Université de Montréal

# Développement d'une méthode bio-informatique permettant de relier les gènes aux métabolites

par
Sarah Cherkaoui

Département de Biochimie et Médecine Moléculaire
Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade de Maîtrise
en Bio-informatique
option Recherche

Décembre, 2015

Université de Montréal
Faculté des études supérieures

Ce mémoire est intitulé :

# Développement d'une méthode bio-informatique permettant de relier les gènes aux métabolites

Présentée par
Sarah Cherkaoui

a été évalué par un jury composé des personnes suivantes :

Sébastien Lemieux
Président-rapporteur

Christine Des Rosiers
Directeur de recherche

John D. Rioux
Codirecteur

Guillaume Lettre
Codirecteur

Gertraud Burger
Membre du jury

# Résumé

L'objectif de ce projet était de faire le lien entre gènes et métabolites afin d'éventuellement proposer des métabolites à mesurer en lien avec la fonction de gènes. Plus particulièrement, nous nous sommes intéressés aux gènes codant pour des protéines ayant un impact sur le métabolisme, soit les enzymes qui catalysent les réactions faisant partie intégrante des voies métaboliques. Afin de quantifier ce lien, nous avons développé une méthode bio-informatique permettant de calculer la distance qui est définie comme le nombre de réactions entre l'enzyme encodée par le gène et le métabolite dans la carte globale du métabolisme de la base de données *Kyoto Encyclopedia of Genes and Genomes* (KEGG). Notre hypothèse était que les métabolites d'intérêt sont des substrats/produits se trouvant à proximité des réactions catalysées par l'enzyme encodée par le gène. Afin de tester cette hypothèse et de valider la méthode, nous avons utilisé les études d'association pangénomique combinées à la métabolomique (mGWAS) car elles rapportent des associations entre variants génétiques, annotés en gènes, et métabolites mesurés. Plus précisément, la méthode a été appliquée à l'étude mGWAS par Shin *et al*. Bien que la couverture des associations de Shin *et al.* était limitée (24/299), nous avons pu valider de façon significative la proximité entre gènes et métabolites associés ($P<0,01$). En somme, cette méthode et ses développements futurs permettront d'interpréter de façon quantitative les associations mGWAS, de prédire quels métabolites mesurer en lien avec la fonction d'un gène et, plus généralement, de permettre une meilleure compréhension du contrôle génétique sur le métabolisme.

**Mots-clés** : Métabolomique, Génomique, Voies métaboliques, mGWAS, KEGG

# Abstract

The objective of this project was to link genes and metabolites in order to ultimately predict which metabolites to measure in order to adequately reflect the function of a given gene. Specifically, we were interested in genes, which code for proteins that regulate substrate metabolism, hence enzymes that catalyze reactions that are part of metabolic pathways. In order to quantify this link, we have developed a bioinformatics method to calculate a distance, which is defined as the number of reactions separating a given selected gene-encoded enzyme and its metabolite of interest in Kyoto Encyclopedia of Genes and Genomes (KEGG) database's metabolic overview map. Our hypothesis was that metabolites of interest are products/substrates found at proximity of the reactions catalyzed by the selected gene-encoded enzyme. In order to test our hypothesis and validate the method, we have used genome-wide association study of metabolites levels (mGWAS) because these studies report associations between genetic variants, annotated to genes, and measured metabolites. More specifically, we used the mGWAS conducted by Shin *et al.* Even though the coverage of the associations reported by Shin *et al*. was limited (24/299), we significantly validated the proximity between gene-metabolite associated pairs ($P<0.01$). Overall, the method and its future developments will allow the quantitative interpretation of mGWAS associations, predict which metabolite to measure with regards to the function of a gene and, in general, enable a better understanding of the genetic control of metabolism.

**Keywords**: Metabolomics, Genomics, Metabolic Pathways, mGWAS, KEGG

# Table des matières

# Liste des tableaux

## Article

# Liste des figures

## Introduction



## Article

# Liste des documents spéciaux

Tableaux supplémentaires de l'article : MatSup-Article.xlsx

Supplementary Table 1. Genes Identification in KEGG

Supplementary Table 2. Metabolites Identification in KEGG

Supplementary Table 3. Associations Identification in KEGG

Supplementary Table 4. Difference between results from PathQuant compared to Shin *et al.* biological annotations

# Liste des sigles et abréviations

Tous les mots écrits en italique dans ce mémoire sont dans une autre langue que le français.

A : Adénine

ADN : Acide Désoxyribonucléique

ARNm : Acide Ribonucléique messager

BioPAX : *Biological Pathway Exchange*

C : Cytosine

COSMOS : *COordination of Standards in MetabolomicS*

CNV : *Copy Number Variant*

Da : Daltons

EHMN : *Edinburg Human Metabolic Newtork*

ENCODE *: The Encyclopedia of DNA Elements*

G : Guanine

GC-MS : *Gas Chromatography coupled with Mass Spectrometry*

GWAS : *Genome-Wide Association Study*

HapMap : *Haplotype Map*

HMDB: *Human Metabolome Database*

InChI : IUPAC International Chemical Identifier

KEGG *: Kyoto Encyclopedia of Genes and Genomes*

KGML *: KEGG Markup Language*

LC-MS : *Liquid Chromatography coupled with Mass Spectrometry*

mGWAS : *Genome-Wide Association Study combined with metabolomics*

*MS : Mass Spectrometry*

OTC : Ornithine carbamoyl transférase (OTC)

*REST : Representational State Transfer*

RMN*:* Résonnance Magnétique Nucléaire

RPAIR : *Reactant pair*

SBML : *Systems Biology Markup Language*

SMILES : S*implified Molecular-Input Line-Entry System*

SNP : *Single Nuleotide Polymorphism*

T : Thymine

U : Uracile

XML *: Extensible Markup Language*

# Remerciements

Je tiens tout d'abord à remercier ma directrice, Christine Des Rosiers, pour m'avoir accueillie au sein de son laboratoire à ma première année de baccalauréat et pour avoir partagé avec moi sa passion pour la recherche, ce qui m'a poussé à continuer aux études supérieures. Christine a su à travers ces 4 années me guider dans mon parcours académique, m'encourager à chaque étape et me donner les moyens de mes ambitions. Je suis honorée d'avoir été dirigée par un chercheur aussi passionné, talentueux et enthousiaste. Grâce à elle, j'ai pu présenter mes travaux au Congrès *Metabolomics* à San Francisco, suivre des formations en Californie et en Angleterre et me trouver un poste de doctorante en Suisse. Je garderai un excellent souvenir de ces années passées dans son laboratoire et des valeurs qui m'y ont été inculquées.

Je voudrais remercier mes deux codirecteurs, Guillaume Lettre et John D. Rioux, pour les rencontres productives et pour leurs conseils par rapport à mon projet. Merci à Gabrielle Boucher pour ses conseils statistiques.

J'aimerais également remercier Matthieu Ruiz pour ses réponses à toutes mes questions, scientifiques (ou non), pour les corrections constructives de ce mémoire ainsi que pour le partage amical, à la limite du loufoque, de notre bureau. Merci également à tous les membres du laboratoire pour les discussions et l'agréable compagnie lors de ces années. Particulièrement, merci à Julie Thompson Legault et Anik Forest pour les corrections de ce mémoire.

Merci à toutes les personnes rattachées au programme de Bio-informatique de l'Université de Montréal, qui m'ont beaucoup appris tout au long de mon baccalauréat et de ma maîtrise. Je remercie Elaine Meunier pour ses nombreux conseils administratifs ainsi que tous mes collègues de cours et de l'AEBINUM, particulièrement Armande Ang-Houle et Marc-André Legault, qui ont rendu ces années plus agréables et qui ont été d'un grand soutien scolaire et moral.

De plus, je tiens à manifester ma gratitude à ma famille qui m'a toujours poussé dans mes études et plus spécialement à mes parents qui sont pour moi une source constante de motivation et d'inspiration. J'aimerais tout particulièrement remercier ma mère pour m'avoir aidé de toutes les façons qu'elle pouvait dans mon parcours académique, y compris dans l'amélioration de mon style d'écriture.

J’aimerais aussi mentionner que je n’aurais sûrement pas fini (sainement) cette maîtrise sans le support de ma deuxième famille élargie, mes amis de Marie-De-France. Amélia, Léa, Constance, Ghinwa, Julien, Thomas, Shady et cie., il y a beaucoup de vous dans tout ce que je fais et vous m’incitez au quotidien à me dépasser. Merci de m’avoir écouté me plaindre des aléas de la recherche et de m’avoir supporté dans les bons et mauvais moments.

Finalement, je tiens à remercier Sébastien Lemieux et Gertraud Burger d’avoir accepté de constituer, avec ma directrice et mes 2 codirecteurs, le jury de ce mémoire.

“Strive not to be a success, but rather to be of value”

- Albert Einstein

# Avant-Propos

Ce travail est divisé en 3 chapitres présentant l'introduction du sujet de recherche, la méthodologie et les résultats de mes travaux, sous format d'article, ainsi que la discussion des résultats obtenus et des perspectives futures.

En outre, le présent projet s'inscrit dans une plus grande étude multidisciplinaire dont l'objectif global et à long terme est l'identification d'un patron de biomarqueurs associés à la réponse aux médicaments chez des patients atteints de maladies inflammatoires de l'intestin. Afin de déterminer les principales fonctions biologiques perturbées dans ces maladies, une approche de génomique fonctionnelle a été utilisée sur des cellules pour identifier l'impact biologique de gènes sélectionnés à partir des 163 loci associés au risque de développer ces maladies (Jostins et al., 2012). Cette approche est combinée à des tests cliniques effectués sur les cellules circulantes, le plasma ou le sérum de patients, lesquels incluent : la protéomique, l'immunologie, l'histologie ainsi que la métabolomique. Ces tests cliniques doivent cibler la mesure de marqueurs moléculaires reliés à la fonction des gènes associés à ces maladies. Pour la métabolomique, ceci constituait un défi étant donné qu'il s'avère impossible, à ce jour, de mesurer l'ensemble des métabolites. Par conséquent, il fallait sélectionner un sous-ensemble de métabolites à mesurer en lien avec la fonction des gènes afin de guider les analyses métabolomiques subséquentes.

Dans cette optique, l'objectif de ce projet de maîtrise était de développer une méthode bio-informatique permettant de faire le lien entre gènes et métabolites.

Dans ce mémoire, je me suis plus particulièrement intéressée aux voies métaboliques humaines et plus généralement à l'Homme car l'étude principale porte sur une maladie humaine. Néanmoins, la méthode pourrait être appliquée à n'importe quel organisme ayant des voies métaboliques dans KEGG.

# 1 Introduction

La découverte présentée en 1908 par Archibald Garrod (Garrod, 1923) du concept de maladies innées du métabolisme a grandement contribué à la compréhension du contrôle génétique sur le métabolisme, ce dernier étant défini comme l'ensemble des transformations chimiques essentielles au fonctionnement d'un organisme. Ces maladies causées par la mutation d'un seul gène entraînent le défaut d'une protéine et provoquent ainsi des troubles métaboliques sévères. Depuis les premières recherches par Garrod, de nouvelles erreurs innées du métabolisme sont continuellement découvertes et elles demeurent un axe important en recherche clinique, particulièrement dans le dépistage systématique de ces maladies à la naissance (Schulze et al., 2003; Saudubray, Berghe, & Walter, 2011).

Afin de caractériser ces maladies impliquant le défaut d'un seul gène, il a été nécessaire de faire différentes mesures, du point de vue génétique, protéique ainsi que métabolique. Plusieurs approches permettant d'identifier les variants génétiques et de mesurer les variations au niveau du métabolisme ont été développées afin de mieux comprendre l'étiologie de ces maladies et de proposer des marqueurs afin de les dépister (Miller et al., 2015) et plus généralement, pour étudier l'impact de mutations génétiques dans un organisme.

La section suivante introduira les bases de la théorie de conservation et d'utilisation de l'information génétique ainsi que ses récentes avancées et abordera de manière générale la théorie du dogme central de la biologie moléculaire pour l'ensemble des êtres vivants. Ce dernier permet d'établir le lien entre la génétique et le phénotype et implique le contrôle génétique sur les différentes molécules, soit protéiques et métaboliques, grâce à différents processus biologiques.

## 1.1 Dogme central de la biologie moléculaire

Historiquement, la première version du dogme central de la biologie moléculaire a été décrite par Francis Crick en 1958 et publiée en 1970 (Crick, 1970). Crick a démontré que l'information génétique encodée par l'acide désoxyribonucléique (ADN) est transcrite en acide ribonucléique messager (ARNm), puis l'ARNm est traduit en protéine pour chaque organisme vivant. Le dogme permet d'expliquer le transfert séquentiel de l'information entre gène, transcrit

(ou ARNm) et protéine. Ainsi, il existe trois niveaux de transfert dits « généraux » qui seront abordés pour la suite de ce mémoire.

La transcription réfère au premier processus de l'expression des gènes, où les régions codantes (segments d'ADN) sont copiées en ARNm. L'ARNm est la copie de l'ADN qui utilise les mêmes bases de nucléotides: adénine (A), cytosine (C) et guanine (G). La seule différence est que la thymine (T) de l'ADN est transcrite en uracile (U) pour l'ARNm.

La traduction réfère au processus où la séquence nucléotidique de l'ARNm est traduite en séquence d'acides aminés constituant la protéine. L'ARNm, qui contient 4 bases nucléotidiques, sera utilisé comme modèle pour la traduction en protéine. Cette dernière est constituée d'une chaîne dont la longueur varie en fonction de la séquence d'ARNm. Ainsi, une protéine est une combinaison d'acides aminés, chacun étant le reflet d'un triplet de nucléotides, appelé codon. Le code génétique répertorie l'ensemble des possibilités de ces combinaisons de triplets de nucléotides, formant un codon, et étant associées à un acide aminé donné. Au total, 64 combinaisons ou codons sont possibles, mais seulement 22 acides aminés existent chez tous les organismes vivants (incluant les Archées). Ceci s'explique par le fait que le code génétique est dégénéré, à savoir que plusieurs codons peuvent coder pour un même acide aminé. La traduction est initiée par le codon AUG, qui code pour l'acide aminé méthionine, et est arrêtée par un des 3 codons stops suivants: UAA, UGA et UAG.

### 1.1.1 Avancées du dogme

Depuis que ce dogme fut énoncé en 1958, plusieurs études ont tenté de faire le lien entre ces 3 différents niveaux de transfert, qui incluent les gènes, les transcrits et les protéines. Ainsi, le génotype d'un individu, qui correspond au premier niveau de transfert, est défini par sa constitution génétique (*e.g.*, les différentes formes des gènes d'un individu). Son phénotype, qui résulte des différents niveaux de transfert, est défini comme l'ensemble des caractères observables (Lesk, 2012). Ces derniers incluent les propriétés macroscopiques telles que le poids, la taille, et la couleur des yeux et des cheveux, ainsi que les caractères microscopiques, qui sont observables au niveau cellulaire et moléculaire.

De ces niveaux de transfert que l'on qualifie de « généraux » s'est ajouté un quatrième niveau qui inclue les métabolites, l'ensemble de ces derniers formant le métabolome. En effet, le

phénotype est basé sur les caractéristiques observables, et les dernières observées du point de vue moléculaire sont les métabolites. Les métabolites, qui réfèrent aux molécules organiques de faible poids moléculaire, entre 50 et 15000 daltons (Da), sont transformés dans des réactions catalysées par des enzymes, protéines ayant des propriétés catalytiques. Afin de préciser le lien entre génotype et phénotype, il a été suggéré de faire une description précise du métabolome afin d'étudier la fonction des gènes codant pour des protéines (Fiehn, 2002). Il est important de signaler que contrairement aux gènes, transcrits et protéines, les métabolites ne sont pas directement encodés par les gènes (Baker, 2011).

La Figure 1 (p. 3) représente la nouvelle version simplifiée et linéaire des 3 niveaux généraux de transfert d'information dans la cellule, auxquels ont été ajoutés les métabolites ainsi que le lien entre génotype et phénotype. Dans cette figure, le terme « ome », représente l'ensemble des entités de chaque niveau de transfert. Le génome est défini comme l'ensemble de l'ADN comprenant les gènes, le transcriptome comme l'ensemble des transcrits d'ARNm, le protéome comme l'ensemble des protéines, et le métabolome comme l'ensemble des métabolites d'un organisme.

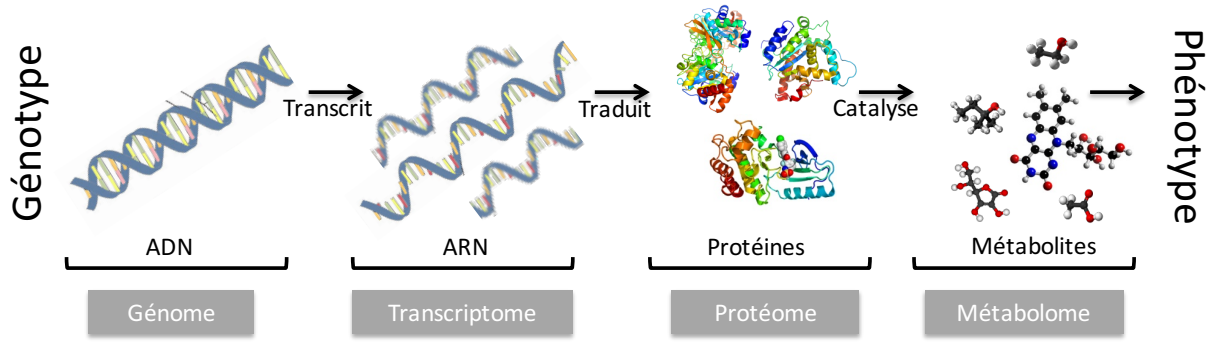


**Figure 1: Le dogme central de la biologie moléculaire revisité.** Dans cette version, le métabolome est intégré aux autres « omes » énoncés par Crick (Crick, 1970). Ici, le lien entre génotype et phénotype est illustré par les 4 niveaux de transfert.

Néanmoins, il est important de noter que l'on connaît maintenant l'existence de multiples boucles de rétroaction partant des métabolites, protéines et/ou transcrits, qui viennent complexifier les transferts illustrés à la Figure 1. Ces boucles démontrent la réelle complexité du transfert d'information, qui est en fait un réseau, et devrait être représenté sous forme de

complexe d'interactions entre différents niveaux de régulation contrôlés de façon dynamique (Goodacre, 2005).

