnol, 14(9), 1146-1148

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. et Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. U.S.A., 98 (8), 4569–4574.

Iyer, V.R., et al., (2001). *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF*. Nature, 409(6819): p. 533-8.

Jaenisch, R., (1988) Transgenic Animals, Science, 240, 1468-1474

Jallat, S., (1991) Les souris transgéniques, Biofutur, 45, 3-1.

Janabi, M, Yamashita, S, Hirano, K, Sakai, N, Hiraoka, H, Matsumoto, K, Zhang, Z, Nozaki, S & Matsuzawa, Y. (2000). Oxidized LDL-induced NF-kappa B activation and subsequent expression of proinflammatory genes are defective in monocyte-derived macrophages from CD36-deficient patients. *Arterioscler Thromb Vasc Biol*, 20, 1953-1960.

Janne, J., Alhonen, L., (1996) Transgenic animals : pratical aspects, management, logistics, In : Proceedings of the Sixth Symposium of the Federation of European Laboratory Animal Science Associations, Harmonization of Laboratory Animal Husbandry, Basel, Switzerland, 19-21, London : The Royal Society of Medicine Press Limited, 41-44.

Janowski, BA, Willy, PJ, Devi, TR, Falck, JR & Mangelsdorf, DJ. (1996). An oxysterol signalling pathway mediated by the nuclear receptor LXR alpha. *Nature*, 383, 728-731.

Janowski, BA, Grogan, MJ, Jones, SA, Wisely, GB, Kliewer, SA, Corey, EJ & Mangelsdorf, DJ. (1999). Structural requirements of ligands for the oxysterol liver X receptors LXRalpha and LXRbeta. *Proc Natl Acad Sci U S A*, 96, 266-271.

Jensen L., Kuhn M., Stark M., Chaffron S., Creevey C., Muller J., Doerks T., Julien P., Roth A., Simonovic M., Bork P. et Mering C.(2008): String 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research.
Jeong,H., Mason,S.P., Barabas Nature, 411 (6833), 41–42.

Jeong, K.S., J. Ahn, and A.B. Khodursky, (2004). *Spatial patterns of transcriptional activity in the chromosome of Escherichia coli.* Genome Biol, 5(11): p. R86.

Jeppesen, J, Albers, PH, Rose, AJ, Birk, JB, Schjerling, P, Dzamko, N, Steinberg, GR & Kiens, B. (2011). Contraction-induced skeletal muscle FAT/CD36 trafficking and FA uptake is AMPK independent. *J Lipid Res*, 52, 699-711.

Jimenez, B, Volpert, OV, Crawford, SE, Febbraio, M, Silverstein, RL & Bouck, N. (2000). Signals leading to apoptosis-dependent inhibition of neovascularization by thrombospondin-1. *Nat Med*, 6, 41-48.1.

Kafatos, F.C., C.W. Jones, and A. Efstratiadis, (1979). *Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure.* Nucleic Acids Res, 7(6): p. 1541-52.

Kallioniemi, A., et al., (1992). *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.* Science, 258(5083): p. 818-21.

Kampa, D., et al., (2004). *Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.* Genome Res, 14(3): p. 331-42.

Kapranov, P., et al., (2002). *Large-scale transcriptional activity in chromosomes 21 and 22.* Science, 296(5569): p. 916-9.

Kane, M.D., et al., (2000). *Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.* Nucleic Acids Res, 2000. 28(22): p. 4552-7.

Kar, NS, Ashraf, MZ, Valiyaveettil, M & Podrez, EA. (2008). Mapping and characterization of the binding site for specific oxidized phospholipids and oxidized low density lipoprotein of scavenger receptor CD36. *J Biol Chem*, 283, 8765-8771.

Kaski, S., et al., (2003). *Trustworthiness and metrics in visualizing similarity of gene expression.* BMC Bioinformatics, 4: p. 48.

Kerr, M.K. and G.A. Churchill, (2001). *Statistical design and the analysis of gene expression microarray data.* Genet Res, 77(2): p. 123-8.

Kerr, M.K., M. Martin, and G.A. Churchill, (2000). *Analysis of variance for gene expression microarray data.* J Comput Biol, 7(6): p. 819-37.

Kepler, T.B., L. Crosby, and K.T. Morgan, (2002). *Normalization and analysis of DNA microarray data by self-consistency and local regression.* Genome Biol. 3(7): p. RESEARCH0037.

Kernbaum, S., (1998) Dictionnaire de Médecine, Ed Médecine-Sciences , 6ème édition, Flammarion, Paris.

Kim HW, Cho YS, Lee HR, Park SY, Kim YH (2001) Diabetic alterations in cardiac sarcoplasmic reticulum Ca^{2+} -ATPase and phospholamban protein expression. Science Life 70: 367-379.

Kim, K.Y., B.J. Kim, and G.S. Yi, (2004). *Reuse of imputed data in microarray analysis increases imputation efficiency.* BMC Bioinformatics,. 5: p. 160.

Koonen, DP, Jacobs, RL, Febbraio, M, Young, ME, Soltys, CL, Ong, H, Vance, DE & Dyck, JR. (2007). Increased hepatic CD36 expression contributes to dyslipidemia associated with diet-induced obesity. *Diabetes*, 56, 2863-2871.

Kooperberg, C., et al., (2002). *Improved background correction for spotted DNA microarrays.* J Comput Biol, 9(1): p. 55-66.

Kuang M, Febbraio M, Wagg C, Lopaschuk GD, Dyck JRB. (2004) Fatty acid translocase/CD36 deficiency does not energetically or functionally compromise hearts before or after ischemia. Circulation.;109:1550–1557.

Kothapalli, R., et al., (2002). *Microarray results: how accurate are they?* BMC Bioinformatics, 3: p. 22.