Il a précédemment été suggéré par Teusink et al. (Teusink, Walsh, van Dam, & Westerhoff, 1998) que les métabolites peuvent servir comme mesure de la régulation cellulaire et, plus récemment, comme signature directe de l'activité biochimique, d'où l'importance d'étudier le métabolome (Patti, Yanes, & Siuzdak, 2012).

En outre, l'expression de ces « omes » est propre à l'organisme, au tissu, à la cellule ainsi qu'au compartiment cellulaire étudié. Les conditions environnementales influencent aussi grandement certains « omes », particulièrement le protéome (à plus long terme) et le métabolome (de manière instantanée). De plus, ces conditions peuvent causer des changements au niveau des bases qui composent l'ADN, des protéines qui s'y attachent ou encore de la chromatine, modulant ainsi l'expression des gènes et donc le phénotype observé. Cette discipline, qui est nommée l'épigénétique, est un champ de recherche qui est très actif à l'heure actuelle.

Suite à la découverte du dogme et des avancées technologiques des dernières décennies, de nouveaux champs de recherche ayant pour objectif d'étudier ces « omes » ont été créés. Ces sciences, appelées « omiques », étudient les constituants de ces niveaux de transfert dans leur totalité.

Dans un premier temps, nous nous intéresserons plus particulièrement à la première et la plus mature de ces sciences « omiques », la génomique, qui est l'étude du génome. Ainsi, la section suivante abordera la génomique, ses développements ainsi que ses applications.

## 1.2 Génomique

Suite à de nombreux projets de séquençage de bactériophage et de mitochondries, un projet ambitieux a été créé en 1990 dans l'objectif de découvrir la séquence complète du génome humain. Le projet du génome humain (*Human Genome Project*), dont la version préliminaire de la séquence fut obtenue en 2001 (Venter et al, 2001; Lander et al, 2001), a permis de séquencer les $3\times10^9$ paires de bases d'ADN (Lesk, 2012). Ce projet a fourni des outils puissants aux chercheurs afin de mieux comprendre les facteurs génétiques influençant les maladies humaines, ouvrant la voie à de nouvelles stratégies pour le diagnostic, le traitement et la prévention des maladies.

### 1.2.1 Variants génétiques et génome

Les variations dans le génome sont dues à des modifications de l'information génétique appelées mutations. Celles-ci peuvent survenir au cours de la réplication de l'ADN ou de la division cellulaire. Cependant, des mécanismes de contrôle efficaces de réparation de l'ADN corrigent la très grande majorité de ces erreurs. Il est aussi possible que d'autres facteurs puissent engendrer ces mutations, par exemple l'exposition à des agents mutagènes ou à des virus présents dans l'environnement. Certaines de ces mutations sont transmises à la génération suivante, tandis que d'autres se produisent dans un individu d'une génération et sont donc présentes pour la première fois chez une famille.

De façon générale, si ces mutations sont transmises, le site de ces mutations est alors appelé site polymorphe car il contient différents nucléotides dans une population (Nachman, 2001). Il existe plusieurs types de mutations qui peuvent avoir lieu dans la partie codante ou non-codante du génome. La partie codante est présente dans l'ARNm mature. Initialement, l'ARNm contient les séquences codantes, nommées exons, et les séquences non-codantes, nommées introns, qui seront par la suite épissés. Les mutations dans la région codante peuvent modifier les protéines encodées par ces gènes, tandis que les mutations dans les régions non-codantes ne modifieront pas directement la protéine, mais peuvent avoir un impact sur la régulation de la transcription. Les mutations dans la région codante n'affectant pas les acides aminés de la protéine résultante, en conséquence du code génétique dégénéré, sont dites « synonymes » tandis que celles affectant l'acide aminé sont dites « non-synonymes ». De ces mutations non-synonymes, il en existe deux types, soit (i) les mutations faux-sens, qui impliquent la substitution d'un acide aminé par un autre, et (ii) les mutations de perte de fonction, qui incluent les mutations non-sens changeant l'acide aminé par un codon stop, produisant ainsi une protéine tronquée. En outre, il existe d'autres types de variants génétiques ayant un impact sur la protéine résultante, comme les insertions et délétions de quelques nucléotides.

Un variant génétique qui n'affecte qu'un seul nucléotide et qui est fréquent dans une population (soit plus de 1%) est appelé un polymorphisme d'un seul nucléotide, ou plus communément *Single Nuleotide Polymorphism* (SNP) (Willard, Ginsburg, & Geoffrey S. Ginsburg, 2010). D'ailleurs, la distribution de ces SNPs n'est pas homogène sur le génome humain; elle varie selon le taux de mutation, qui est la mesure de la vitesse d'apparition des

mutations dans le temps, et la position sur le génome. Les SNPs peuvent survenir, par exemple, de façon plus fréquente dans les régions non-codantes que dans les régions codantes. De façon générale, la sélection naturelle agira de façon à éliminer les SNPs qui seront génétiquement défavorables et vice-versa (Barreiro, Laval, Quach, Patin, & Quintana-Murci, 2008).

Le projet du génome humain a permis de découvrir qu'il existe environ 20 000 gènes codant pour des protéines, ce qui représente seulement 2% du génome entier, ce qui explique le plus grand nombre de SNPs dans les régions non-codantes car ces dernières représentent 98% du génome. Pour cette raison, des projets tels que *The Encyclopedia of DNA Elements* - ENCODE (Bernstein et al., 2012) ont été créés dans l'objectif d'identifier les éléments fonctionnels du génome dans les régions non-codantes. Par exemple, ce projet a mis en évidence qu'il y a un enrichissement de SNPs associés à des pathologies dans des régions non-codantes, lesquelles peuvent contrôler l'expression des gènes.

## 1.2.2 Plateformes d'analyse

Depuis le premier séquençage du génome humain en 2001, qui s'est étalé sur plus de 10 ans, l'émergence de nouvelles technologies a rendu le séquençage moins coûteux et excessivement plus rapide. Ainsi, les facteurs limitants ne sont plus le temps de séquençage mais plutôt l'analyse des grandes quantités de données provenant des génomes séquencés. Comparativement au séquençage qui lit la sequence complète, le génotypage permet de révéler le génotype d'un individu, soit ses différents allèles qui sont définis comme une des formes alternatives d'un gène ou d'une position sur la séquence (locus génétique).

### 1.2.2.1 **Génotypage par puce**

Les méthodes actuelles de génotypage permettent d'obtenir une couverture rapide et de grande taille de variants connus du génome, et ce, à moindre coût. Suite au projet du génome humain, d'autres projets, tels que *Haplotype Map* – HapMap (International HapMap Consortium, 2003), ont vu le jour. Ce dernier avait pour objectif d'identifier les variations génétiques les plus fréquentes et les positions polymorphes du génome humain. HapMap a permis de fournir de l'information génétique de plusieurs populations européennes, africaines et asiatiques afin d'aider dans la conception (design) des études subséquentes et dans l'interprétation des données de génotypage à grande échelle. Plusieurs techniques existent pour le génotypage par puce, et

celles-ci sont capables d'analyser plusieurs centaines de milliers de SNPs, voir millions. Ces techniques utilisent les propriétés d'hybridation de l'ADN, et se distinguent par leurs méthodes de détection du polymorphisme (Ragoussis, 2009). Pour une position donnée, chaque puce a une séquence d'ADN particulière qui permet d'identifier le génotype. Ce génotypage par puces permet d'identifier des génotypes de façon robuste, rapide et à moindre coût. Néanmoins, cette méthode ne couvre pas l'ensemble des polymorphismes d'un génome; elle identifiera surtout les variants communs déjà connus dans la population, et il sera difficile de détecter des variants plus rares ou nouveaux.

#### 1.2.2.2 **Séquençage**

Plusieurs technologies émergentes permettent de fournir à la génomique la capacité d'identifier la séquence d'ADN, et contrairement au génotypage par puce, donnent accès à la séquence complète et non à des variants individuels. Cette section présentera les deux technologies les plus couramment utilisées, et mettra de l'avant leurs avantages et inconvénients respectifs.

*Sanger*

Le premier génome séquencé fut celui du bactériophage ΦX-174, virus à simple brin d'ADN, par Frederick Sanger en 1977 (Sanger et al., 1977). Celui-ci a développé l'une des premières méthodes de séquençage, nommée Sanger, ce qui a permis une avancée majeure en génétique. Cette méthode est d'ailleurs encore utilisée à ce jour. Cette technologie est considérée comme étant le séquençage de première génération (Metzker, 2010). Brièvement, le séquençage selon Sanger consiste à déterminer précisément l'ordre des nucléotides dans une séquence d'ADN donnée lors de l'élongation de son brin complémentaire (Sanger & Coulson, 1975). Une des premières méthodes de détection utilisait l'ajout de traceurs radioactifs, mais de nos jours sont utilisés des fluorophores, petits composés chimiques dégageant des lumières colorées selon le nucléotide incorporé, et permettent d'effectuer le séquençage en une seule réaction. La couleur émise indique le nucléotide de chaque base du fragment d'ADN.

Le séquençage Sanger est encore utilisé afin de valider la séquence d'une petite région d'un génome d'intérêt ou pour confirmer des observations identifiées dans des études à haut débit. Cette technologie a pour principal inconvénient de séquencer seulement de 300 à 1000

nucléotides. Malgré ses limitations techniques, cette méthode est considérée comme la plus fiable étant donné son faible taux d'erreurs.

*Séquençage de nouvelle génération*

L'avènement de nouvelles technologies de séquençage, nommées de seconde ou nouvelle génération, a permis de réduire considérablement le prix et d'augmenter la vitesse du séquençage du génome complet. La différence entre le séquençage de Sanger et de nouvelle génération réside dans la capacité à séquencer un grand nombre de séquences simultanément. Le séquençage de Sanger se limite à une centaine de séquences lues en parallèle, comparativement aux plusieurs millions par les technologies de nouvelle génération qui offrent une analyse à haut débit (Mardis, 2008). Le séquençage de nouvelle génération s'effectue par des méthodes similaires à la méthode de Sanger mais sur des petits fragments d'ADN. Ainsi sont produits des millions de petites séquences, appelées « courtes lectures » ou *short reads.*

Depuis les dernières années, les récentes technologies ont permis une diminution du prix du séquençage et donc une augmentation du nombre de génomes séquencés. Ainsi, étant donné la quantité faramineuse de petites lectures générées par ces séquenceurs, il a été nécessaire de développer des outils computationnels afin de traiter ces données.

### 1.2.3 Études d'association pangénomique

L'approche des études d'association pangénomique permet de tester l'association entre des variants génétiques et des phénotypes. Les phénotypes étudiés peuvent ainsi être des maladies ou des traits mesurables, comme la taille de l'organisme étudié par exemple. Un trait est un caractère d'un organisme qui est influencé par des gènes et/ou par l'environnement. Ces études utilisent plus communément le génotypage à haut débit afin d'identifier des centaines de milliers, voire des millions de variants pour les associer au phénotype étudié (Pearson & Manolio, 2008).

De plus, ces études tirent avantage de la répartition non aléatoire des SNPs sur le génome. En étudiant les variants génétiques chez différents individus, il a été montré qu'il existait des corrélations, soit des relations, entre les variants qui se situaient dans des régions rapprochées. Il est connu que certaines régions du génome regroupant plusieurs variants sont transmises de façon intacte. Ainsi, le génotype d'une position peut permettre d'inférer (imputer) le génotype

des positions avoisinantes. Ce principe est dû au déséquilibre de liaison (Hirschhorn & Daly, 2005), qui crée des motifs de corrélation entre variants. Ces motifs sont utilisés lors d'approche d'association pangénomique afin d'obtenir une meilleure couverture des variants du génome sans avoir à tous les génotyper.

En outre, ces études nécessitent un très grand nombre d'individus afin d'obtenir des variants génétiques significativement associés. En effet, le principal enjeu de ces études consiste à obtenir des associations franchissant des seuils de significativité qui sont corrigés par rapport aux multiples tests statistiques effectués (pour l'ensemble des variants testés). Afin que les valeurs d'association franchissent ces seuils très faibles, il est nécessaire d'analyser plusieurs milliers d'individus. De plus, des réplications de ces études dans d'autres populations distinctes sont aussi nécessaires afin de confirmer les associations obtenues.

Néanmoins, étant donné le déséquilibre de liaison, il est difficile de confirmer le variant causal d'une région associée en utilisant seulement les résultats d'études d'association. Il est ainsi complexe de savoir exactement quel gène et quel variant génétique ont un impact direct sur le phénotype. Des études fonctionnelles sont nécessaires afin de raffiner la position et la causalité du variant associé. Même si ces études d'association pangénomique ont permis d'identifier de nombreux loci associés, les variants rapportés ne permettent pas de prédire entièrement ces phénotypes (McClellan & King, 2010).

### 1.2.4 Applications de la génomique

De façon historique, les scientifiques ont tenté d'expliquer le lien entre le phénomène d'hérédité et les traits physiques. Ces traits peuvent être des observations anatomiques, morphologiques et/ou moléculaires chez un organisme vivant. Certains traits peuvent avoir un caractère pathologique. Cette section présentera les traits monogéniques et complexes, qui sont les traits les plus communément étudiés dans les études cliniques.

#### 1.2.4.1 **Traits monogéniques**

Dans les cas les plus simples, un trait est dû à la variation d'un seul gène, appelé alors trait monogénique. On parle de caractères à transmission mendélienne, car les principes de transmission sont conformes aux lois d'hérédité de Mendel. Cela implique que la variation d'un seul gène est suffisante pour entraîner une variation observable du phénotype. Chez l'humain, les

caractères mendéliens les plus étudiés sont les maladies monogéniques, telles les erreurs innées du métabolisme présentées plus tôt. Généralement, les maladies à transmission mendélienne ont des manifestations cliniques importantes, mais n'affectent qu'un faible pourcentage de la population. Un autre exemple de maladie monogénique est la fibrose kystique, qui découle de la mutation du gène *CFTR* codant pour la protéine du même nom (*CFTR – cystic fibrosis transmembrane conductance regulator)*. Pour étudier ce type de maladies, des familles multi-générationnelles comportant des individus affectés et non affectés ont été recrutées afin d'identifier la mutation impliquée. Ceci a été le cas pour la découverte de la mutation la plus fréquente pour la fibrose kystique (Kerem et al., 1989). De plus, le gène *CFTR* entraînant cette pathologie fut le premier à être identifié par le projet du génome humain (Tolstoi & Smith, 1999). La découverte des différentes mutations de ce gène a permis une meilleure compréhension de la base génétique de ce défaut ainsi qu'une meilleure compréhension de la physiopathologie qui en découle.

#### 1.2.4.2 **Traits complexes**

Les traits complexes résultent d'un ensemble de facteurs de risques, lesquels incluent des variations génétiques impactant plusieurs gènes et des facteurs environnementaux. Ces traits complexes peuvent se refléter au niveau du phénotype macroscopique, comme la taille d'un individu, ainsi que dans le développement de maladies dites « complexes » car elles sont causées par l'effet de plusieurs variants de gènes en combinaison avec le mode de vie et l'environnement d'un individu. De façon générale, ces traits sont retrouvés de façon plus fréquente dans la population comparativement aux traits monogéniques. Comme exemple de maladie complexe, on retrouve les maladies cardiaques, les cancers ainsi que les maladies inflammatoires de l'intestin. Pour les maladies inflammatoires de l'intestin, plusieurs études génomiques se sont intéressées à trouver des variants génétiques contribuant à une portion de la variance totale de ces traits complexes. Par exemple, une récente étude d'association pangénomique a utilisé le génotypage par puce et l'imputation de variants (Jostins et al., 2012) chez des patients atteints de maladies inflammatoires de l'intestin et chez des témoins sains (n = 40 660). Cela a permis la découverte de 163 loci associés à cette pathologie. Néanmoins, ces 163 loci n'expliquent qu'une partie de la susceptibilité d'être atteint d'une maladie inflammatoire de l'intestin (de 4.1% à 13.6% ) (Jostins et al., 2012).

Néanmoins, il s'avère que les variants génétiques identifiés par ces études n'expliquent pas l'ensemble du trait complexe en question. En conséquence, les autres sciences « omiques » ont pris de l'importance pour améliorer nos connaissances sur l'explication des traits complexes, la dernière en date étant la métabolomique. Cette dernière est la plus récente des sciences « omiques » et a comme but d'étudier les métabolites, qui peuvent être des marqueurs moléculaires plus proches du phénotype observé et représentent donc un atout considérable pour la caractérisation des traits complexes. Dans l'objectif de mieux comprendre le lien entre génotype et phénotype, les sections subséquentes présenteront un sujet clé de ce mémoire, qui est de complémenter la génomique par la métabolomique.

## 1.3 Métabolomique

Historiquement, la métabolomique a été proposée comme outil pour étudier l'impact de la délétion de gènes chez l'Homme(Raamsdonk et al., 2001). Les premières études de ce domaine, nommées à l'époque études d'empreinte ou de profilage métabolique (Westall, 1960; Allan, Cusworth, Dent, & Wilson, 1958), mesuraient seulement quelques métabolites. Elles permettaient d'identifier l'impact du défaut d'un gène sur les niveaux de ces métabolites dans le contexte des maladies innées du métabolisme, et servaient ainsi d'outil diagnostique. Outre l'impact de la génétique sur les niveaux de métabolites dans ce type de maladie, il a été précédemment démontré que des perturbations enzymatiques (et donc protéiques) pouvaient également influencer directement le métabolome intracellulaire (Ewald, Matt, & Zamboni, 2013).

### 1.3.1 Métabolites et métabolome

Le métabolome, dernier niveau de régulation après la transcription et la traduction (Figure 1, p. 3), est plus proche de la fonction et est souvent utilisé comme reflet du phénotype d'un individu. L'ensemble des métabolites retrouvés dans un échantillon biologique à un instant donné est appelé le métabolome (Oliver, 1998). On estime le nombre de métabolites présents dans le métabolome humain entre $10^4$ et $10^5$ (Liesenfeld, Habermann, Owen, Scalbert, & Ulrich, 2013). Les métabolites réfèrent à des molécules endogènes, *i.e.* qui sont impliquées ou résultent du métabolisme primaire, et aux métabolites exogènes, qui ne sont pas produits naturellement dans l'organisme (Kaddurah-Daouk, Kristal, & Weinshilboum, 2008). Les métabolites primaires

sont impliqués dans les processus physiologiques fondamentaux tels que la croissance, le développement et la reproduction. Ils sont divisés en différentes classes lesquelles incluent : les glucides, les lipides, les acides aminés et les acides nucléiques.