Krulewski, T.F., Neumann, P.E., Gordon, J.W., (1989) Insertional mutation in a transgenic mouse allelic with Purkinje cell degeneration, Proc. Natl. Acad. Sci. USA, 86(10), 3709-3712

Kuo, W.P., et al., *Analysis of matched mRNA measurements from two different microarray technologies.* Bioinformatics, 18(3): p. 405-12.

Kutalik, Z., et al., (2004). *Advanced significance analysis of microarray data based on weighted resampling: a comparative study and application to gene deletions in Mycobacterium bovis.* Bioinformatics, 20(3): p. 357-63.

Lake, J.P., Haines, D., Linder, C., Davisson, M., (1999) Dollars and Sense : Time and Cost factors critical to establishing genetically engineered mouse colonies, Lab. Animal, 28 (8), 24-3

Langmann, T, Liebisch, G, Moehle, C, Schifferer, R, Dayoub, R, Heiduczek, S, Grandl, M, Dada, A & Schmitz, G. (2005). Gene expression profiling identifies retinoids as potent inducers of macrophage lipid efflux. *Biochim Biophys Acta*, 1740, 155-161.

Lashkari, D.A., et al., (1997). *Yeast microarrays for genome wide parallel genetic and gene expression analysis.* Proc Natl Acad Sci U S A, 94(24): p. 13057-62.

Lawrence, N.D., et al., (2004). *Reducing the variability in cDNA microarray image processing by Bayesian inference.* Bioinformatics, 20(4): p. 518-26.

Li, AC, Binder, CJ, Gutierrez, A, Brown, KK, Plotkin, CR, Pattison, JW, Valledor, AF, Davis, RA, Willson, TM, Witztum, JL, Palinski, W & Glass, CK. (2004). Differential inhibition of macrophage foam-cell formation and atherosclerosis in mice by PPARalpha, beta/delta, and gamma. *J Clin Invest*, 114, 1564-1576.

Le Naour, F., et al., (2001). *Profiling changes in gene expression during differentiation and maturation of monocyte-derived dendritic cells using both oligonucleotide microarrays and proteomics.* J Biol Chem, 276(21): p. 17920-31.

Lee, C. and M. Roy, (2004). *Analysis of alternative splicing with microarrays: successes and challenges.* Genome Biol, 5(7): p. 231.

Lee, P.D., et al., (2002). *Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies.* Genome Res, 12(2): p. 292-7.

Leung, Y.F., (2002). *Unravelling the mystery of microarray data analysis.* Trends Biotechnol, 20(9): p. 366-8.

Leung, Y.F. and D. Cavalieri, (2003). *Fundamentals of cDNA microarray data analysis.* Trends Genet, 19(11): p. 649-59.

Li, AC, Brown, KK, Silvestre, MJ, Willson, TM, Palinski, W & Glass, CK. (2000). Peroxisome proliferator-activated receptor gamma ligands inhibit development of atherosclerosis in LDL receptor-deficient mice. *J Clin Invest*, 106, 523-531.

Li, C. and W.H. Wong, (2001). *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 98(1): p. 31-6.

Li, AC & Glass, CK. (2002). The macrophage foam cell as a target for therapeutic intervention. *Nat Med*, 8, 1235-1242.

Li, AC & Glass, CK. (2004). PPAR- and LXR-dependent pathways controlling lipid metabolism and the development of atherosclerosis. *J Lipid Res*, 45, 2161-2173.

Li, CM, Chung, BH, Presley, JB, Malek, G, Zhang, X, Dashti, N, Li, L, Chen, J, Bradley, K, Kruth, HS & Curcio, CA. (2005). Lipoprotein-like particles and cholesteryl esters in human Bruch's membrane: initial characterization. *Invest Ophthalmol Vis Sci*, 46, 2576-2586.

Liang, P. and A.B. (1992). Pardee, *Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction.* Science, 257(5072): p. 967-71.

Lipsky, RH, Eckert, DM, Tang, Y & Ockenhouse, CF. (1997). The carboxyl-terminal cytoplasmic domain of CD36 is required for oxidized low-density lipoprotein modulation of NF-kappaB activity by tumor necrosis factor-alpha. *Recept Signal Transduct*, 7, 1-11.

Lipshutz, R.J., et al., (1999). *High density synthetic oligonucleotide arrays.* Nat Genet, 21(1Suppl): p. 20-4.

Lockhart, D.J., et al., (1996). *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. 14(13): p. 1675-80.

Loganathan R, Bilgen M, Al-Hafez B, Zhero SV, Alenezy MD, Smirnova IV (2007) Exercise training improves cardiac performance in diabetes: in vivo demonstration with quantitative cine-MRI analyses. J Appl Physiol 102: 665-672.

Lopaschuk GD, Barr R, Thomas PD, Dyck JRB. (2003); Beneficial effects of trimetazidine in ex vivo working ischemic hearts are due to a stimulation of glucose oxidation secondary to inhibition of long-chain 3-ketoacyl coenzyme a thiolase. Circ Res.;93:e33–37.

Lopaschuk GD, Ussher JR, Folmes CDL, Jaswal JS, Stanley WC. (2010) Myocardial fatty acid metabolism in health and disease. Physiol Rev.;90:207–258.

Luiken JJ, Schaap FG, van Nieuwenhoven FA, van der Vusse GJ, Bonen A, Glatz JF. (1999) Cellular fatty acid transport in heart and skeletal muscle as facilitated by proteins. Lipids.;34 Suppl:S169–175

Luiken, JJ, Dyck, DJ, Han, XX, Tandon, NN, Arumugam, Y, Glatz, JF & Bonen, A. (2002). Insulin induces the translocation of the fatty acid transporter FAT/CD36 to the plasma membrane. *Am J Physiol Endocrinol Metab*, 282, E491-E495.

Luiken, JJ, Coort, SL, Willems, J, Coumans, WA, Bonen, A, Van Der Vusse, GJ & Glatz, JF. (2003). Contraction-induced fatty acid translocase/CD36 translocation in rat cardiac myocytes is mediated through AMP-activated protein kinase signaling. *Diabetes*, 52, 1627-1634.

Lyng, H., et al., (2004). *Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction.* BMC Genomics, 5(1): p. 10.

Maere,S., Heymans,K. et Kuiper,M. (2005) BiNGO : a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 21 (16), 3448–3449.

Madan Babu M., Nicholas M Luscombe, L. Aravind, Mark Gerstein, et Sarah A Teichman, 2004.. Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol, 14(3) :283–291.

Madrazo JA, Kelly DP. (2008); The PPAR trio: regulators of myocardial energy metabolism in health and disease. J Mol Cell Cardiol.;44:968–975.

Malaud, E, Hourton, D, Giroux, LM, Ninio, E, Buckland, R & McGregor, JL. (2002). The terminal six amino-acids of the carboxy cytoplasmic tail of CD36 contain a functional domain implicated in the binding and capture of oxidized low-density lipoprotein. *Biochem J*, 364, 507-515.

Mandosi, E, Fallarino, M, Gatti, A, Carnovale, A, Rossetti, M, Lococo, E, Buchetti, B, Filetti, S, Lenti, L & Morano, S. (2010). Atorvastatin downregulates monocyte CD36 expression, nuclear NFkappaB and TNFalpha levels in type 2 diabetes. *J Atheroscler Thromb*, 17, 539-545.

Martin W., Rujan T., Richly E., Hansen A., Cornelsen S., Lins T., Leister D., Stoebe B., Hasegawa M. et Penny D.(2002): Evolutionary analysis of arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA, 99(19) :12246–51.

Martoglio, A.M., et al., (2002). *A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer.* Bioinformatics, 18(12): p. 1617-24.

Maskos, U. and E.M. Southern, (1992). *Parallel analysis of oligodeoxyribonucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation.* Nucleic Acids Res, 20(7): p. 1675-8.

Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H (2012): The intAct molecular interaction database . Nucleic Acids Res 2012, 40(Database issue):D841–D846.

Mata, J., S. Marguerat, and J. Bahler, (2005). *Post-transcriptional control of gene expression: a genome-wide perspective.* Trends Biochem Sci, 30(9): p. 506-14.

Matsumoto, K, Hirano, K, Nozaki, S, Takamoto, A, Nishida, M, Nakagawa-Toyama, Y, Janabi, MY, Ohya, T, Yamashita, S & Matsuzawa, Y. (2000). Expression of macrophage (Mphi) scavenger receptor, CD36, in cultured human aortic smooth muscle cells in association with expression of peroxisome proliferator activated receptor-gamma, which regulates gain of Mphi-like phenotype in vitro, and its implication in atherogenesis. *Arterioscler Thromb Vasc Biol*, 20, 1027-1032.

McGarry JD. Banting lecture 2001: dysregulation of fatty acid metabolism in the etiology of type 2 diabetes. Diabetes 2002;51:7–18.

McLaren, JE, Michael, DR, Salter, RC, Ashlin, TG, Calder, CJ, Miller, AM, Liew, FY & Ramji, DP. (2010). IL-33 reduces macrophage foam cell formation. *J Immunol*, 185, 1222-1229.

Medeiros, LA, Khan, T, El Khoury, JB, Pham, CL, Hatters, DM, Howlett, GJ, Lopez, R, O'Brien, KD & Moore, KJ. (2004). Fibrillar amyloid protein present in atheroma activates CD36 signal transduction. *J Biol Chem*, 279, 10643-10648.

Mencarelli, A, Renga, B, Distrutti, E & Fiorucci, S. (2009). Antiatherosclerotic effect of farnesoid X receptor. *Am J Physiol Heart Circ Physiol*, 296, H272-H281.

Mencarelli, A, Cipriani, S, Renga, B, Francisci, D, Palladino, G, Distrutti, E, Baldelli, F & Fiorucci, S. (2010). The bile acid sensor FXR protects against dyslipidemia and aortic plaques development induced by the HIV protease inhibitor ritonavir in mice. *PLoS One*, 5, e13238.

Michelle L Green et Peter D Karp, 2004. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinformatics, 5 :76

Miki, S, Horikawa, K, Nishizumi, H, Suemura, M, Sato, B, Yamamoto, M, Takatsu, K, Yamamoto, T & Miki, Y. (2001). Reduction of atherosclerosis despite hypercholesterolemia in lyn-deficient mice fed a high-fat diet. *Genes Cells*, 6, 37-42.

Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., et Alon U., 2002. Network motifs : simple building blocks of complex networks. Science, 298(5594) :824–827.

Moody, D.E., (2001). *Genomics techniques: An overview of methods for the study of gene expression.* J. Anim. Sci., 79: p. E128-E135.

Mockler, T.C., et al., (2005). *Applications of DNA tiling arrays for whole-genome analysis.* Genomics, 85(1): p. 1-15.

Monnereau, L.,(1997) Cours de Biologie du Développement, 2ème année de 1 er cycle, Ecole Nationale Vétérinaire de Toulouse, Communication personnelle.

Moore, KJ, Rosen, ED, Fitzgerald, ML, Randow, F, Andersson, LP, Altshuler, D, Milstone, DS, Mortensen, RM, Spiegelman, BM & Freeman, MW. (2001). The role of PPAR-gamma in macrophage differentiation and cholesterol uptake. *Nat Med*, 7, 41-47.

Moore, KJ & Freeman, MW. (2006). Scavenger receptors in atherosclerosis: beyond lipid uptake. *Arterioscler Thromb Vasc Biol*, 26, 1702-1711.