Chez l'Homme, afin de mieux explorer son métabolome, des bases de données telles que *Human Metabolome Database* (HMDB) (Wishart et al., 2007) ont été créées dans l'objectif de répertorier tous les métabolites mesurés en y ajoutant de l'information uniformisée et dérivée de la littérature. De plus, la dernière version la plus à jour de cette base de données, HMDB 3.0 (Wishart et al., 2013), fait la distinction entre métabolites mesurés et métabolites attendus, qui sont des molécules dont on connaît le rôle dans des réactions du métabolisme même s'ils n'ont pas encore été mesurés. Cette ressource, contenant près de 42 000 métabolites (http://www.hmdb.ca/), est dédiée à fournir aux scientifiques la couverture la plus actuelle et complète du métabolome humain.

Comme la génomique pour le génome, la métabolomique est l'étude de l'ensemble des métabolites, le métabolome, dans un système biologique (Mamas, Dunn, Neyses, & Goodacre, 2011; German, Hammock, & Watkins, 2005). La métabolomique permet l'étude complète, qualitative et quantitative, des métabolites d'un échantillon biologique grâce à des techniques analytiques de plus en plus performantes (Oliver, 1998; Fiehn, 2002).

Les métabolites sont des molécules aux propriétés physico-chimiques diverses. Des exemples sont illustrés à la Figure 2 (p. 13). La diversité de ces molécules explique ainsi la complexité des techniques analytiques permettant de mesurer l'ensemble de ces molécules. De plus, comparativement aux autres sciences « omiques », la métabolomique est confrontée au problème d'identification des métabolites mesurés (Saghatelian & Cravatt, 2005). Comparativement à la génomique qui identifie de très longues séquences de 4 acides nucléiques connus et à la protéomique qui identifie des millions de protéines dont on connaît les 22 acides aminés qui les composent, la métabolomique mesure des milliers de métabolites ayant chacun une structure différente. Néanmoins, à ce jour, bien que des milliers de metabolites peuvent être mesurés lors d'analyses métabolomiques, ceci représente mois de 5% du metabolome, et de plus une fraction encore plus faible est identifiable et connue.

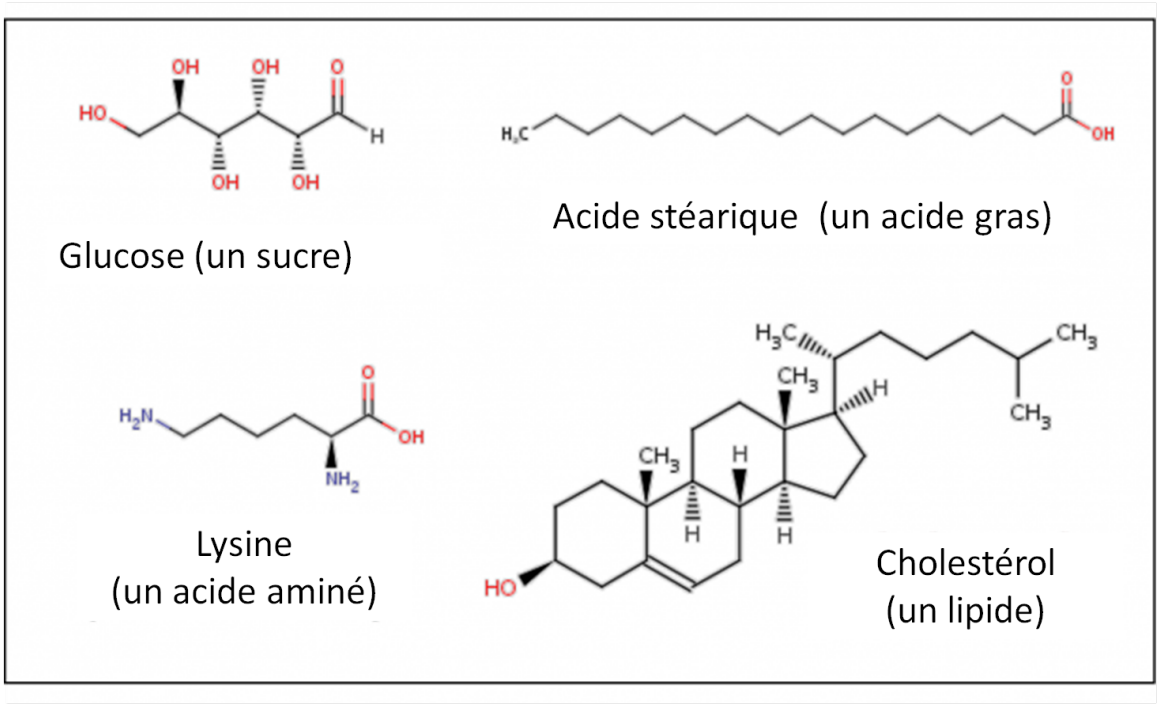


**Figure 2: Exemples de métabolites.** La classe des différents métabolites est indiquée entre parenthèses. Adapté de « Introduction to Metabolomics » par *European Bioinformatic Institute* (http://www.ebi.ac.uk/training/online/course/introduction-metabolomics/).

Une expérience de métabolomique exige une importante planification du design expérimental. Cette expérience peut être divisée en 5 grandes étapes illustrées à la Figure 3 (p. 14), soit la collecte des échantillons, la préparation des échantillons, l'acquisition des données, le traitement des données, l'analyse des données et leur interprétation.

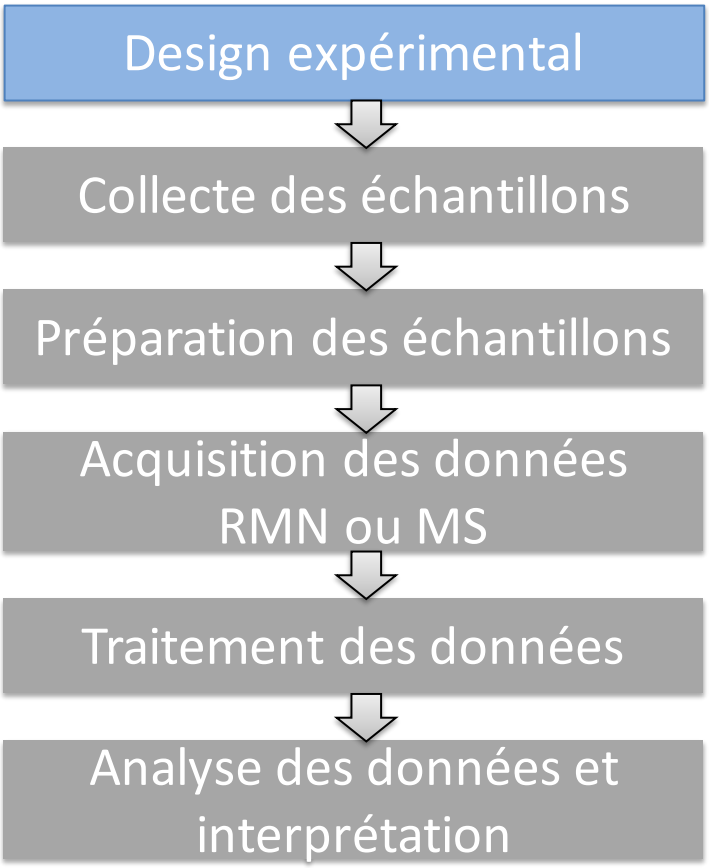


**Figure 3: Les étapes à suivre lors d'une étude métabolomique.** La conception de l'étude (en bleu) précède et englobe toutes les étapes subséquentes reliées plus directement à l'analyse métabolomique (en gris).

### 1.3.2 Design expérimental

Le design expérimental représente la planification d'une étude qui peut, pour celle métabolomique, se diviser en 5 étapes (Figure 3, p. 14): de la collecte des échantillons jusqu'à l'interprétation des données, afin de répondre à une question biologique donnée. Pour ce faire, plusieurs facteurs doivent être pris en compte pour réduire les variations externes et expérimentales afin de révéler uniquement les variations biologiques. Ces facteurs incluent de manière non exhaustive l'uniformité de la méthode de collecte, la méthode de conservation des échantillons, le nombre d'échantillons analysés, la randomisation (définie comme l'analyse aléatoire des échantillons de différentes conditions), les méthodes de préparation d'échantillons, les variations techniques dues aux appareils utilisés, les choix de traitement de données, les

contrôles de qualité pour plusieurs de ces étapes (Sugimoto, Kawakami, Robert, Soga, & Tomita, 2012), etc.

À titre d'exemple, la stabilité des métabolites dans des liquides biologiques peut être un facteur de variation externe. Ainsi, lors de la collecte d'échantillons d'urine, il a été proposé de mettre immédiatement les échantillons au froid (à 4° C) afin d'éviter la dégradation des métabolites et la croissance bactérienne (Roux, Thévenot, Seguin, Olivier, & Junot, 2014), puis de les conserver à plus long terme à -80 °C.

#### 1.3.2.1 **Approches**

Selon la question biologique posée, deux différentes approches peuvent être utilisées lors de l'analyse métabolomique, l'approche « ciblée » et « non-ciblée ». Dans certains cas, lorsqu'une hypothèse est définie et qu'on veut la tester, il est souhaitable de définir un ensemble précis de métabolites à mesurer. Pour ce faire, il faut utiliser une approche dite « ciblée » qui quantifiera de façon absolue et quantitative (concentration) ou relative un ensemble de métabolites sélectionnés. Dans d'autres cas, lorsque l'objectif est de générer de nouvelles hypothèses, une approche dite « non-ciblée » peut être utilisée afin de mesurer le plus de métabolites, favorisant ici la découverte de nouvelles molécules. Cette dernière méthode permet une quantification relative des métabolites observés en comparant les résultats obtenus selon les conditions étudiées. De fait, le choix de l'approche est un attribut définissant l'expérience métabolomique (Patti et al., 2012).

### 1.3.3 Plateformes analytiques

Il existe différents instruments permettant d'identifier et de quantifier les métabolites d'un échantillon biologique. Néanmoins, malgré de récentes avancées technologiques importantes, le métabolome en son entier ne peut pas à ce jour être mesuré à l'aide d'une seule plateforme (Vorkas et al., 2015). Chaque plateforme ayant des spécificités différentes, celles-ci peuvent être utilisées de façon complémentaire afin d'élargir le nombre de métabolites mesurables. Les deux principales technologies analytiques utilisées en métabolomique sont la spectrométrie de masse (nommée MS pour *mass spectrometry*) et la résonnance magnétique nucléaire (nommée RMN) (Dettmer, Aronov, & Hammock, 2007).

#### 1.3.3.1 **Spectrométrie de masse (MS)**

De manière globale, la MS est une technologie qui permet d'analyser, grâce à un analyseur de masse, la masse d'une molécule préalablement ionisée. Cette technique est souvent couplée à des méthodes de séparation par chromatographie. Cette dernière permet de séparer les molécules d'un mélange complexe selon leurs propriétés physico-chimiques (Dettmer et al., 2007). Une des approches est la chromatographie gazeuse couplée à la spectrométrie de masse (nommée GC-MS).

Pour la chromatographie en phase gazeuse, l'échantillon doit le plus souvent être d'abord traité avec un agent chimique afin de dériver les molécules et les rendre ainsi plus volatiles et compatibles avec l'analyse sur un GC. L'échantillon est ensuite chauffé et vaporisé grâce à un gaz porteur dans une colonne. Les molécules seront séparées dans cette colonne en fonction de leur affinité pour celle-ci. Une seconde approche est la chromatographie liquide couplée à la spectrométrie de masse (nommée LC-MS). En chromatographie liquide, la méthode est similaire à la chromatographie en phase gazeuse, à la différence que les molécules peuvent être analysées sans dérivation et l'échantillon n'est donc pas chauffé, mais mélangé à un liquide permettant son transport dans une colonne.

Les deux approches de chromatographie permettront donc la séparation des métabolites selon leur temps de rétention. Le couplage au spectromètre de masse permet d'obtenir le spectre caractéristique de l'échantillon selon trois axes : le ratio masse/charge (d'où l'on peut dériver la masse), le temps de rétention qui dépend notamment du ratio masse/charge (d'où l'on peut dériver la masse) et l'intensité du signal obtenu, qui représente l'abondance du métabolite.

La GC-MS est reconnue pour sa robustesse (Liesenfeld et al., 2013) et elle est plus largement utilisée pour quantifier les métabolites étant donné son coût relativement abordable (Kaddurah-Daouk et al., 2008). Les métabolites mesurés par GC-MS sont principalement volatiles, moins polaires et de faible masse moléculaire. Une limitation de la GC-MS est qu'elle est moins adaptée aux composés non volatiles. En comparaison, la LC-MS offre plusieurs avantages sur la GC-MS, telle qu'une plus grande couverture de métabolites de différentes masses moléculaires (Cacciatore & Loda, 2015), et elle est plus adaptée pour les composés polaires et non volatiles. La variété et le nombre de composés analysables par la LC-MS peuvent être considérés comme un avantage pour l'analyse non-ciblée.

#### 1.3.3.2 Résonnance Magnétique Nucléaire (RMN)

La RMN est une technique qui utilise les propriétés magnétiques des noyaux atomiques des molécules. Cette technique donne des renseignements sur la structure de la molécule grâce aux spectres de résonance des atomes soumis à un champ magnétique. Cette méthode est aussi utilisée afin d'identifier la structure de macro-molécules telles que les protéines, mais est plus précise pour les petites molécules telles que les métabolites car les spectres sont uniques, bien résolus et hautement prévisibles (Lindon, Nicholson, Holmes, & Everett, 2000). La détermination de la structure des métabolites connus en utilisant diverses méthodes de RMN est simple et directe tandis que l'analyse *de novo* de structures imprévues ou encore de métabolites inconnus l'est moins, mais tout de même faisable (Rolin, 2012).

#### 1.3.3.3 Avantages et inconvénients de la MS et RMN

Un avantage de la MS, comparativement à la RMN, est qu'elle nécessite très peu d'échantillons. La MS est une technique plus sensible et permet une couverture métabolique plus grande comparativement à la RMN (Mal, Koh, Cheah, & Chan, 2012). Toutefois, l'avantage de la RMN est qu'elle ne nécessite pas ou très peu de préparation d'échantillon et donc est plus rapide. De plus, elle est non destructive et se prête à plus facilement à des études sur des tissus intacts ou des organismes complets (Reo, 2002). L'inconvénient majeur de la RMN est sa sensibilité relativement faible.

Le choix d'instruments utilisés pour une analyse métabolique dépend de l'objectif de la question posée, chacun ayant des limitations spécifiques. Étant donné la complexité des données générées, des approches bio-informatiques et statistiques puissantes ont été développées afin de permettre l'analyse de ces données.

### 1.3.4 Applications

La nature non invasive, par l'analyse de liquide biologique, de la métabolomique et son lien étroit avec le phénotype rendent cette discipline idéale en médecine préventive, en pharmaceutique et pour les industries alimentaires. Ainsi, l'information métabolomique est de grande valeur, car les réactions métaboliques reflètent la fonction de la cellule. La section suivante présentera donc quelques exemples des applications de la métabolomique, en particulier la recherche de biomarqueurs et la médecine personnalisée.

#### 1.3.4.1 **Recherche de biomarqueurs**

La métabolomique a surtout été appliquée à la recherche de biomarqueurs, caractéristique biologique qui est mesurée objectivement et évaluée comme un indicateur biologique, liés au diagnostic d'un phénotype. En effet, ces biomarqueurs aident à une meilleure compréhension des processus biologiques, physiologiques et pathologiques d'un phénotype.

Ces biomarqueurs sont des métabolites qui peuvent être utilisés afin de distinguer deux groupes, typiquement associés au phénotype malade ou sain. Par exemple, un métabolite différentiellement exprimé dans des échantillons de patients malades comparativement à sujets sains, sera classifié comme biomarqueur de la pathologie en question. Chez l'Homme, des échantillons biologiques tels que l'urine, la salive, le sang, le fluide séminal ou spinal, peuvent être utilisés pour la découverte de ces biomarqueurs.

Néanmoins, dans la majorité des cas, il est rare qu'un seul métabolite soit caractéristique d'un état pathologique; on considérera donc plutôt une signature de plusieurs métabolites (Kaddurah-Daouk et al., 2008). Par exemple, dans une récente étude ayant pour objectif l'identification de biomarqueurs chez des patients ayant une maladie rénale chronique, un profil de 8 métabolites détectés dans l'urine des patients atteints, soit le 5-oxoproline, le glutamate, le guanidoacétate, l'α-phenylacétylglutamine, la taurine, le citrate et le triméthylamine N-oxide, a été associé au développement de cette maladie (Posada-Ayala et al., 2014).

#### 1.3.4.2 **Médecine personnalisée**

La médecine personnalisée est la personnalisation des soins de santé à l'échelle de l'individu et non de la population. Celle-ci a grandement utilisé la génomique, mais a aussi plus récemment été aidée par l'introduction de biomarqueurs métabolomiques dans le diagnostic rapide de maladies. En santé, il existe actuellement des tests biochimiques qui mesurent des concentrations de métabolites afin d'identifier les différents stades de la maladie. Un exemple connu est le diabète, où l'on mesure les niveaux du métabolite glucose dans le sang (Valeri, Pozzilli, & Leslie, 2004). La métabolomique offre le potentiel d'identifier de façon rapide des centaines de métabolites, marqueurs des différents stades précoces de maladie. À l'avenir, avec l'avenue de la médecine personnalisée, il sera possible d'étudier les variations dans notre propre métabolome afin de personnaliser le type de médication ainsi que les doses utilisées et

d'améliorer les stratégies de traitement. En effet, un facteur majeur de la réponse interindividuelle à des médications est dû à des variations du phénotype métabolique, qui est influencé en partie par le génotype mais aussi par des facteurs environnementaux tels que le statut nutritionnel, l'âge et la co-administration d'autres médicaments. L'étude de profils métaboliques permettant les traitements personnalisés risque ainsi d'être plus efficace que les procédures actuellement utilisées, qui sont basées sur les populations.