Muller, H, Deckers, K & Eckel, J. (2002). The fatty acid translocase (FAT)/CD36 and the glucose transporter GLUT4 are localized in different cellular compartments in rat cardiac muscle. *Biochem Biophys Res Commun*, 293, 665-669.

Munteanu, A, Ricciarelli, R & Zingg, JM. (2004). HIV protease inhibitors-induced atherosclerosis: prevention by alpha-tocopherol. *IUBMB Life*, 56, 629-631.

Munteanu, A, Zingg, JM, Ricciarelli, R & Azzi, A. (2005). CD36 overexpression in ritonavir-treated THP-1 cells is reversed by alpha-tocopherol. *Free Radic Biol Med*, 38, 1047-1056.

Nagy, L, Tontonoz, P, Alvarez, JG, Chen, H & Evans, RM. (1998). Oxidized LDL regulates macrophage gene expression through ligand activation of PPARgamma. *Cell*, **93**, 229-240.

Nassir, F, Wilson, B, Han, X, Gross, RW & Abumrad, NA. (2007). CD36 is important for fatty acid and cholesterol uptake by the proximal but not distal intestine. *J Biol Chem*, **282**, 19493-19501.

Nauli, AM, Nassir, F, Zheng, S, Yang, Q, Lo, CM, Vonlehmden, SB, Lee, D, Jandacek, RJ, Abumrad, NA & Tso, P. (2006). CD36 is important for chylomicron formation and secretion and may mediate cholesterol uptake in the proximal intestine. *Gastroenterology*, **131**, 1197-1207.

Navazo, MD, Daviet, L, Savill, J, Ren, Y, Leung, LL & McGregor, JL. (1996). Identification of a domain (155-183) on CD36 implicated in the phagocytosis of apoptotic neutrophils. *J Biol Chem*, **271**, 15381-15385.

Netticadan T, Temsah RM, Kent A, Elimban V, Dhalla NS (2001) Depressed levels of Ca2+-cycling proteins may underlie sarcoplasmic reticulum dysfunction in the diabetic heart. Diabetes 50: 2133-2138.

Newton, M.A., et al., (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol, 8(1): p. 37-52.

Niehrs, C. and N. (1999). Pollet, *Synexpression groups in eukaryotes.* Nature, **402**(6761): p. 483-7.

Nicholson, AC, Febbraio, M, Han, J, Silverstein, RL & Hajjar, DP. (2000). CD36 in atherosclerosis. The role of a class B macrophage scavenger receptor. *Ann N Y Acad Sci*, **902**, 128-131.

Nicholson, AC, Frieda, S, Pearce, A & Silverstein, RL. (1995). Oxidized LDL binds to CD36 on human monocyte-derived macrophages and transfected cell lines. Evidence implicating the lipid moiety of the lipoprotein as the binding site. *Arterioscler Thromb Vasc Biol*, **15**, 269-275.

Nimgaonkar, A., et al., (2003). *Reproducibility of gene expression across generations of Affymetrix microarrays.* BMC Bioinformatics, **4**: p. 27.

Oba, S., et al., (2003). *A Bayesian missing value estimation method for gene expression profile data.* Bioinformatics, 19(16): p. 2088-96.

Oh, J, Weng, S, Felton, SK, Bhandare, S, Riek, A, Butler, B, Proctor, BM, Petty, M, Chen, Z, Schechtman, KB, Bernal-Mizrachi, L & Bernal-Mizrachi, C. (2009). 1,25(OH)2 vitamin d inhibits foam cell formation and suppresses macrophage cholesterol uptake in patients with type 2 diabetes mellitus. *Circulation*, 120, 687-698.

Ohgami, N, Nagai, R, Ikemoto, M, Arai, H, Kuniyasu, A, Horiuchi, S & Nakayama, H. (2001). CD36, a member of class B scavenger receptor family, is a receptor for advanced glycation end products. *Ann N Y Acad Sci*, 947, 350-355.

Okumura T, Jamieson GA. Platelet glycocalicin. I. Orientation of glycoproteins of the human platelet surface. *J Biol Chem.* 1976; 251(19): 5944-9.

Opie LH. The heart (1998), Physiology, from cell to circulation. Lippincott-Raven.

Opie L.(1991), The heart Physiology and metabolism. New York : Raven Press.

Oquendo, P, Hundt, E, Lawler, J & Seed, B. (1989). CD36 directly mediates cytoadherence of Plasmodium falciparum parasitized erythrocytes. *Cell*, 58, 95-101.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. et Maltsev,N. (1999) Use of contiguity on the chromosome to predict functional coupling. In Silico Biol. (Gedrukt), 1 (2), 93–108.

Ozer, NK, Negis, Y, Aytan, N, Villacorta, L, Ricciarelli, R, Zingg, JM & Azzi, A. (2006). Vitamin E inhibits CD36 scavenger receptor expression in hypercholesterolemic rabbits. *Atherosclerosis*, 184, 15-20.

Pang,K., Sheng,H. et Ma,X. (2010) Understanding gene essentiality by finely characterizing hu the yeast protein interaction network. Biochem. Biophys. Res. Commun., 401 (1), 112–116.

Patriotis, P.C., et al., (2001). *ArrayExplorer, a program in Visual Basic for robust and accurate filter cDNA array analysis.* Biotechniques, 31(4): p. 862, 864, 866-8, 870, 872.

Pearce, SFA, Wu, J & Silverstein, RL. (1995). Recombinant GST/CD36 fusion proteins define a thrombospondin binding domain. Evidence for a single calcium-dependent binding site on CD36. *J Biol Chem*, 270, 2981-2986.