Un exemple de cette application concerne l'utilisation de médicaments tels que les statines, qui sont utilisés pour réduire le cholestérol LDL (lipoprotéines de basse densité), afin de traiter certaines maladies cardiovasculaires. En premier lieu, plusieurs variants génétiques contribuant à la variabilité du cholestérol LDL en réponse aux statines avaient été identifiés (Mangravite, Wilke, Zhang, & Krauss, 2008). Toutefois, seulement une petite proportion de la variance était expliquée par ces facteurs génétiques (Trupp et al., 2012). Subséquemment, une étude de pharmacométabolomique (Kaddurah-Daouk et al., 2011) a démontré que certains métabolites, dont trois acides biliaires, contribuaient à prédire l'ampleur de la réduction du cholestérol LDL par les statines. Ces métabolites sont en fait les produits de réactions qui, dans leur ensemble, constituent des voies essentielles pour le fonctionnement cellulaire.

La prochaine section présentera ces voies métaboliques qui font le lien entre les deux derniers niveaux de transfert, soit le protéome et le métabolome.

## 1.4 Voies métaboliques

Une voie métabolique est une série de réactions chimiques se déroulant dans une cellule. Ces réactions constituent la dernière interaction entre protéines et métabolites présentée dans les avancées du dogme central (Figure 1, p. 3). Les métabolites et les protéines ayant une fonction enzymatique sont les principaux acteurs de ces voies ; les métabolites sont modifiés dans des séries de réactions catalysées par ces enzymes. *Escherichia coli* fut l'un des organismes ayant permis de définir les premières cartes globales métaboliques (Karp, Riley, Paley, Pellegrini-Toole, & Krummenacker, 1997). La définition et la taille (le nombre de réactions) des voies métaboliques peuvent grandement varier selon l'utilité voulue de leurs créateurs. En outre, la métabolomique vise à mettre en évidence les changements dans les réseaux et les voies métaboliques selon certaines conditions. Ces voies ont permis de mieux comprendre le

fonctionnement cellulaire, le métabolisme énergétique cellulaire, les voies de dégradation des métabolites et de façon générale, d'expliquer les phénomènes permettant l'homéostasie cellulaire, définie comme processus physiologique maintenant de façon constante l'équilibre du milieu intérieur, malgré les variations du milieu extérieur (Richards et al., 2010) .

L'annotation de voies métaboliques fait partie d'un effort de compilation des connaissances génomiques et post-génomiques, soit transcriptomiques, protéomiques ainsi que métabolomiques.

### 1.4.1 Réactions enzymatiques

Les réactions enzymatiques font le lien entre les métabolites qui sont leurss produits et substrats. Ces réactions, catalysées par les enzymes, sont des mécanismes chimiques constituant le fonctionnement du métabolisme. Les enzymes sont classifiées après validation expérimentale grâce à des numéros appelés *Enzyme Commission number* (EC) selon les réactions qu'elles catalysent (Tipton & Boyce, 2000). Comme système de nomenclature, chaque numéro EC est associé à un nom spécifique pour son enzyme respective. Ces numéros EC permettent de regrouper des classes de réactions ayant la même fonction. En outre, l'ensemble des réactions inconnues d'un organisme et leur caractérisation est le sujet de nombreuses recherches. L'information manquante de ces réactions expliquerait la grande quantité de métabolites non attendus dans les analyses métabolomiques (Fiehn, Barupal, & Kind, 2011).

Les voies métaboliques sont donc une série de réactions enzymatiques où le produit d'une réaction devient le substrat de la réaction suivante.

### 1.4.2 Bases de données de voies métaboliques

Il existe une multitude de bases de données permettant la description des voies métaboliques de plusieurs organismes, chacune ayant des définitions différentes. La section suivante présente les deux plus grandes bases de données organisées et gérées manuellement (Altman, Travers, Kothari, Caspi, & Karp, 2013).

#### 1.4.2.1 **KEGG**

*Kyoto Encyclopedia of Genes and Genomes* (KEGG; (Ogata et al., 1999)) est l'une des bases de données bio-informatiques largement utilisées pour ses voies métaboliques, de

signalisation et des processus cellulaires selon différents organismes. KEGG rassemble les informations concernant la séquence génomique de ces organismes, les gènes annotés, les protéines telles que les enzymes associées aux réactions, ainsi que les métabolites. Plus particulièrement, pour les voies métaboliques, KEGG illustre un ensemble de voies de référence où sont projetées les réactions spécifiques de chaque organisme. KEGG est donc construit à partir de voies métaboliques multi-organismes. Afin de rendre ces voies spécifiques à un organisme, KEGG vérifie la présence des gènes par homologie. Ces gènes sont ensuite cartographiés sur les réactions des voies multi-organismes, rendant ainsi ces voies métaboliques spécifiques à l'organisme. L'information relative aux voies métaboliques est représentée selon trois niveaux de résolution : (i) les cartes métaboliques telles que la carte globale, (ii) les voies métaboliques spécifiques à une voie, comme le cycle de Krebs ou la glycolyse, et (iii) les modules spécifiques à des voies métaboliques données, lesquels sont séparés en plus petites unités fonctionnelles (Minoru Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012). À titre d'exemple, les cartes métaboliques (Figure 4, p. 21) sont définies par KEGG comme une vue d'ensemble qui inclut plusieurs voies métaboliques regroupées. À sa dernière mise à jour, KEGG contenait 4 007 organismes, 9 910 réactions, 17 448 métabolites et 476 voies métaboliques. Cette base de données comprend des outils de visualisation permettant une superposition des voies métaboliques spécifiques à un organisme. KEGG est annotée de façon manuelle et mise à jour régulièrement par ses créateurs.

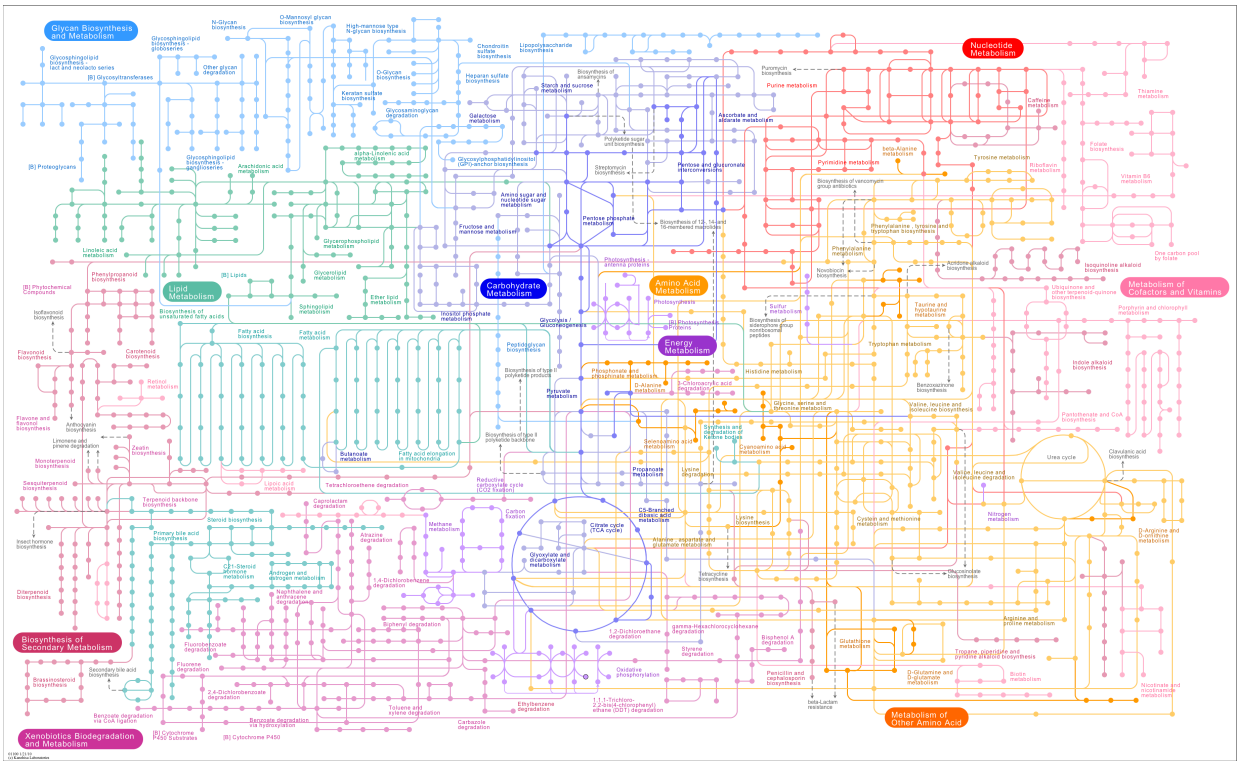


**Figure 4: Carte globale multi-organisme de KEGG intitulée *Metabolic Pathways*.** Tirée de http://www.genome.jp/kegg. Sur cette carte, les classes de voies métaboliques, telles que le métabolisme des glucides, des acides aminés et des lipides sont représentés selon les différentes couleurs, et les voies métaboliques de ces classes sont indiquées.

### 1.4.2.2 BioCyc

BioCyc (Caspi et al., 2012) est une collection de voies métaboliques de plusieurs organismes permettant aussi la construction et la publication de nouvelles données métaboliques (Baker, 2011). Biocyc contient MetaCyc qui est une représentation multi-organisme, ce qui serait l'équivalent des cartes de référence de KEGG, contenant des éléments vérifiés par essais biochimiques ou collectés manuellement à partir de bases de données spécifiques à un organisme. Toutes les autres bases de données de BioCyc sont une description des voies métaboliques et du génome d'un seul organisme déterminé par des expériences et des méthodes prédictives computationnelles. Par exemple, HumanCyc (Trupp et al., 2010) est la base de données pour l'humain. À sa dernière mise à jour, BioCyc contenait 5 711 organismes, 12 702 réactions, 12 361 métabolites et 2 363 voies métaboliques.

BioCyc définit ses voies métaboliques plus petites que d'autres bases de données, telles que KEGG (Altman et al., 2013), car pour cette base de données, il est important que chaque voie métabolique ne représente qu'une fonction biologique. En comparaison, les voies métaboliques de KEGG, qui sont plus grandes, dressent un portrait des réactions possibles dans un grand nombre d'organismes, des représentations souvent utilisées dans les livres de biochimie. Néanmoins, ces deux bases de données ne donnent pas d'information sur la localisation ou les compartiments où se déroulent ces réactions.

L'information de ces voies métaboliques combine donc de l'information protéique et métabolique souvent déduite par la séquence génomique des organismes. Bien que chaque science « omique » a été développée de façon isolée, les approches actuelles tendent à intégrer l'ensemble de ces sciences « omiques ». La section suivante présentera donc une nouvelle approche basée sur la combinaison de l'ensemble des données « omiques » où sont souvent utilisées les voies métaboliques.

## 1.5 Intégrer les « omiques » : vers une approche des systèmes

Une approche dite « des systèmes » permet l'analyse conjointe de différentes données dont l'objectif est l'identification de modèles élucidant les fonctions biologiques. L'approche des systèmes est un champ d'étude interdisciplinaire qui vise à explorer des systèmes biologiques complexes en utilisant des méthodes holistiques telles que les sciences « omiques » (Hood, Heath, Phelps, & Lin, 2004; Gu & Chen, 2014). L'intérêt de cette approche est d'aborder des questions biologiques plus complexes grâce à l'intégration de ces différentes données (Tilton et al., 2015). Les sciences « omiques » forment la base fondamentale de la biologie des systèmes.

### 1.5.1 Avantage et défis

Une approche des systèmes permet une dissection plus approfondie et informative des relations génotype-phénotype qu'une analyse utilisant seulement un type de données. Combiner plusieurs types de données « omiques » peut compenser pour de l'information manquante d'un type de données (Barallobre-Barreiro, Chung, & Mayr, 2013). Le postulat derrière ces analyses est que si des résultats de différentes données « omiques » convergent vers le même résultat, ce dernier est moins susceptible de conduire à des conclusions erronées.

Dans une approche des systèmes, la fonction encore inconnue de chaque élément peut parfois être inférée grâce aux associations avec d'autres éléments connus, ce qui permet une meilleure interprétation du système biologique. La compréhension du système peut se diviser en quatre propriétés (Kitano, 2002) : 1) la structure du système, telle que la topographie des voies métaboliques; 2) la dynamique du système, à savoir comment il réagit à différentes conditions; 3) ses méthodes de contrôle, c'est-à-dire les mécanismes mis en jeu pour minimiser les effets de dysfonctionnement cellulaire, et 4) le design de méthodes afin de modifier ou construire un système biologique ayant des propriétés désirées. Par exemple, en biologie synthétique, des systèmes sont créés afin de tester et d'améliorer notre compréhension des principes biologiques qui les gouvernent.

Néanmoins, de nombreux défis peuvent survenir lors d'analyses de ce genre. Parmi ceux-ci, on compte l'accessibilité aux échantillons pertinents, tels que des tissus spécifiques ou de sang récolté, dont la qualité affectera différemment les données « omiques » (Medina-Cleghorn & Nomura, 2014). Il est important de prendre en compte qu'une mauvaise préservation des échantillons aura un impact majeur sur les métabolites, voire les protéines, mais moindre sur le génome. De plus, comparativement aux transcrits et aux protéines, les métabolites ne partagent pas de lien direct avec l'information génétique et sont plutôt un produit de l'interaction des différents niveaux de régulation.

Ainsi, la prochaine section s'intéressera plus particulièrement aux approches développées pour l'intégration de deux sciences « omiques » : la génomique et la métabolomique. Cette intégration permettra d'élucider la fonction encore inconnue de gènes et de mieux comprendre le lien entre génotype et métabolome, le métabolome étant souvent plus proche du phénotype.

### 1.5.2 Intégrer génomique et métabolomique

La disponibilité croissante de séquences de génome complet contraste avec les connaissances actuellement limitées de la fonction des gènes. Cette découverte effectuée lors du projet du génome humain a montré qu'un fort investissement devrait être fait afin d'identifier les fonctions de ces gènes. En outre, une partie de ces derniers codent pour des enzymes non caractérisées qui ont un impact sur le métabolisme. Ces découvertes ont aussi révélé que les connaissances du métabolisme cellulaire étaient beaucoup moins complètes que ce qui était

pensé, ce qui a ouvert la porte à de nouvelles possibilités de voies métaboliques encore inconnues (Medina-Cleghorn & Nomura, 2014).

La métabolomique a permis d'aider à caractériser la fonction de certains de ces gènes orphelins et ainsi de découvrir de nouvelles voies métaboliques (Prosser, Larrouy-Maumus, & de Carvalho, 2014). Combiner génomique et métabolomique offre de multiples bénéfices, lesquels incluent une meilleure compréhension du lien complexe entre génotype et phénotype. Cette section présentera deux types d'approches où génomique et métabolomique ont été combinées.

#### 1.5.2.1 **Génomique fonctionnelle**

Un axe majeur de recherche biologique est d'identifier la fonction des gènes, ce qui permettrait une annotation complète du génome et aiderait à mieux comprendre les fonctions cellulaires. Néanmoins, bien que les avancées technologiques telles le séquençage et les approches computationnelles aient considérablement contribué à l'annotation des fonctions des gènes, une grande partie du génome demeure non annotée. On estime cette proportion à 40% (Galperin & Koonin, 2004). À l'aide de méthodes computationnelles, c'est-à-dire en utilisant une approche informatique, il a été possible de valider des hypothèses relativement à l'activité de certaines enzymes grâce à l'analyse ciblée de certains métabolites (Plata, Fuhrer, Hsiao, Sauer, & Vitkup, 2012). De la même façon, la validation de réactions appartenant à une voie métabolique précise a pu être effectuée par des études ciblées de métabolomique (Reaves, Young, Hosios, Xu, & Rabinowitz, 2013).

De plus, grâce à la couverture offerte par les approches non-ciblées, il a été possible de découvrir la fonction non suspectée d'enzymes en utilisant des techniques de profilage métabolomique basées sur l'activité enzymatique (*Activity based metabolomic profiling*) (Larrouy-Maumus et al., 2013; de Carvalho et al., 2010). Ces techniques permettent la découverte de nouvelles activités enzymatiques et de voies métaboliques par le criblage simultané et rapide d'une centaine de métabolites. À titre d'exemple, en utilisant cette technique, une étude a identifié le métabolite n-acyl taurine comme substrat de l'enzyme *fatty acid amide hydrolase* (*FAAH*; Saghatelian et al., 2004). Ultimement, avec les développements futurs des études métabolomiques permettant un meilleur débit et couverture, il sera possible de faire des prédictions automatiques de la fonction des enzymes grâce à des méthodes computationnelles et

un grand ensemble de données métabolomiques. Toutefois, les méthodes permettant d'effectuer ces prédictions restent encore à être développées (Sévin, Kuehne, Zamboni, & Sauer, 2015).

Cette dernière section présentera les données utilisées dans ce mémoire. Celles-ci proviennent de nouvelles études combinant génomique et métabolomique. Comparativement à la section précédente, ces études s'intéressent aux variants sur l'ensemble du génome.

#### 1.5.2.2 **Études d'associations pangénomique et métabolomique**

Alors que les études classiques de génomique fonctionnelle s'intéressent à caractériser la fonction d'un gène individuel, d'autres, telles que les études pangénomiques, se sont plutôt intéressées à identifier une multitude de variants génétiques directement associés à un phénotype. De fait, l'information obtenue par une approche pangénomique (GWAS – *Genome-Wide Association Study*) ne comprenait pas nécessairement assez de détails sur le mécanisme à l'origine du phénotype observé, comme dans le cas de traits complexes. Afin d'élucider les fonctions engendrant ces phénotypes, des études ont proposé de combiner une approche GWAS avec des études de métabolomique dans le but d'identifier la relation fonctionnelle entre variants génétiques et variations métaboliques (Adamski & Suhre, 2013; Suhre & Gieger, 2012). Ces études, nommées mGWAS, utilisent ainsi les concentrations de métabolites afin d'expliquer l'effet de ces variants génétiques et possiblement expliquer les mécanismes en jeu.