Pearce, SF, Roy, P, Nicholson, AC, Hajjar, DP, Febbraio, M & Silverstein, RL. (1998). Recombinant glutathione S-transferase/CD36 fusion proteins define an oxidized low density lipoprotein-binding domain. *J Biol Chem*, 273, 34875-34881.

Peter D Karp, Suzanne Paley, et Pedro Romero, 2002. The Pathway Tools software. Bioinformatics, 18 Suppl 1 :S225–S232.

Pease, A.C., et al., (1994). *Light-generated oligonucleotide arrays for rapid DNA sequence analysis.* Proc Natl Acad Sci U S A, 91(11): p. 5022-6.

Pietsch, A, Erl, W & Lorenz, RL. (1996). Lovastatin reduces expression of the combined adhesion and scavenger receptor CD36 in human monocytic cells. *Biochem Pharmacol*, 52, 433-439.

Prieto, J, Eklund, A & Patarroyo, M. (1994). Regulated expression of integrins and other adhesion molecules during differentiation of monocytes into macrophages. *Cell Immunol*, 156, 191-211.

Podrez, EA, Poliakov, E, Shen, Z, Zhang, R, Deng, Y, Sun, M, Finton, PJ, Shan, L, Febbraio, M, Hajjar, DP, Silverstein, RL, Hoff, HF, Salomon, RG & Hazen, SL. (2002a). A novel family of atherogenic oxidized phospholipids promotes macrophage foam cell formation via the scavenger receptor CD36 and is enriched in atherosclerotic lesions. *J Biol Chem*, 277, 38517-38523.

Podrez, EA, Poliakov, E, Shen, Z, Zhang, R, Deng, Y, Sun, M, Finton, PJ, Shan, L, Gugiu, B, Fox, PL, Hoff, HF, Salomon, RG & Hazen, SL. (2002b). Identification of a novel family of oxidized phospholipids that serve as ligands for the macrophage scavenger receptor CD36. *J Biol Chem*, 277, 38503-38516.

Pollack, J.R., et al., (1999). *Genome-wide analysis of DNA copy-number changes using cDNA microarrays.* Nat Genet, 23(1): p. 41-6.

Pollock, J.D., (2002). *Gene expression profiling: methodological challenges, results, and prospects for addiction research.* Chem Phys Lipids, 121(1-2): p. 241-56.

Puente Navazo, MD, Daviet, L, Ninio, E & McGregor, JL. (1996). Identification on human CD36 of a domain (155-183) implicated in binding oxidized low-density lipoproteins (Ox-LDL). *Arterioscler Thromb Vasc Biol*, 16, 1033-1039.

Puccetti, L, Sawamura, T, Pasqui, AL, Pastorelli, M, Auteri, A & Bruni, F. (2005). Atorvastatin reduces platelet-oxidized-LDL receptor expression in hypercholesterolaemic patients. *Eur J Clin Invest*, 35, 47-51.

Quackenbush, J., (2001). *Computational analysis of microarray data.* Nat Rev Genet, 2(6): p. 418-27.

Quackenbush, J., (2002). *Microarray data normalization and transformation.* Nat Genet, 32 Suppl: p. 496-501.

Querec, T.D., et al., (2004). *A novel approach for increasing sensitivity and correcting saturation artifacts of radioactively labeled cDNA arrays.* Bioinformatics, 20(12): p. 1955-61.

Rac, ME, Safranow, K & Poncyljusz, W. (2007). Molecular basis of human CD36 gene mutations. *Mol Med*, 13, 288-296.

Rahaman S.O., Lennon D.J., M. Febbraio, E.A. Podrez, S.L. Hazen, and R.L. Silverstein (2006)- A CD36-dependent signaling cascade is necessary for macrophage foam cell formation ; PMC.

Randle PJ, Garland PB, Hales CN, Newsholme EA.(1993) The glucose fattyacid cycle. Its role in insulin sensitivity and the metabolic disturbances of diabetes mellitus.1:785–9.

Randle PJ. (1998); Regulatory interactions between lipids and carbohydrates: the glucose fatty acid cycle after 35 years. Diabetes Metab Rev.;14:263–283.

Rahaman, SO, Lennon, DJ, Febbraio, M, Podrez, EA, Hazen, SL & Silverstein, RL. (2006). A CD36-dependent signaling cascade is necessary for macrophage foam cell formation. *Cell Metab*, 4, 211-221.

Ren, S., Li , M., Cai, H., Hudgins, S., Furth, P.A., (2001) A Simplified Method to Prepare PCR Template DNA for Screening of Transgenic and Knockout Mice, Contemporary Topics in Laboratory Animal Science, 40 (2), 27-30.

Ren, J, Jin, W & Chen, H. (2010). oxHDL decreases the expression of CD36 on human macrophages through PPARgamma and p38 MAP kinase dependent mechanisms. *Mol Cell Biochem*, 342, 171-181.

Ren, Y, Silverstein, RL, Allen, J & Savill, J. (1995). CD36 gene transfer confers capacity for phagocytosis of cells undergoing apoptosis. *J Exp Med*, 181, 1857-1862.

Ren, Y, Silverstein, RL, Allen, J & Savill, J. (1995). CD36 gene transfer confers capacity for phagocytosis of cells undergoing apoptosis. *J Exp Med*, 181, 1857-1862.

Renn, S.C., N. Aubin-Horth, and H.A. Hofmann, (2004). *Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray.* BMC Genomics, 5(1): p. 42.

Ricciarelli, R, Zingg, JM & Azzi, A. (2000). Vitamin E reduces the uptake of oxidized LDL by inhibiting CD36 scavenger receptor expression in cultured aortic smooth muscle cells. *Circulation*, 102, 82-87.