Les mGWAS consistent en une étude métabolomique combinée à des GWAS. Elles rapportent les mesures d'un très grand nombre de métabolites ainsi que les génotypes des variants dans une grande cohorte de patients. De cette façon, il est possible de faire l'association entre les niveaux de métabolites et les génotypes. Dans une approche pangénomique, les phénotypes classiquement mis en corrélation avec les variants sont remplacés par des phénotypes intermédiaires, les métabolites, dans les mGWAS. L'avantage de ces mGWAS et des GWAS de façon générale est qu'il n'est pas nécessaire d'avoir d'hypothèses préexistantes. De plus, les propriétés dynamiques du métabolome au cours du temps fournissent un moyen d'identifier l'impact des gènes et de l'environnement (Adamski, 2012). En règle générale, lors d'études mGWAS, l'objectif de l'analyse métabolomique est de faire l'acquisition du plus grand nombre de métabolites possible.

Néanmoins, puisque la métabolomique ne permet qu'une capture à un instant donné d'une partie de l'état métabolique d'un échantillon, les auteurs de certains des mGWAS (Suhre & Gieger, 2012) se sont demandé si le concept de phénotype métabolique individuel (appelé métabotype), soit de prouver que le métabolome est en partie contrôlé par des variants du génome, est réellement significatif et concrètement mesurable. Pour évaluer l'existence de ce métabotype, une étude longitudinale sur sept ans a estimé la conservation de celui-ci (Yousri et al., 2014). Les résultats de cette étude indiquent que plus de 40% des individus étudiés peuvent être identifiés de façon unique sur la base de leur profil métabolique. De plus, 95% des individus ont montré un haut degré de conservation du métabotype.

En utilisant les GWAS, approche de génomique « non-ciblée », avec des mesures métabolomiques, il est possible d'identifier les variants génétiques situés dans des gènes, ayant un impact sur le métabolisme et codant pour des enzymes et des transporteurs. En outre, les associations rapportées dans ces études fournissent des indications fonctionnelles sur les mécanismes impliqués. La régulation du métabolisme, soit les réactions métaboliques, sont entre autres influencées par la variation génétique qui affecte l'expression ou la fonction de ces protéines. De ce fait, les mGWAS ont eu un grand succès pour associer des gènes à des fonctions métaboliques. Chez l'homme, ces associations sont effectuées par l'analyse d'échantillons de sang et d'urine (Sévin et al., 2015). Par exemple, par une analyse de métabolomique ciblée de 363 métabolites utilisant la MS, une étude effectuée à partir d'échantillons de sang de 284 Allemands a identifié 4 loci (FADS1, LIPC, SCAD, MCAD) associés à des métabolites, dont un loci identifié au gène *FADS1* (*fatty acid desaturase 1*) (Gieger et al., 2008). Celui-ci était associé avec une phosphatidylcholine comprenant deux chaînes d'acide gras de 36 carbones et de 4 insaturations, ce qui est en accord avec le rôle de l'enzyme encodée par le gène *FADS1,* laquelle métabolise les acides gras polyinsaturés. Une autre étude effectuée à partir d'échantillons d'urine de 835 allemands et 601 brésiliens a identifié 11 loci associés à des métabolites (dont *FUT2* qui était associé à la lysine et au fucose) par une analyse métabolomique de 1 276 pics utilisant la RMN (Rueedi et al., 2014).

Particulièrement, une récente étude mGWAS effectuée par Shin *et al.*(Shin et al., 2014) a augmenté de façon substantielle le nombre de loci rapportés comme ayant une influence sur le métabolisme et, à ce jour, est considéré comme la plus complète (Kastenmüller, Raffler, Gieger, & Suhre, 2015). Cette étude a été réalisée dans deux cohortes d'origine européenne, soit la

*German KORA* (*Kooperative Gesundheitsforschung in der Region Augsburg*) et la *Twin UK*. Au total, 7 824 individus ont été recrutés et leurs échantillons de plasma ont été analysés sur des plateformes génomique et métabolomique. Pour l'analyse GWAS, 2,1 millions de SNPs ont directement été génotypés par une puce ou imputés grâce aux données du projet HapMap 2. Pour l'analyse métabolomique, plusieurs plateformes et approches ont été combinées, soit la GC-MS et la LC-MS, avec des approches ciblées et non-ciblées. L'analyse métabolomique a permis de mesurer 529 métabolites, dont 63% (n = 333) ont pu être annotés. Au total, cette étude a rapporté 299 associations significatives entre SNPs et métabolites, impliquant 145 loci. De ces associations, 84 loci ont été identifiés pour la première fois, tandis que 64 avaient déjà été identifiés par des études antérieures (Shin et al., 2014).

En outre, si les voies métaboliques sont contrôlées de façon coordonnée par des déterminants génétiques communs. Il est donc attendu que les résultats de mGWAS permettront d'identifier des loci génétiques qui seront associés à un groupe de métabolites appartenant à la même voie métabolique (Chan, Rowe, Hansen, & Kliebenstein, 2010). Il sera donc intéressant de relier ces associations entre variants génétiques et métabolites directement aux voies métaboliques. La section suivante présentera donc les besoins bio-informatiques afin de conduire ces études mGWAS.

*Nécessité d'outils bio-informatiques*

Du côté génomique, l'approche pangénomique des mGWAS identifient des millions de génotypes dans chaque échantillon. Sachant que ces études sont effectuées dans de grandes populations, soit plusieurs centaines voire des milliers d'individus, le nombre de génotypes obtenu sera donc considérable. Afin d'effectuer les tests statistiques permettant de statuer si chacun de ces génotypes est associé avec un trait, il a été essentiel de développer des outils bio-informatiques, tels que PLINK, logiciel permettant l'analyse des associations de GWAS (Purcell et al., 2007). De plus, une grande quantité d'algorithmes bio-informatiques, tels que IMPUTE 2 (Howie, Donnelly, & Marchini, 2009), ont aussi été développés afin d'effectuer l'imputation de variants génétiques en déséquilibre de liaison afin d'obtenir une meilleure couverture du génome.

La métabolomique génère un grand nombre de données, appelées signaux métaboliques. La gestion de ces données, leur traitement et leur analyse nécessitent donc des outils bio-

informatiques appropriés pour ces études (Shulaev, 2006). Le développement de ces outils, qui permettront le progrès de cette science « omique », comprennent : la gestion et le traitement des signaux métaboliques bruts, des analyses statistiques complexes (Sugimoto et al., 2012) ainsi qu'une exploration des résultats obtenus dans un cadre biologique. Le logiciel le plus utilisé pour le traitement et l'analyse statistique des données métabolomiques (Patti et al., 2012) est XCMS (C. A. Smith, Want, O'Maille, Abagyan, & Siuzdak, 2006).

Ainsi, combiner ces deux sciences « omiques » augmente le nombre de données à analyser et à interpréter, et donc le besoin de développer des outils bio-informatiques sophistiqués. Puisque le nombre d'échantillons analysés, de variants identifiés et de métabolites mesurés ne fait qu'augmenter avec les avancées technologiques, le rôle de la bio-informatique sera de plus en plus crucial pour analyser ces ensembles de données efficacement. De plus, l'interprétation de ces grandes quantités de données reste encore un défi de taille. Puisque ces données sont nouvelles, il est donc nécessaire de créer des outils bio-informatiques permettant leur interprétation dans le contexte des connaissances biologiques actuelles.

Afin de mettre en place une approche des systèmes, il est nécessaire de développer des stratégies d'analyse permettant l'intégration de données « omiques ». Ces stratégies permettront d'identifier les véritables associations avec les phénotypes étudiés et d'ainsi réduire le nombre de fausses découvertes (Ritchie, Holzinger, Li, Pendergrass, & Kim, 2015). La section suivante présentera les méthodes bio-informatiques utilisées pour l'intégration des « omiques », plus particulièrement dans le contexte du métabolisme.

## 1.6 Méthodes bio-informatiques pour l'intégration des « omiques » dans le contexte du métabolisme

Le terme « intégration » réfère à la situation où, pour un système biologique choisi, différentes données sont mises en commun et étudiées de façon conjointe afin d'augmenter la découverte (Ritchie et al., 2015).

L'intégration de ces données pourrait s'effectuer de façon conceptuelle (manuelle) après avoir identifié les éléments importants de chaque donnée (Cavill et al., 2011). Par exemple, les changements dans le niveau d'expression d'une enzyme et la concentration d'un métabolite de la même voie pourraient être expliqués par l'hypothèse d'une régulation différentielle de cette voie

métabolique. Bien que cette approche subjective puisse conduire à des explications biologiques plausibles, celle-ci peut négliger certains nouveaux mécanismes et s'avérer chronophage. Pour cette raison, des méthodes bio-informatiques ont été développées afin d'intégrer ces données de façon systématique.

### 1.6.1 Défis computationnels de l'intégration des données

Ces défis se posent en raison de différentes tailles, formats et dimensionnalités (*i.e.* le nombre de variables) des données à intégrer. De plus, la complexité des données générées ainsi que le niveau de bruit (éléments indésirables s'ajoutant au signal) de ces données sont des facteurs à prendre en compte (Gligorijević & Pržulj, 2015).

Par exemple, des effets confondants, provenant de variables indépendantes, peuvent mener à de fausses associations (R. Smith, Ventura, & Prince, 2014). Ces effets confondants peuvent par exemple être des facteurs démographiques, environnementaux, génétiques et techniques. Ils peuvent représenter un défi de taille lors de l'analyse et l'interprétation d'études intégrant différents types de  données.

De plus, étant donné la nature hétérogène des données « omiques », le balancement des données peut avoir un impact direct sur l'intégration. Par exemple, une expérience transcriptomique peut rapporter des dizaines de milliers de transcrits comparativement à une expérience métabolomique qui ne comprendra que quelques centaines de métabolites (Richards et al., 2010).

### 1.6.2 Intégration basée sur les voies métaboliques

L'intégration sur les voies métaboliques permet le développement rapide de méthodes visant la découverte du fonctionnement de la cellule et de l'organisme à partir de l'information moléculaire connue. Ces voies métaboliques sont utilisées car elles contiennent de l'information génétique, protéique et métabolique, ce qui permet la cartographie des données de génomique, transcriptomique, protéomique et métabolomique sur les mêmes voies. Des outils bio-informatiques ont été développées afin d'identifier quelles voies métaboliques sont mises en jeu par des données rapportées. Cela permet de découvrir les mécanismes impliqués dans les conditions étudiées en identifiant quelle voie métabolique est perturbée.

Ces méthodes basées sur les voies métaboliques facilitent donc l'interprétation biologique des résultats, car elles sont directement intégrées dans les connaissances du domaine. Ces méthodes sont des éléments-clés pour l'intégration des « omiques » même si elles sont très sensibles aux définitions d'experts de ce qui constitue une voie biochimique.

#### 1.6.2.1 **Analyse d'enrichissement de gènes et de métabolites**

*Gènes et métabolites*

L'analyse d'enrichissement provient de l'analyse de données de l'expression des gènes, *gene set enrichment analysis* (GSEA – Mootha et al., 2003; Subramanian, Tamayo, Mootha, Mukherjee, & Ebert, 2005). GSEA évalue si les données d'expression montrent un enrichissement pour un groupe (*sets*) prédéfini de gènes. Ces groupes sont des listes de gènes créées selon différentes informations, telles que l'appartenance à la même voie métabolique. Par exemple, l'analyse d'enrichissement attribuera un score à chaque voie métabolique selon les gènes retrouvés et l'expression de ceux-ci. Cela permet d'évaluer si des changements subtils, mais coordonnés, de l'expression de gènes d'une même voie métabolique sont observés. Cette analyse permet d'identifier des changements dans des voies métaboliques sans utiliser de seuil (*cut-off*) arbitraire pour la sélection des gènes, comme c'est le cas pour d'autres approches. GSEA est basée sur le test statistique non paramétrique Kolmogorov-Smirnov qui est utilisé afin de déterminer s'il y a des différences dans la distribution reliée aux gènes d'un groupe comparativement au reste des gènes (Irizarry, Wang, Zhou, & Speed, 2009). Depuis le premier article d'écrivant GSEA (Mootha et al., 2003), différentes versions de l'algorithme et du test statistique utilisé ont été développées (Nam & Kim, 2008).

En métabolomique, une méthode se basant sur cette analyse a été introduite afin de calculer l'enrichissement de métabolites (*metabolite set enrichment analysis* – MSEA; Xia & Wishart, 2010). Cette méthode utilise les concentrations de métabolites au lieu des valeurs d'expression de gènes et des groupes (*sets*) de métabolites créés à partir des métabolites appartenant à la même voie métabolique. De plus, MSEA permet le calcul d'enrichissement en utilisant la topologie des voies métaboliques. Par exemple, si des métabolites retrouvés sont des « *hubs* », soit des entrées/sorties de plusieurs réactions, cela augmente le score de la voie métabolique. Néanmoins, une des limitations de cette technique est qu'à ce jour, aucune des

plateformes métabolomiques n'effectue une couverture complète du métabolome. Cela amène des biais dans le calcul d'enrichissement car certains métabolites ne sont pas mesurables.

*L'intégration de plusieurs données « omiques »*

L'analyse d'enrichissement a aussi été utilisée afin d'intégrer des données « omiques » en combinant gènes, protéines et métabolites dans chacun des groupes (*sets*). Plus particulièrement, le logiciel IMPaLA (Kamburov, Cavill, Ebbels, Herwig, & Keun, 2011) a été créé afin d'intégrer des données de transcriptomique et de métabolomique. IMPaLA effectue l'analyse d'enrichissement de voies métaboliques en utilisant à la fois l'expression des gènes et les métabolites. IMPaLA permet l'identification de voies métaboliques additionnelles dont l'activité n'aurait pas été mise en évidence si une seule analyse (transcriptomique ou métabolomique) avait été utilisée.

Une méthode similaire à GSEA, appelée *Over-representation analysis* (ORA) (Tavazoie, Hughes, Campbell, Cho, & Church, 1999) a été développée afin d'évaluer si des éléments précédemment sélectionnés selon un seuil (*cut-off*) étaient surreprésentés dans des groupes (*sets*). Pour ce type d'analyse, seulement une liste de gènes ou de métabolites est sélectionnée sans prendre en compte les données quantitatives. Cette méthode utilise le test exact de Fisher, qui est non paramétrique, afin d'identifier les voies métaboliques surreprésentées. Par la suite, une correction telle que celle de Bonferroni (Dunn, 1959) est appliquée afin de corriger les valeurs de P selon le nombre de tests effectués. Cette technique a l'avantage de ne nécessiter qu'une liste de métabolites, sans concentrations ou abondances.

Étant donné que les approches basées sur les voies métaboliques comptent sur la définition de celles-ci, ces voies peuvent biaiser les résultats obtenus. À titre d'exemple, sachant que la taille et la complexité des voies métaboliques peut différer selon la base de données utilisée, le choix de la base de données influencera donc grandement les résultats d'enrichissement et de surreprésentation obtenus.

### 1.6.3 Intégration basée sur la construction de réseau biologique

Plusieurs approches de réseau biologique peuvent être mises en œuvre pour améliorer notre compréhension des signatures métaboliques. Les réseaux biologiques sont créés afin de représenter les liens complexes entre divers types de composantes cellulaires, tels que les

données « omiques » reliées aux connaissances actuelles. Ces réseaux sont représentés par des graphes contenant des nœuds et des arrêtes définis par l'utilisateur. Ces réseaux peuvent être utilisés afin d'identifier des altérations obtenues selon différentes conditions, sans dépendre uniquement des voies métaboliques préexistantes. Ces réseaux peuvent être créés sur la base des relations biochimiques des éléments le constituant. Par exemple, un lien (représenté par une arrête) peut être créé entre des gènes, protéines ou métabolites appartenant à la même réaction ou voie métabolique. Utilisant cette méthode, le logiciel MetaMapR (Grapov, Wanichthanarak, & Fiehn, 2015) permet la création de réseaux de métabolites selon leur similarité des structures, de l'activité enzymatique, de la masse ou des associations empiriques.

De plus, ces liens biochimiques, limités par les connaissances du domaine, peuvent être étendus par l'incorporation de relations empiriques ou statistiques (Grapov, Fahrmann, & Wanichthanarak, 2015). L'intégration empirique et statistique a l'avantage d'être plus objective que l'intégration biologique. Ces liens entre les ensembles de données sont réalisés en utilisant des procédures telles que la corrélation, la régression ou des techniques plus sophistiquées (Richards et al., 2010). Par exemple, plusieurs modèles empiriques utilisent des réseaux de corrélation entre différentes données « omiques » (Moco, Forshed, Vos, Bino, & Vervoort, 2008) (Adourian et al., 2008). L'analyse de co-expression permettra de trouver des fonctions partagées entre les données et représentera des régulations biologiques qui resteraient à être validées. De plus, certains de ces réseaux sont créés seulement sur la base des corrélations (réseaux de corrélations; Dumas, 2012) et permettent d'identifier des biomarqueurs (Adourian et al., 2008).

Bien que ces analyses de corrélation soient très utilisées pour l'intégration de données, ces approches peuvent donner un aperçu limité dans le cas où un grand nombre de données sont fortement corrélées entre elles, formant ainsi un réseau trop dense et difficilement interprétable. Dans ce cas, il s'avère qu'appliquer des méthodes plus sophistiqués telles que la modélisation graphique gaussienne (*Gaussian graphical models* GGM), utilisant des corrélations partielles et des réseaux bayésiens, ont la capacité de différencier les associations directes et indirectes (Grapov, Fahrmann, et al., 2015).

Un exemple d'outil est 3Omics (Kuo, Tian, & Tseng, 2013), qui permet l'intégration de données de trancriptomique, protéomique et métabolomique. Ce logiciel calcule les corrélations entre ces différentes données et les modélise grâce à un réseau biologique. Cet outil effectue

également l'analyse d'enrichissement de voies métaboliques avec les bases de données KEGG et HumanCyc, présentées plus tôt. De plus, 3Omics permet à l'utilisateur d'entrer seulement deux des trois données « omiques ». Le logiciel tentera alors de compléter l'information manquante (métabolites, protéines ou gènes) en effectuant du *text mining* (Clegg & Shepherd, 2008), qui est l'extraction de connaissances à partir de textes, comme les résumés des articles de PubMed.

De fait, les associations des mGWAS sont aussi des méthodes permettant l'intégration de données génomiques et métabolomiques qui peuvent être ajoutées au réseau biologique au même titre que les autres méthodes empiriques et statistiques.

Tout comme les études de Garrod (Garrod, 1923) sur les maladies innées du métabolisme énoncées dans le début de cette introduction, ces nouvelles études combinant plusieurs données « omiques » ont le potentiel de mieux expliquer le lien complexe entre génotype et phénotype. Plus particulièrement, les mGWAS permettront d'approfondir nos connaissances des contrôles génétiques sur le métabolisme.. Néanmoins, bien que les applications de ces études soient prometteuses, les mGWAS sont limités par leur  intégration dans un contexte biologique. De fait, il existe très peu d'outils permettant l'analyse des associations et leur interprétation biologique reste encore un défi de taille.

# Objectif et hypothèse

L'objectif de ce projet de maîtrise était de développer une méthode bio-informatique permettant de faire le lien entre gènes et métabolites.

Plus particulièrement, nous nous sommes intéressés aux gènes codant pour des protéines ayant un impact sur le métabolisme, soit les enzymes qui catalysent les réactions faisant partie intégrante des voies métaboliques. Les voies métaboliques de la base de données KEGG ont été utilisées comme référence afin de faire le lien entre gène et métabolite d'intérêt. Afin de quantifier ce lien, nous avons calculé la distance, définie comme le nombre de réactions entre l'enzyme encodée par le gène et le métabolite dans une voie métabolique. L'hypothèse de travail était que les métabolites d'intérêt sont des substrats/produits se trouvant à proximité, dans les voies métaboliques, des réactions catalysées par l'enzyme encodée par le gène. Afin de tester cette hypothèse, nous avons choisi d'utiliser les résultats d'études mGWAS. Plus particulièrement, nous avons appliqué la méthode à l'étude mGWAS par Shin *et al.* (Shin et al., 2014).

# 2 Article

Ce chapitre présente un manuscrit en préparation pour soumission au journal *Genome Medicine*. La méthodologie utilisée est bio-informatique et statistique. Ainsi, toutes les analyses sont effectuées de façon computationnelle.

Ma contribution spécifique à cet article est le développement de la méthode, son application sur des données de la littérature ainsi que l'analyse et l'interprétation des résultats obtenus. J'ai écrit l'ébauche de ce manuscrit et effectué les corrections suggérées par les co-auteurs à chaque version. Utilisant les scripts que j'ai développés pour l'analyse du mGWAS sélectionné, un *package* R (PathQuant) est présentement en cours de développement par une étudiante à la maîtrise, Sandra Therrien-Laperrière, et sera librement accessible en ligne afin d'être utilisé pour d'autres jeux de données.

# A systematic method to quantify mGWAS associations by metabolic pathway mapping

S. Cherkaoui[1,2], S. Therrien-Laperriere[1,2], K. Wanichthanarak[5,6], Dmitry Grapov[7], G. Boucher[2], The iGenoMed Consortium, G. Lettre[1,2,4] , J.D. Rioux[1,2,4] & C. Des Rosiers[1,2,3]

[1] Département de Biochimie, Université de Montréal, Québec, Canada

[2] Montreal Heart Institute, Québec, Canada

[3] Département de Nutrition, Université de Montréal, Québec, Canada

[4] Département de Médecine, Université de Montréal, Québec, Canada

[5] National Institutes of Health West Coast Metabolomics Center, California, USA

[6] University of California, Davis, California, USA

[7] CDS Creative Data Solutions, Ballwin, Missouri, USA

## 2.1 Abstract

The integration of the genome-wide association study approach with metabolomics analysis - termed mGWAS – offers tremendous opportunity to gain insights of the genetic control on metabolism. The interpretation of these mGWAS associations is currently time-consuming as it relies on literature searches. Therefore, an immediate bottleneck of mGWAS is the lack of a systematic and quantitative method to interpret associations. Here, we proposed a method to evaluate SNP-metabolite associations through mapping onto KEGG's pathways. As a proof-of-concept, we used a recently published mGWAS, calculated the number of metabolic reactions (distance) between SNPs annotated genes and metabolites pairs, and showed that it is shorter for associated pairs vs. randomly selected. Our method provides a high-throughput framework to link genes-metabolites pairs.

## 2.2 Background

The advent of high throughput technologies and approaches such as genome-wide association studies (GWAS) has contributed significantly to the identification of the genetic risk factors for many complex phenotypes. An advantage of this approach is that it is unbiased since it assesses genetic variables that are distributed across the genome instead of focusing only on specific loci of interest (Inouye & Abraham, 2013). GWAS have commonly been used for dissecting the genetic determinants of clinical endpoints and other quantitative traits. However, identifying the underlying mechanisms remains a challenge (McClellan & King, 2010). One way to tackle this issue is to combine GWAS with other omics data (Silverman & Loscalzo, 2012). In this regard, recent developments in technologies such as mass spectrometry (MS) and nuclear magnetic resonance (NMR) has enabled the measurement of several hundreds of metabolites, small molecules intermediates of metabolism, in a high-throughput manner (Patti et al., 2012). Similar to GWAS, metabolomics is an unbiased approach as it is designed to capture as much metabolic signals as possible from a non-pre-defined set of metabolites. Recent studies illustrate the value of combining classical GWAS with metabolomics analysis, referred to as mGWAS. These studies use metabolites levels as intermediate phenotypes to quantify the impact of genetic variants, thus expanding the window of phenotypes analyzed to the number of metabolites measured (Adamski, 2012).

Hence, the combination of genetic and metabolite analysis offers a tremendous opportunity to enhance our understanding of the genetic regulation of *in vivo* metabolism (Mootha & Hirschhorn, 2010). These mGWAS report associations between metabolites and single nucleotide polymorphism (SNP), associations which are also referred to as metabolite quantitative trait loci (mQTL). These mGWAS can identify multiple loci of interest, which need to be annotated to a gene according to the gene function (Suhre & Gieger, 2012). Although these studies are very promising, the biological interpretation of the identified gene-metabolite associations remains a major challenge (Sévin et al., 2015). To date, a few methods have been developed for the interpretation of mGWAS associations. These include: 1) Gaussian Graphical Modeling (GGM) (Krumsiek, Suhre, Illig, Adamski, & Theis, 2011) provides a reconstruction of metabolic pathways from metabolomics data; and 2) *Mining the unknown* (Krumsiek et al., 2012), which enables the identification of unknown metabolites using mGWAS, GGM and

metabolic pathway databases, with the assumption that associations reveal biological functions. However, these methods do not quantitatively evaluate individual gene-metabolite associations in the context of current biochemical knowledge.

Thus, the objective of this study was to develop a method that would enable a systematic and quantitative evaluation of the biochemical relevance of SNP-metabolite associations, where the SNP is replaced by his annotated gene by the user. Our method involves mapping of these pairs of gene-metabolite on metabolic pathways, namely form KEGG database (Ogata et al., 1999). Chosen pathway was parsed to create a biochemical network in order to compute distance path between pairs of associated gene-metabolite, corresponding respectively to edges and nodes of the network. Our hypothesis is that the calculated distance will be shorter between associated compared to randomly selected pairs of gene-metabolite. The method was tested and its utility demonstrated using the associations from Shin *et al.*(Shin et al., 2014), which represents, to date, the most complete mGWAS on human blood (Kastenmüller et al., 2015).

## 2.3 Method

The method was developed using R programming language (www.r-project.org). The package, PathQuant (for Pathway Quantify) is available at https://github.com/sandraTL/PathQuant. Figure I depicts the package workflow for the following steps:

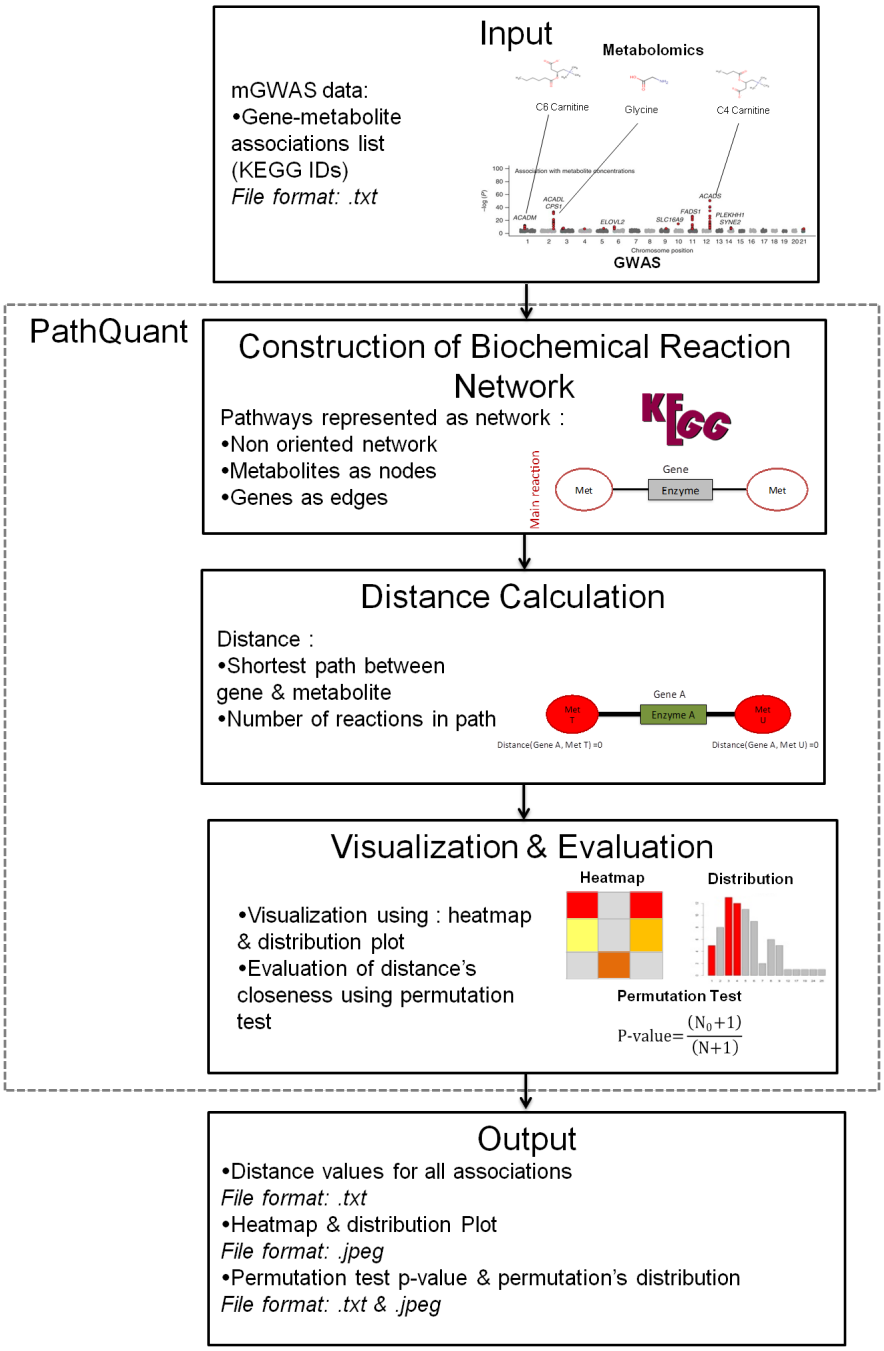


**Figure I : Method workflow.**

### 2.3.1 Input data

PathQuant accepts gene-metabolite associations as input in a text-delimited tab file (.txt), which should include the following columns: 1) the associated SNPs annotated to genes; and 2) the associated metabolites, each with their specific KEGG identifiers (IDs). Each row represents a gene-metabolite association. Only associations between genes and single metabolites are taken as entry, hence associations between a gene and a ratio of metabolites have to be separated prior to analysis.

### 2.3.2 Construction of the biochemical reaction network

The method was built using the publicly available KEGG pathway databases. KEGG pathways, encoded in KEGG XML file format (KGML), are downloaded when selected using KEGG REST API (http://www.kegg.jp/kegg/docs/keggapi.html). The pathway selected by the user is then converted to a network of biochemical reactions (gene –metabolite network) with: metabolites as nodes and genes, mapped to their corresponding encoded enzymes, as edges. Genes that encode for an enzyme that catalyzes multiple reactions are mapped to multiple edges. Our network includes only metabolites that are the main substrate-product pairs (reactant pairs; information contained in KEGG RPAIRS), which are defined for all reactions. Specifically, these pairs do not include cofactors, such as NAD, or co-substrates/products, such as ATP or $H_2O$. The latter are defined as ubiquitous metabolites and are not major components of a transformation series (Cottret et al., 2010); they are part of many reactions, which if included in the network, can cause artefacts when computing distances since they are found in multiple reactions. Finally, we have used an undirected network to give equal importance to changes in levels of both upstream and downstream metabolites of a given reaction catalyzed by a gene's encoded enzyme.

### 2.3.3 Distance calculation

Our method evaluates distances, which are defined as the shortest path between a given gene and a metabolite in the biochemical reaction network. This distance is computed as follows:

D(gene,metabolite)= number of edges between a given gene (edge) and metabolite (node)

A distance of 0 is assigned to metabolites, which are the main substrates or products of the reaction catalyzed by a given enzyme, which is encoded by the selected gene of interest. The distance from all other metabolites are obtained using the breadth-first search algorithm (Skiena, 2008), which finds the shortest path between the substrate and product, thereby selecting the shortest distance between these two. Figure II represents an example of distance calculation for the gene *FH*, which encodes for the enzyme fumarase, which is part of the Krebs cycle. The main reactant pairs of this reaction, fumarate and malate, are set at a distance of 0 (coloured in red). For other Krebs's cycle metabolites, the shortest distance between them and fumarate or malate are calculated. For example, citrate is two reactions away from malate and six from fumarate, thus a distance of two is assigned to citrate.

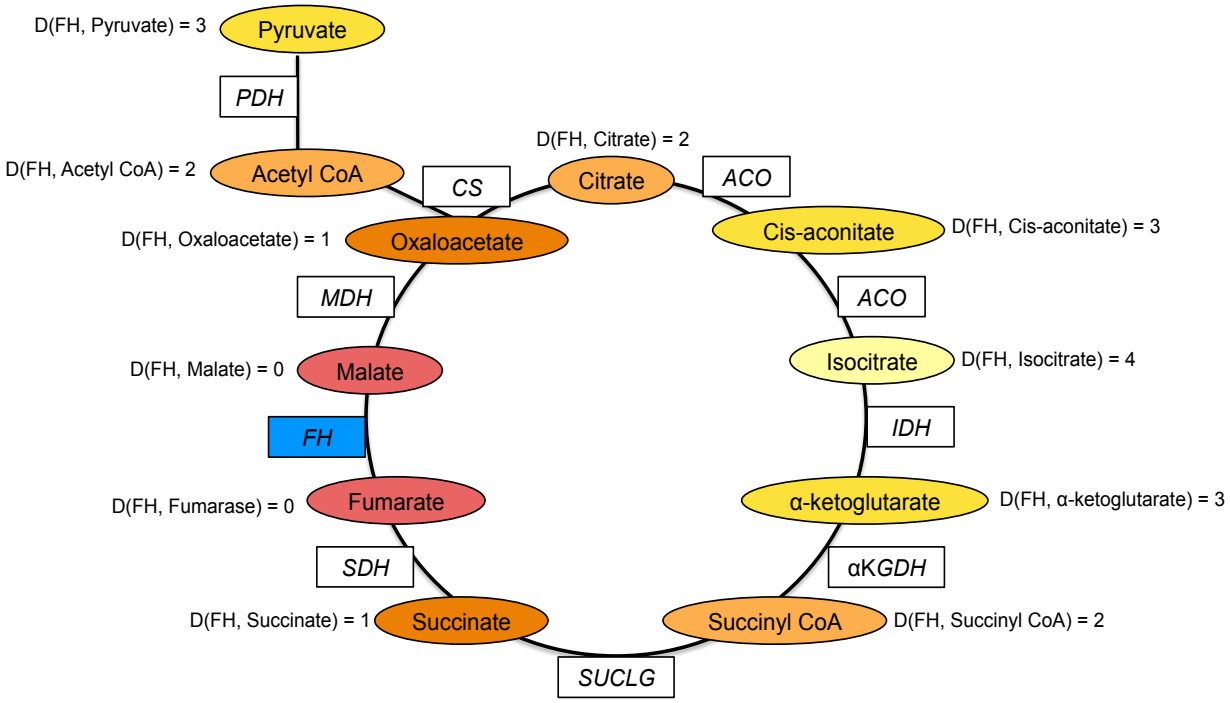


**Figure II : Principle of distance calculation using as example the gene *FH* (in blue), which encodes for the enzyme fumarase in the Krebs Cycle.** Metabolites are represented as nodes and genes mapped to their corresponding enzyme as edges (rectangles). For simplification, some genes are not shown but included in their enzymatic complex (for e.g. DLD is in αKGDH and PDH). Metabolites are coloured according to their distance to reaction catalyzed by fumarase; corresponding color code (from red – closest; to yellow - farthest).

### 2.3.4 Distance visualization and statistical analysis

The distances for all gene-metabolite associations are presented as heatmap and distribution plots. A permutation test was built into the method to assess if distances for gene-metabolite associations (referred to as associated set) are significantly smaller than that obtained by shuffling genes and metabolites (referred to as permuted set). This permutated set was obtained by randomly selecting genes and metabolites that were measured in the mGWAS and which were in the overview map. For each permutated set, distances are computed between the same number of gene-metabolite pairs as the associated set. The statistical significance is obtained by comparing the median distance of the associated set to median distances of permutated sets using the following formula:

(1) $$\text{P-value}=\frac{(N_0+1)}{(N+1)}$$

$N_0$ denotes the number of permutations with median value smaller or equal to the median of the associated sets. N denotes the number of permutations. The precision level of the obtained P-value will be set according to the number of permutations performed.

### 2.3.5 Data outputs

PathQuant gives as output the distance values for all associations in a text-delimited tab file, heatmap and distribution plot, in JPEG file format, as well as the permutation test p-value along with the distribution of the permutation median values in JPEG file format.

## 2.4 Results & discussion

We applied the method to the recent mGWAS by Shin *et al.* (Shin et al., 2014), which was conducted in 7824 human blood samples and for which 2.1 million SNPs and 529 metabolites (assessed by untargeted metabolomics) were analysed. Shin *et al.* reported 299 SNP-metabolite associations and 145 loci, which were annotated by the authors of the study to 132 genes (Table 1)(Shin et al., 2014).