Riek, AE, Oh, J & Bernal-Mizrachi, C. (2010). Vitamin D regulates macrophage cholesterol metabolism in diabetes. *J Steroid Biochem Mol Biol*, 121, 430-433.

Rigotti, A, Acton, SL & Krieger, M. (1995). The class B scavenger receptors SR-BI and CD36 are receptors for anionic phospholipids. *J Biol Chem*, 270, 16221-16224.

Rigotti, A, Trigatti, BL, Penman, M, Rayburn, H, Herz, J & Krieger, M. (1997). A targeted mutation in the murine gene encoding the high density lipoprotein (HDL) receptor scavenger receptor class B type I reveals its key role in HDL metabolism. *Proc Natl Acad Sci U S A*, 94, 12610-12615.

Riva, A., et al., (2004). *The difficult interpretation of transcriptome data: the case of the GATC regulatory network.* Comput Biol Chem, 28(2): p. 109-18.

Rong, X, Li, Y, Ebihara, K, Zhao, M, Kusakabe, T, Tomita, T, Murray, M & Nakao, K. (2010). Irbesartan treatment up-regulates hepatic expression of PPARalpha and its target genes in obese Koletsky (fa(k)/fa(k)) rats: a link to amelioration of hypertriglyceridaemia. *Br J Pharmacol*, 160, 1796-1807.

Roths, J.B., Foxworth, W.B., Mcarthur, M.J., Montgomery, C.A., Kier, A.B., (1999) Spontaneous and Engineered Mutant Mice as Models for Experimental and Comparative Pathology : History, Comparison, and Developmental Technology, Laboratory Animal Science, 49 (1), 12-34.

Rubic, T & Lorenz, RL. (2006). Downregulated CD36 and oxLDL uptake and stimulated ABCA1/G1 and cholesterol efflux as anti-atherosclerotic mechanisms of interleukin-10. *Cardiovasc Res*, 69, 527-535.

Ruiz-Velasco, N, Dominguez, A & Vega, MA. (2004). Statins upregulate CD36 expression in human monocytes, an effect strengthened when combined with PPAR-gamma ligands Putative contribution of Rho GTPases in statin-induced CD36 expression. *Biochem Pharmacol*, 67, 303-313.

Rung, J., et al., (2002). *Building and analysing genome-wide gene disruption networks.* Bioinformatics, 18 Suppl 2: p. S202-10.

Rodriguez, B.A. and T.H. Huang,(2005). *Tilling the chromatin landscape: emerging methods for the discovery and profiling of protein-DNA interactions.* Biochem Cell Biol, 83(4): p. 525-34.

Ryeom, SW, Sparrow, JR & Silverstein, RL. (1996). CD36 participates in the phagocytosis of rod outer segments by retinal pigment epithelium. *J Cell Sci*, 109 (Pt 2), 387-395.

Sahoo, D & Phillips, MC. (2008). CD36 mediates both cellular uptake of very long chain fatty acids and their intestinal absorption in mice. *J Biol Chem*, 283, 13108-13115.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., Arnheim, N., 1985) Enzymatic amplification of β-Globin Genomic Sequences and restriction Site Analysis for Diagnosis of Sickle Cell Anemia, Scienc, 230, 13501354

Sampson, MJ, Davies, IR, Braschi, S, Ivory, K & Hughes, DA. (2003). Increased expression of a scavenger receptor (CD36) in monocytes from subjects with Type 2 diabetes. *Atherosclerosis*, 167, 129-134.

Savill, J, Hogg, N, Ren, Y & Haslett, C. (1992). Thrombospondin cooperates with CD36 and the vitronectin receptor in macrophage recognition of neutrophils undergoing apoptosis. *J Clin Invest*, 90, 1513-1522.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. (2009): PID: the Pathway Interaction Database. Nucleic Acids Res, 37(Database issue):D674–679.

Schena, M., et al., (1995). *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 270(5235): p. 467-70.

Sherman BT, Huang DW, Tan Q, Guo Y, Bour S, et al. (2007) DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics 8: 426.

Silverstein, RL & Febbraio, M. (2009). CD36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior. *Sci Signal*, 2, re3.

Simantov, R, Febbraio, M, Crombie, R, Asch, AS, Nachman, RL & Silverstein, RL. (2001). Histidine-rich glycoprotein inhibits the antiangiogenic effect of thrombospondin-1. *J Clin Invest*, 107, 45-52.

Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L. et Ideker,T. (2011) Cytoscape 2.8 : new features for data integration and network visualization. Bioinformatics (Oxford, England), 27 (3), 431–432.

Southern, E.M., U. Maskos, and J.K. Elder, (1992). *Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models.* Genomics, 13(4): p. 1008-17.

Stuart,J.M., Segal,E., Koller,D. et Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science, 302 (5643), 249–255.

Su, A.I., et al., (2001). *Molecular classification of human carcinomas by use of gene expression signatures.* Cancer Res, 61(20): p. 7388-93.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005) Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression

profiles. Proceedings of the National Academy of Sciences of the United States of America, 102 (43), 15545–50.

Sun, X.J., Rothenberg, P., Kahn, C.R., Backer, J.M., Araki, E., Wilden, P.A., Cahill, D.A., Goldstein, **B.J. and White, M.F.** (1991) - Structure of the insulin receptor substrate IRS-1 defines a unique signal transduction protein. *Nature*, **352** (6330) : 73-7.

Swerlick, RA, Lee, KH, Wick, TM & Lawley, TJ. (1992). Human dermal microvascular endothelial but not human umbilical vein endothelial cells express CD36 in vivo and in vitro. *J Immunol*, 148, 78-83.

Taegtmeyer H., (2004) Cardiac metabolism as a target for the treatment of heart failure. Circulation;110:894–896.

Taegtmeyer H, Salazar R. (2007); Myocardial metabolism: a new target for thetreatment of heart disease? CurrHypertensRep. 2004 Dec;6(6):414-5.