### 2.4.1 Building input data - Identification

Since only genes and metabolites names were provided in supplementary files of Shin *et al*., KEGG IDs had to be assigned to build the input data file. While for genes, the identification was automatically achieved using KEGG IDs lists (obtained at : http://www.genome.jp/linkdb/), for metabolites this was problematic and ambiguous because metabolites have multiple synonymous names. Thus, the identification for metabolites was achieved manually using KEGG compound database (M Kanehisa & Goto, 2000) and double checked using the Human Metabolome database (HMDB) (Wishart et al., 2013). We recommend for future mGWAS publications that investigators provide common identification for metabolites (e.g., InChI or SMILES)(Salek, Steinbeck, Viant, Goodacre, & Dunn, 2013). Furthermore, many of the listed metabolites did not have KEGG IDs, among which were acylcarnitine derivatives. These metabolites are found predominantly in plasma, but arise from intracellular metabolism of their acyl-CoA counterparts (Friolet, Hoppeler, & Krähenbühl, 1994). Since KEGG database encompasses predominantly intracellular enzymatic reactions and their corresponding metabolites, we created a rule to bridge the gap whereby, all 20 listed metabolites with names ending with 'carnitine' (shown in red in Tables S2) that do not have KEGG IDs were replaced by their acyl-CoA counterparts (ending with 'CoA'). As a result, we assigned KEGG IDs to 10 carnitine derivatives.

In summary, all 132 genes listed in the study were identified and had KEGG IDs (Table 1 and Table S1). For metabolites, out of the 529 that were measured in this study, 333 were annotated (chemically identified), while 196 were listed as unknown (Shin et al., 2014). From those 333 annotated metabolites, 229 metabolites were identified (Table 1 and Table S2).

Finally, Shin *et al.* reported 299 associations. Among these, 245 were associations between a single gene and a metabolite whereas 54 were associations between a gene and metabolite ratios. For the latter, it was assumed that the gene was associated with both metabolites from the ratio; therefore, associations with ratios were separated in two distinct associations. At this step, we thus lose information about ratios, which could be merge back afterward for further evaluation. In total, there were 129 gene-metabolite associations of which genes and metabolites were identified with KEGG IDs (Table 1).

**Table 1. Numbers of genes, metabolites and gene-metabolite associations reported in Shin *et al.* which were identified in KEGG and map on its overview map.**

| | Genes | Metabolites | Associations |
|---|---|---|---|
| **In mGWAS Study** | 132 | 529 *(333 annotated and 196 unknowns)* | 299 *(245 with single metabolites and 54 with ratios)* |
| **Identified KEGG database** | 132 | 229 | 129 |
| **Mapped overview map** | 49 | 98 | 24 |

### 2.4.2 Mapping on biochemical reaction network

We have selected KEGG's most complete metabolism's map for the biochemical reaction network: named metabolic pathways that is part of the global and overview maps (referred thereafter as 'overview map' throughout this article). Other KEGG metabolic pathways could be used with PathQuant, but are smaller and, hence, fewer associations will be mapped. Genes and metabolites with KEGG IDs were then mapped to the constructed biochemical network representing the overview map. Out of the 132 genes and 229 metabolites with KEGG IDS, 49 and 98, respectively were mapped on the overview map (see Table 1 and Figure S I).

Unmapped genes may possibly be those coding for transporters, which are not part of the network (29 transporters reported in Shin & al. 's biological annotation) or enzymes not in the overview map. Furthermore, as can be seen in Figure S I, which depicts the 49 genes and 98 metabolites mapped to KEGG's original overview map (using KEGG mapper: http://www.genome.jp/kegg/tool/map_pathway2.html), measured metabolites and genes are concentrated in the same region of this map, mostly related to 'Amino Acid' and 'Energy'. For these two pathway classes, defined in Shin et al. as 'Super-Pathways' (Table 2), 76 and 83%, respectively, measured metabolites had a KEGG ID. In contrast, other regions such as that related to "Lipid Metabolism" (shown in green on the left of Figure S I) include only a few metabolites, which indicates their poor coverage. In fact, while 'Lipids' was the most represented

pathway class of metabolites measured (n=130), only 22% of them were mapped to the overview map.

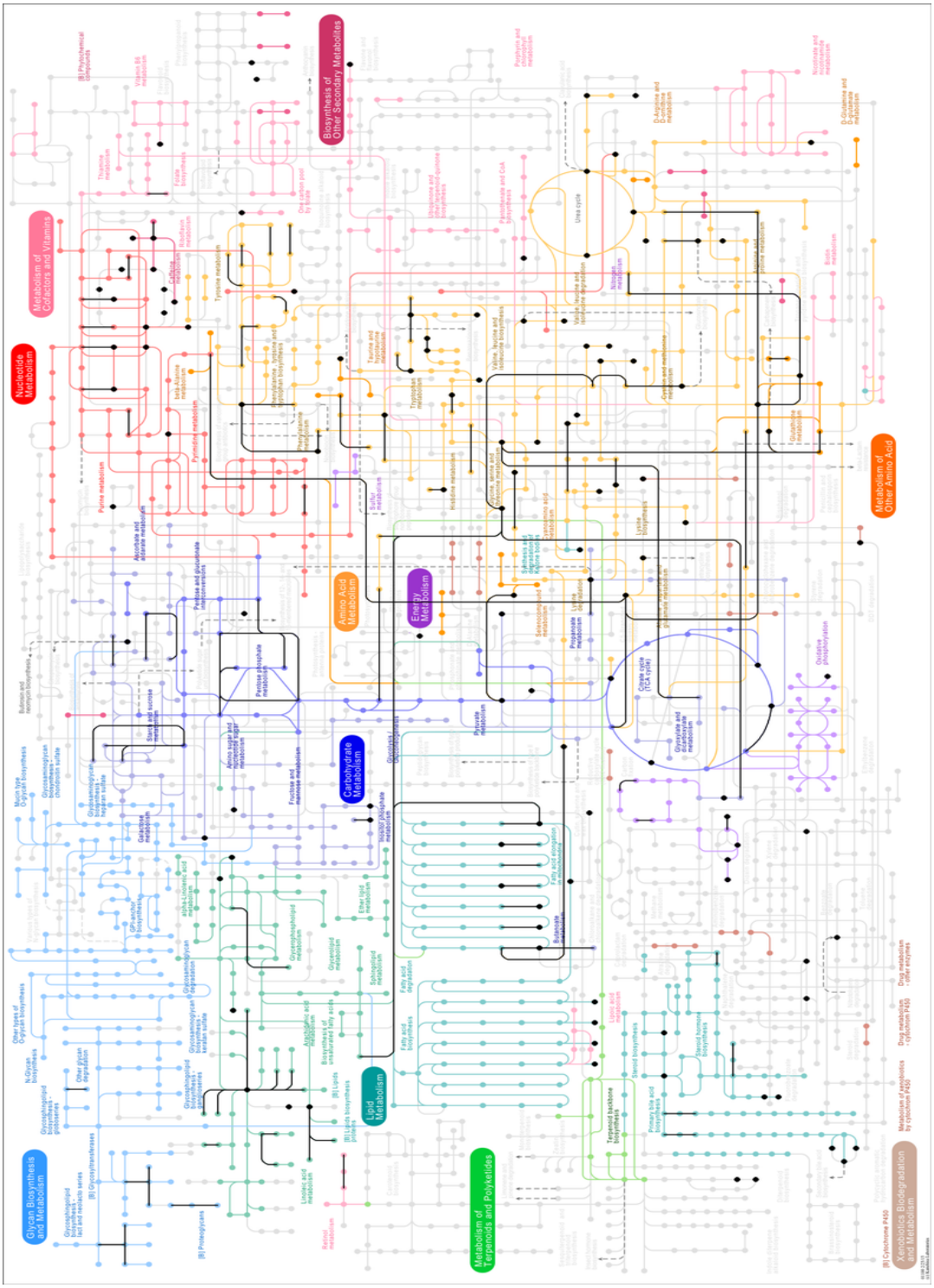

**Figure S I : Overview of measured genes and metabolites mapped to KEGG overview map.** Genes and metabolites measured from the mGWAS study were mapped by black edges and nodes respectively. The Figure was generated using KEGG mapper (http://www.genome.jp/kegg/tool/map_pathway2.html).

Hence, these examples demonstrate that various pathway classes are not equally represented in the overview map; thereby highlighting the coverage limitation of this database and map. As a result, the number of gene-metabolite associations that could be mapped to this overview was also limited. Indeed, from the 129 associations that had KEGG IDs, only 24 associations were mapped on the overview map (Table 1) and were used to compute distances. A challenge of using only one map, like the overview, is the limited coverage. However, this challenge could be tackle by using other pathways in KEGG, feature available in PathQuant but not done in this application since it was a proof-of-concept.

**Table 2. Number of annotated metabolites in each pathway class, referred to as 'Super-Pathway' by Shin *et al.*, and coverage percentage on KEGG and on KEGG overview map.**

| | Amino acids | Carbohy-drates | Cofactors & vitamins | Energy | Lipids | Nucleotides | Peptides | Xenobiotics |
|---|---|---|---|---|---|---|---|---|
| **Number of measured** | 80 | 14 | 15 | 6 | 130 | 14 | 27 | 47 |
| **% of measured metabolites (333)** | 24% | 4% | 5% | 2% | 39% | 4% | 8% | 14% |
| **Coverage Percentage** | | | | | | | | |
| **KEGG** | 76% | 86% | 73% | 100% | 65% | 86% | 15% | 79% |
| **Overview map** | 51% | 43% | 27% | 83% | 22% | 64% | 0% | 11% |

### 2.4.3 Distance calculation and visualization

#### 2.4.3.1 Calculation for all mapped associations on the biochemical network

We calculated distances between the 18 genes and 19 metabolites involved in the 24 resulting mapped associations. Figure III shows these distances, with the 24 gene-metabolite associations highlighted by a thick black border. Out of these 24 gene-metabolite associations, 10 have a distance of 0, implying that these gene-metabolite pairs, e.g. *PSPH (phosphoserine phosphatase)*-L-Serine, are from the same enzymatic reaction. The longest distances are found between *TYMP (thymidine phosphorylase)*-Uridine (distance = 6), and between *NAT2(N-acetyltransferase 2)*-4-acetamidobutanotate (distance = infinite). Infinite distance indicates that there are no series of reactions that can link this specific gene and metabolite. This is attributed to the incompletion of biological knowledge, which affects pathway-based integration (Grapov, Fahrmann, et al., 2015). This incompleteness impacts the biochemical network. As a result, all reactions are not connected as one big network, but rather makes multiple sub-networks, containing gaps which result from missing reactions (Orth & Palsson, 2010). Because of these infinite values, the descriptive statistic use is a median instead of a mean. For all 24 computed distances, the median was found to be 1. Hence for most associations, metabolites are one reaction further than that catalyzed by the gene-encoded enzyme.

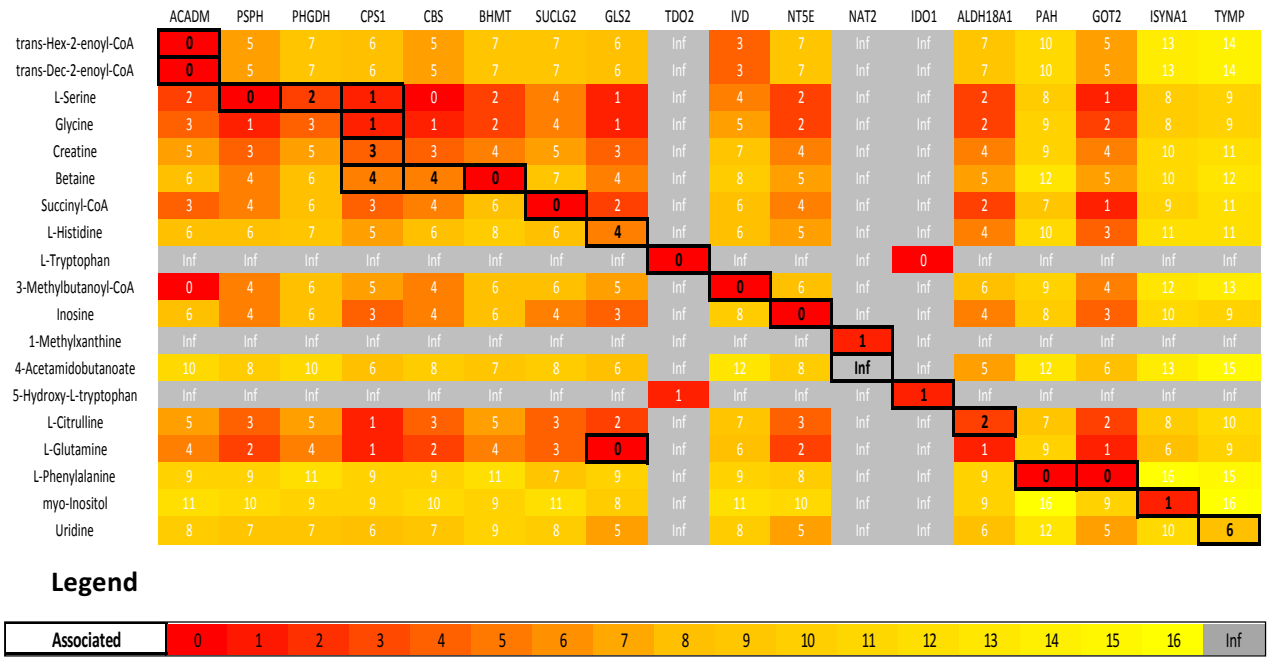


**Figure III : Heatmap of distances calculated using PathQuant between the associated 18 genes and 19 metabolites reported by Shin *et al.* (Shin et al., 2014).** Columns represent genes and rows represent metabolites. The calculated distance is shown in each cell with the corresponding color code (from red – closest; to yellow - farthest). The distance calculated for the 24 gene-metabolite associations are reported in cells with a thick black border. Inf = infinite value, which means that there is no known path between the gene and the metabolite.

### 2.4.3.2 Distance visualisation for individual genes: Examples with *GLS2* and *TYMP*

Figure IV illustrates the distribution of calculated distances between all measured metabolites that could be mapped using two genes as examples: *GLS2* which encodes glutaminase 2 and *TYMP* for thymidine phosphorylase. The distance calculated for the metabolites associated with these two genes are shown as red bars, while all measured but non-associated metabolites are shown as grey bars. Specifically, *GLS2* was associated with L-glutamine and L-histidine (shown in red), which was, respectively, at a distance of 0 and 4. It is noteworthy that there were some measured metabolites that were at a distance closer than 4, but

were not associated with *GLS2*, although most of these metabolites had a distance greater than 4 and up to 25. As for *TYMP*, its associated metabolite uridine is at a distance of 6 (shown in red), which is the shortest distance calculated; all other measured metabolites had a distance > 6.

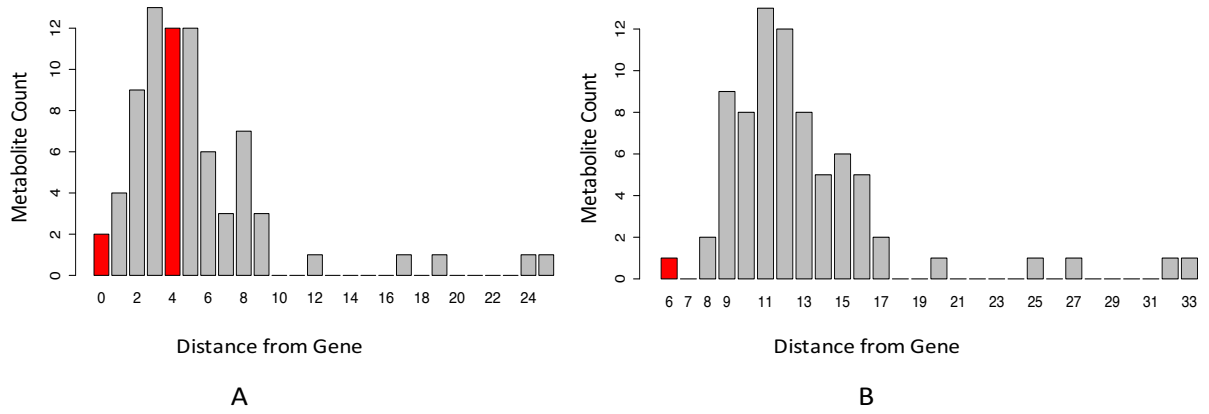


**Figure IV : Distance distribution plots for gene (A) GLS2 and (B) TYMP.** The plots are depicted as frequency bars, which represent the number of metabolites at a given distance for the selected gene. Frequency bars are shown in grey for metabolites that are not associated with the selected gene and in red if there is at least one metabolite associated with this gene.

### 2.4.4 Distance Evaluation – Statistical analysis

A permutation test was used in order to evaluate if the computed distances for associated genes and metabolites are statistically shorter than those obtained at random. Similar to the associated set, each permutated set contained 18 genes and 19 metabolites, which were randomly selected from 49 genes and 98 metabolites reported by Shin *et al.* that were mapped to the overview (Table 1). For gene selection, a condition was built to match the number of reactions of genes from permutated sets to the one of genes from the associated sets, with a difference of +/-1 except for genes with only 1 reaction. This condition was added because the number of reactions catalyzed by an enzyme encoded by a given gene could bias the distance value e.g. the odds for having a short distance to a metabolite is higher if the gene has multiple reactions. For each

permutation sets, 24 distances were computed, identical to the associated set pattern (position of cells with thick black border in Figure III).

The median value of these distances for each 1000 permutations sets, depicted as a distribution plot in Figure V, varied from 4 to 34, compared to the median value of the associated set which was 1 (shown in red). It should be noted that none of the permutation sets had a median distance as short as the one found for the associated set. We further assessed whether the median of the associated set is significantly shorter than the distribution of the permutation sets by computing the p-value using formula (1) and obtained a P-value equals to $9.9\times10^{-4}$. Considering the number of permutations (1000), we can say that the precision level of the P value is smaller or equal to 0.01.