Taketa, K, Matsumura, T, Yano, M, Ishii, N, Senokuchi, T, Motoshima, H, Murata, Y, Kim-Mitsuyama, S, Kawada, T, Itabe, H, Takeya, M, Nishikawa, T, Tsuruzoe, K & Araki, E. (2008). Oxidized low density lipoprotein activates peroxisome proliferator-activated receptor-alpha (PPARalpha) and PPARgamma through MAPK-dependent COX-2 expression in macrophages. *J Biol Chem*, 283, 9852-9862.

Talle MA, Allegar N, Makowski M, Rao PE, Mittler RS, Goldstein G. Classification of human lymphocytes and monocytes with the OK series of monoclonal antibodies. *Diagn Immunol*. 1983; 1(3): 129-35.

Talle MA, Rao PE, Westberg E, Allegar N, Makowski M, Mittler RS, Goldstein G. Patterns of antigenic expression on human monocytes as defined by monoclonal antibodies. *Cell Immunol*. 1983; 78(1): 83-99.

Tandon, NN, Kralisz, U & Jamieson, GA. (1989a). Identification of glycoprotein IV (CD36) as a primary receptor for platelet-collagen adhesion. *J Biol Chem*, 264, 7576-7583.

Tandon, NN, Lipsky, RH, Burgess, WH & Jamieson, GA. (1989b). Isolation and characterization of platelet glycoprotein IV (CD36). *J Biol Chem*, 264, 7570-7575.

Tao, N, Wagner, SJ & Lublin, DM. (1996). CD36 is palmitoylated on both N- and C-terminal cytoplasmic tails. *J Biol Chem*, 271, 22315-22320.

Templin, M.F., et al., (2002). *Protein microarray technology*. Trends Biotechnol, 20(4): p. 160-6.

Thygesen, H.H. and A.H. Zwinderman, (2004). *Comparing transformation methods for DNA microarray data*. BMC Bioinformatics, 5: p. 77.

Townes, T.M., Lingrel, J.B., Chen, H.Y., Brinster, R.L., Palmiter, R.D., (1985) Erythroid-specific expression of human beta-globin genes in transgenic mice, EMBO J, 4(7), 1715-1723.

Tontonoz, P & Mangelsdorf, DJ. (2003). Liver X receptor signaling pathways in cardiovascular disease. *Mol Endocrinol*, 17, 985-993.

Tontonoz, P & Nagy, L. (1999). Regulation of macrophage gene expression by peroxisome-proliferator-activated receptor gamma: implications for cardiovascular disease. *Curr Opin Lipidol*, 10, 485-490.

Tontonoz, P, Nagy, L, Alvarez, JG, Thomazy, VA & Evans, RM. (1998). PPARgamma promotes monocyte/macrophage differentiation and uptake of oxidized LDL. *Cell*, 93, 241-252.

Tontonoz, P & Mangelsdorf, DJ. (2003). Liver X receptor signaling pathways in cardiovascular disease. *Mol Endocrinol*, 17, 985-993.

Tontonoz, P & Nagy, L. (1999). Regulation of macrophage gene expression by peroxisome-proliferator-activated receptor gamma: implications for cardiovascular disease. *Curr Opin Lipidol*, 10, 485-490.

Tontonoz, P, Nagy, L, Alvarez, JG, Thomazy, VA & Evans, RM. (1998). PPARgamma promotes monocyte/macrophage differentiation and uptake of oxidized LDL. *Cell*, 93, 241-252.

Troyanskaya, O., et al., (2001). *Missing value estimation methods for DNA microarrays.* Bioinformatics, 17(6): p. 520-5.

Tschentscher, F., et al., (2003). *Tumor classification based on gene expression profiling shows that uveal melanomas with and without monosomy 3 represent two distinct entities.* Cancer Res, 63(10): p. 2578-84.

Turk, R., et al., (2004). *Gene expression variation between mouse inbred strains.* BMC Genomics, 2004. 5(1): p. 57.

Tusher, V.G., R. Tibshirani, and G. Chu, (2001). *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 98(9): p. 5116-21.

Venugopal, SK, Devaraj, S & Jialal, I. (2004). RRR-alpha-tocopherol decreases the expression of the major scavenger receptor, CD36, in human macrophages via inhibition of tyrosine kinase (Tyk2). *Atherosclerosis*, 175, 213-220.

Von Mering C., Jensen L. J., Kuhn M., Chaffron S., Doerks T., Krüger B., Snel B. et Bork P. (2007): String 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res, 35(Database issue) :D358–62.

Wanders RJ, Vreken P, den Boer ME, Wijburg FA, van Gennip AH, IJlst L. (1999) Disorders of mitochondrial fatty acyl-CoA beta-oxidation. J Inherit Metab Dis.;22:442–487.

Wang, D.G., et al.,(1998). *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 280(5366): p. 1077-82.

Warwick Tucker et Vincent Moulton, 2005. Reconstructing metabolic networks using interval analysis. Proceedings of 5th Workshop on Algorithms for BioInformatics (WABI'05), Lecture Notes in BioInformatics, subseries Lecture Notes in Computer Science, 3692, 192–203.

Wheeler, D.B., A.E. Carpenter, and D.M. Sabatini, (2005). *Cell microarrays and RNA interference chip away at gene function.* Nat Genet, 37 Suppl: p. S25-30.

Whitehead, A. and D.L. (2005). Crawford, *Variation in tissue-specific gene expression among natural populations.* Genome Biol, 6(2): p. R13.

Whitfield, M.L., et al., (2002). *Identification of genes periodically expressed in the human cell cycle and their expression in tumors.* Mol Biol Cell, 13(6): p. 1977-2000.

Winzeler, E.A., et al., (1998). *Direct allelic variation scanning of the yeast genome.* Science,. 281(5380): p. 1194-7.

Wittig U. et De Beuckelaer A., 2001. Analysis and comparison of metabolic pathway databases. Brief Bioinform, 2(2) :126–142.

Yamada, K., et al., (2003). *Empirical analysis of transcriptional activity in the Arabidopsis genome.* Science, 302(5646): p. 842-6.

Yang, G. and S. Komatsu, (2004). *Microarray and proteomic analysis of brassinosteroid- and gibberellin-regulated gene and protein expression in rice.* Genomics Proteomics Bioinformatics, 2(2): p. 77-83.

Yang, Y.H. and T. Speed, (2002). *Design issues for cDNA microarray experiments.* Nat Rev Genet, 3(8): p. 579-88.

Yang, Y, Chen, M, Loux, TJ & Harmon, CM. (2007). Regulation of FAT/CD36 mRNA gene expression by long chain fatty acids in the differentiated 3T3-L1 cells. *Pediatr Surg Int*, 23, 675-683.

Yano, M, Matsumura, T, Senokuchi, T, Ishii, N, Murata, Y, Taketa, K, Motoshima, H, Taguchi, T, Sonoda, K, Kukidome, D, Takuwa, Y, Kawada, T, Brownlee, M, Nishikawa, T & Araki, E. (2007). Statins activate peroxisome proliferator-activated receptor gamma through extracellular signal-regulated kinase 1/2 and p38 mitogen-activated protein kinase-dependent cyclooxygenase-2 expression in macrophages. *Circ Res*, 100, 1442-1451.

Ye, Y, Nishi, SP, Manickavasagam, S, Lin, Y, Huang, MH, Perez-Polo, JR, Uretsky, BF & Birnbaum, Y. (2007). Activation of peroxisome proliferator-activated receptor-gamma (PPAR-gamma) by atorvastatin is mediated by 15-deoxy-delta-12,14-PGJ2. *Prostaglandins Other Lipid Mediat*, 84, 43-53.

Yesner, LM, Huh, HY, Pearce, SF & Silverstein, RL. (1996). Regulation of monocyte CD36 and thrombospondin-1 expression by soluble mediators. *Arterioscler Thromb Vasc Biol*, 16, 1019-1025

Yuen, T., et al., (2002). *Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays.* Nucleic Acids Res, 30(10): p. e48.

Yang, Y.H., et al., (2002). *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Res, 30(4): p. e15.

Yeung, K.Y., M. Medvedovic, and R.E. Bumgarner, (2004). *From co-expression to co-regulation: how many microarray experiments do we need?* Genome Biol, 5(7): p. R48.

Yun, MR, Im, DS, Lee, SJ, Park, HM, Bae, SS, Lee, WS & Kim, CD. (2009). 4-Hydroxynonenal enhances CD36 expression on murine macrophages via p38 MAPK-mediated activation of 5-lipoxygenase. *Free Radic Biol Med*, 46, 692-698.

Yun, MR, Im, DS, Lee, SJ, Woo, JW, Hong, KW, Bae, SS & Kim, CD. (2008). 4-hydroxynonenal contributes to macrophage foam cell formation through increased expression of class A scavenger receptor at the level of translation. *Free Radic Biol Med*, 45, 177-183.

Zecher D, Thuerauf DJ., Hafford DS. (1997) A role for the p38 mitogen-activated protein kinase pathway in myocardial cell growth, sarcomeric organization, and cardiac-specific gene expression. J Cell Biol, 1997, 139, 115-127.

Zhao, Y., M.C. Li, and R. Simon, (2005). *An adaptive method for cDNA microarray normalization.* BMC Bioinformatics, 6(1): p. 28.

Zhao, ZZ, Wang, Z, Li, GH, Wang, R, Tan, JM, Cao, X, Suo, R & Jiang, ZS. (2011). Hydrogen sulfide inhibits macrophage-derived foam cell formation. *Exp Biol Med (Maywood)*, 236, 169-176.

Zingg, JM, Libinaki, R, Lai, CQ, Meydani, M, Gianello, R, Ogru, E & Azzi, A. (2010). Modulation of gene expression by alpha-tocopherol and alpha-tocopheryl phosphate in THP-1 monocytes. *Free Radic Biol Med*, 49, 1989-2000.

Zhou, X., X. Wang, and E.R. Dougherty, (2003). *Missing-value estimation using linear and non-linear regression with Bayesian gene selection.* Bioinformatics, 19(17): p. 2302-7.

Zhou, J, Zhai, Y, Mu, Y, Gong, H, Uppal, H, Toma, D, Ren, S, Evans, RM & Xie, W. (2006). A novel pregnane X receptor-mediated and sterol regulatory element-binding protein-independent lipogenic pathway. *J Biol Chem*, 281, 15013-15020.

Zhou, J, Febbraio, M, Wada, T, Zhai, Y, Kuruba, R, He, J, Lee, JH, Khadem, S, Ren, S, Li, S, Silverstein, RL & Xie, W. (2008). Hepatic fatty acid transporter Cd36 is a common target of LXR, PXR, and PPARgamma in promoting steatosis. *Gastroenterology*, 134, 556-567.

Zhou, C, King, N, Chen, KY & Breslow, JL. (2009). Activation of PXR induces hypercholesterolemia in wild-type and accelerates atherosclerosis in apoE deficient mice. *J Lipid Res*, 50, 2004-2013.

Zuckerman, SH, Panousis, C, Mizrahi, J & Evans, G. (2000). The effect of gamma-interferon to inhibit macrophage-high density lipoprotein interactions is reversed by 15-deoxy-delta12,14-prostaglandin J2. *Lipids*, 35, 1239-1247.