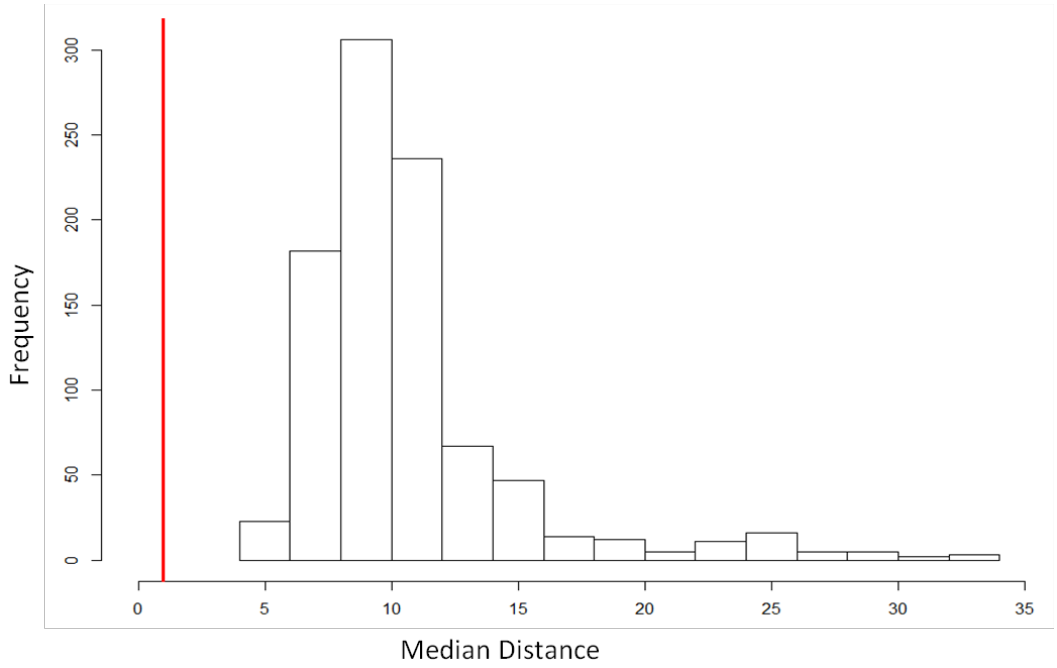


**Figure V : Distribution of median distance values for the 1000 permutation sets**. For comparison, the median value of the associated sets is shown in red.

Collectively, these results demonstrate the utility of our method to assess the biological relevance of gene-metabolite associations reported by mGWAS. Moreover, these results confirm our hypothesis of proximity between gene-metabolite associations.

### 2.4.5 Comparison with other methods

As described in detail below, our method compares advantageously to currently available methods for mGWAS interpretation. To date, mGWAS interpretation has been achieved using three different approaches, namely manual biological annotation, GGM (Krumsiek et al., 2011) and mining the unknown (Krumsiek et al., 2012).

Biological annotation: Table S 4 compares our distance results for the gene-metabolite associations used in the method (rows 2-21) to the manual description of these associations made by Shin et al., referred as biological annotation, (rows 24-40). Metabolites shown in green (in row 2-21 and row 30) are those for which there is agreement between results of our method and Shin *et al*. For these 18 genes (columns), we were able to compute distances for almost all associations except two: *ACADM* (*acyl-CoA dehydrogenase, C-4 to C-12 straight chain)*-acetylcarnitine and *IVD (isovaleryl-CoA dehydrogenase)*-propionylcarnitine were not found by PathQuant. Furthermore, we could compute distances for some metabolites, which were not described by Shin *et al*., namely for *CPS1(carbamoyl-phosphate synthase 1, mitochondrial)* which was associated with L-serine, creatine and betaine at a respective distance of 1, 3 and 4. In summary, while both manual biological annotation and our methods enable similar interpretation of mGWAS datasets, our method offers the advantage of a systematic and quantitative evaluation of these associations.

GGM: This method provides a reconstruction of metabolic pathways from metabolomics data using an undirected network in which each edge represents a pairwise partial correlation between measured metabolites (Krumsiek et al., 2011). However, GGMs only map analysed data such as measured metabolites. In their study, Shin *et al.* have connected metabolites using GGMs and linked genes to the network based on gene-metabolite associations (Shin et al., 2014). It is noteworthy that compared to PathQuant, GGM does not evaluate the relevance of individual associations based on biological knowledge from pathway databases, but rather recovers metabolic pathways in a data-driven approach. The difference between PathQuant and GGM was illustrated using the genes *CPS1*, *CBS (cystathionine-beta-synthase)* and *BHMT (betaine--homocysteine S-methyltransferase)*, which were found associated with the metabolite betaine; using our method, we calculated a distance of 4, 4 and 0, respectively (Figure V), which enables one to confirm that betaine is catalyzed by the protein encoded by BHMT (distance of 0). From

the resulting GGM network, shown in Figure S II, one can conclude that betaine is correlated with metabolites from the 'Amino Acid' (in green) and 'Lipid' (in red) pathway classes as well as associated with the genes CPS1, CBS and BHMT (Figure S II). GGM cannot differentiate between those 3 genes (i.e. showed equally important). Nevertheless it is noteworthy that GGM uses a quantitative approach based on proximity in metabolic pathways, which bears similarity to our method, to assess the relevance of the resulting GGM network. Interestingly, similar to our results, Krumsiek *et al.* reported unconnected metabolite pairs for which the distance was infinite, which they interpreted as potentially representing novel pathways (Krumsiek et al., 2011).

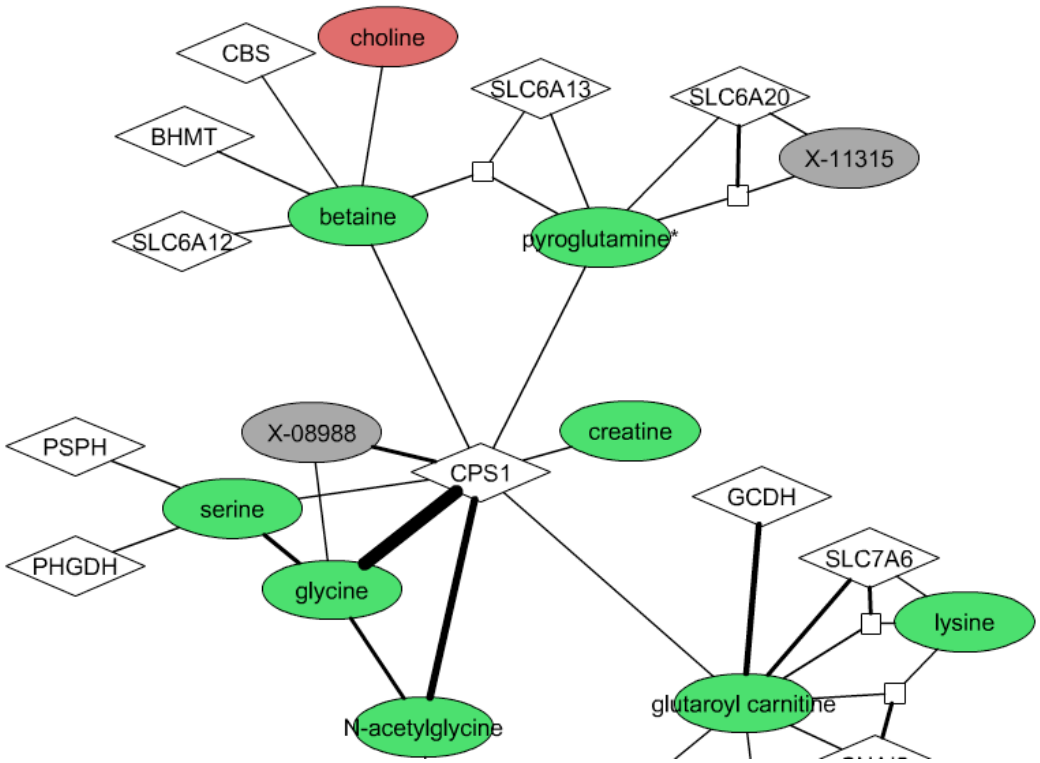


**Figure S II : Selected region of the GGM network representing the associations between CPS1, CBS and BHMT with betaine.** Each node represents either a metabolite (circular nodes) or a gene (diamond-shaped nodes). An edge between a metabolite and a gene was drawn for metabolites that showed an association with the indicated gene. A line between two metabolites nodes indicates a connection in the underlying GGM correlations. Metabolite colors are indicative of the metabolite pathway class: green for amino acid, red for lipid and grey for

unknown. Network was generated using results from Shin *et al.* and their online website (http://metabolomics.helmholtz-muenchen.de/gwa/si/network.php)

Mining the unknown: This method enables the identification of unknown metabolites using mGWAS, GGM and metabolic pathway databases such as KEGG, based on the assumption that gene-metabolite associations reveal biological functions (Krumsiek et al., 2012). This method is based on the premise that if an unknown metabolite displays a similar statistical association with a genetic locus in a GWAS or a known metabolite in a GGM, then this may provide specific information of where it is located in a metabolic pathway. It is specifically used to identify biochemically related metabolites. Similar to PathQuant, this method is based on a proximity-based approach. However, it's very different objective makes any comparison difficult, since PathQuant does not include unknown metabolites.

Overall, by comparing our method to the existing ones, we can assert that, to the best of our knowledge, this is the first method to systematically interpret mGWAS associations using distance calculation on metabolic pathways and testing proximity of gene-metabolite associations as a mean to assess the biochemical relevance of these associations. Nevertheless, GGM (Krumsiek et al., 2011) and mining the unknown could be combined with PathQuant as complementary methods.

### 2.4.6 Limitations and other applications

Beyond the aforementioned coverage limitations by the choice of a single map in this proof-of concept, there are additional constraints linked to the analysis of this mGWAS. In fact, PathQuant uses gene-metabolite associations, while mGWAS report SNP-metabolite associations. For this reason, distances that have been computed may be impacted by the annotation method used by the authors. The SNP-gene annotation approach used ub Shin *et al*., is based on the selection of one of the closest gene from the lead SNP (within 500kb), selection made using biological knowledge from database such as KEGG, EHMN (Edinburgh human metabolic network reconstruction)(Hao, Ma, Zhao, & Goryanin, 2010), PubMed and BRENDA (Schomburg et al., 2013). Since PathQuant calculates distances between the annotated gene and its associated metabolites using pathway mapping on KEGG, we do not exclude the possibility of some redundancy in our distance calculation. To address this issue, our method could be validated with other mGWAS data, which have used a SNP-gene annotation approach that is not

based on metabolic databases (Chan et al., 2010). Interestingly, this limitation suggests that our method could be used as an annotation tool to link the associated SNPs to their corresponding gene. This could be achieved by selecting among the various genes physically close to the lead SNP of interest (500 kb by example), the one that has the shortest distance to the associated metabolite in a given metabolic pathway.

Many mGWAS studies (Suhre et al., 2011) and reviews (Adamski & Suhre, 2013) took inborn error of metabolism as examples to illustrate inherited variations in human metabolism. These disorders are caused by a mutation in a single gene which produces the loss of its enzyme function and results in the alteration of linked reaction metabolites (Mootha & Hirschhorn, 2010). As a straightforward validation, we tested PathQuant with genes and metabolites taken from these disorders, where the enzyme causing the defect and the toxic accumulating metabolites and where SNP-gene annotation is not issue. More precisely, we used urea cycle disorders as examples. Most of these diseases result in hyperammonemia but each disorder affects different enzymes, for which we have calculated distances (Table 3). As expected, most of the distance are short but not equal to zero (0 - 5). Hence, in these disorders, the metabolic phenotype is not necessary immediate substrate/product of the enzyme reaction; this illustrates the relative 'unpredictability' of the metabolism and the applicability of PathQuant.

**Table 3. Distance calculation used for urea cycle disorders (inborn error of metabolism), with reported deficient gene encoded enzyme and its clinically measured metabolites for diagnosis, end product of the single gene defect.** Information on these disorders are take from medscape (http://emedicine.medscape.com/) .

| Disease Name | Gene | Metabolite | Distance |
|---|---|---|---|
| N-Acetylglutamate synthase deficiency | NAGS | Ammonia | 2 |
| Ornithine transcarbamylase deficiency | OTC | Ammonia | 2 |
| Ornithine transcarbamylase deficiency | OTC | Ornithine | 0 |
| Ornithine transcarbamylase deficiency | OTC | Glutamine | 2 |
| Ornithine transcarbamylase deficiency | OTC | Alanine | 5 |
| Ornithine transcarbamylase deficiency | OTC | Citrulline | 0 |
| Argininosuccinic acid synthetase defeciency (citrullinemia) | ASS | Citrulline | 1 |
| Argininosuccinase acid lyase deficiency (argininosuccinic aciduria) | ASL | Argininosuccinate | 3 |
| Argininosuccinase acid lyase deficiency (argininosuccinic aciduria) | ASL | Ammonia | 3 |
| Arginase deficiency (argininemia) | ARG | Arginine | 0 |
| Carbamoyl Phosphate Synthetase Deficiency | CPS1 | Ammonia | 0 |

## 2.5 Conclusion

The availability of relevant samples, large cohorts, and accurate high-throughput technology have allowed researchers to leverage metabolomic profiling to elucidate the genetic influence on intermediate phenotypes, namely metabolites (Inouye & Abraham, 2013). There is, however, a need for methods to interpret the results of these mGWAS in a biological context. To the best of our knowledge, we are proposing the first method that provides a systematic and quantitative evaluation of the biochemical relation of mGWAS gene-metabolite associations through pathway mapping. To achieve this, a specific pathway map, which is an overview representation of the metabolism in KEGG, was selected. The proof-of-concept of this method was demonstrated using mGWAS data by Shin *et al.* Despite some limitations inherent to the poor coverage of some metabolic pathway classes, such as those for lipids, on KEGG overview map, results obtained supported the hypothesis that distances are shorter for gene-metabolite pairs that are associated vs. those selected at random. Taken together, the developed method and its potential applications will enable a better understanding of how genetic variation impacts metabolism by quantifying the relation between mGWAS associations.

## 2.6 Description of additional data files

Supplementary Table 1. Genes Identification in KEGG

Supplementary Table 2. Metabolites Identification in KEGG

Supplementary Table 3. Associations Identification in KEGG

Supplementary Table 4. Difference between results obtained using PathQuant compared to Shin *et al.* biological annotations.

# 3 Discussion

L'objectif de mon travail de maîtrise était de créer et d'utiliser une approche bio-informatique afin de générer de façon systématique une liste prioritaire de métabolites à mesurer à partir d'une liste de gènes sélectionnés, en se basant sur les bases de données disponibles. De fait, dans certaines pathologies, seuls les loci qui leurs sont associé ont été identifiés par des GWAS. Afin de mieux comprendre ces pathologies, des études ont proposé d'intégrer plusieurs données moléculaires en lien avec la fonction de gènes sélectionnés à partir de ces loci associés. Les métabolites ont alors été proposés comme marqueurs moléculaires pouvant être mis en lien avec la fonction des gènes. Ainsi, étant donné qu'il n'est pas possible de mesurer tous les métabolites du métabolome, il a donc fallu sélectionner des métabolites à mesurer en lien avec la fonction des gènes.

Afin de générer cette liste de métabolites à mesurer, il est possible d'utiliser une approche basée sur une recherche manuelle de l'information contenue dans la littérature. Toutefois, cette recherche s'avère fastidieuse et subjective. Par conséquent, le développement d'une méthode basée sur une approche bio-informatique, laquelle permettrait de générer cette liste de métabolites de façon systématique en se basant sur les bases de données disponibles, apparaissait comme une solution judicieuse à ce problème. L'hypothèse de travail était que, considérant l'ensemble des voies métaboliques, les métabolites d'intérêt sont des substrats/produits se trouvant à proximité des réactions catalysées par l'enzyme encodée par le gène étudié. Étant donné qu'il n'existait pas d'outil permettant cette analyse, nous avons développé notre propre méthode. Celle-ci quantifie le lien entre gènes et métabolites par le nombre de réactions les séparant dans des voies métaboliques. Cette mesure, la distance, a été calculée dans des voies métaboliques de la base de données KEGG, plus particulièrement dans sa carte globale du métabolisme. Afin de valider cette méthode, nous l'avons appliquée à l'analyse des mGWAS bien que n'importe quelles listes de gènes et de métabolites peuvent être utilisées. Ces études rapportent des associations directes entre les variants génétiques, annotés en gènes, et les métabolites. Ces associations ont été utilisées afin de tester l'hypothèse de proximité de gènes et de métabolites et, de façon plus générale, afin de mieux comprendre le lien entre gènes et métabolites dans les voies métaboliques. La section suivante

présentera les choix effectués lors du développement de cette méthode ainsi que les postulats sous-jacents de celle-ci.

## 3.1 Considérations méthodologiques

### 3.1.1 Choix pour l'implémentation de la méthode

Initialement, des scripts ont été développés afin d'appliquer la méthode pour l'analyse des données mGWAS de Shin *et al.* (Shin et al., 2014). Le langage R a été choisi, car celui-ci est largement utilisé pour le développement de *package* bio-informatique, ce qui constitue un des objectifs visés à court terme pour cette nouvelle méthode. Il est commun de pouvoir obtenir le code de la méthode librement en ligne pour les articles de méthode de bio-informatique. De cette façon, il est facile de reproduire les résultats obtenus et le *package* pourra ainsi être utilisé pour d'autres données. Un autre langage, tel que Python, aurait aussi pu être choisi et aurait été tout aussi adéquat pour le développement.

#### 3.1.1.1 **Extraction des données KEGG**

La base de données choisie pour cette méthode était KEGG, et plus particulièrement, sa carte *Metabolic pathways - Reference pathway* (nommée *overview map* dans l'article et traduite par « carte globale » pour la suite du mémoire). L'avantage de cette base de données est qu'elle contient un serveur REST (*Representational State Transfer*) qui permet le téléchargement des données de KEGG grâce à des requêtes pouvant être comprises dans des scripts. De plus, l'information de la carte globale est encodée en langage XML (*Extensible Markup Language*) spécifique à KEGG (nommé *KEGG Markup Language* – KGML), ce qui permet une extraction facile des données. Cette information a été extraite grâce à des scripts développés de façon à obtenir l'ensemble des entités de chaque réaction de la carte globale. Seules les informations relatives aux entités pertinentes à la méthode ont été retenues, telles que les produits, les substrats, les réactions et les gènes. Néanmoins, les formats standards utilisés afin de simplifier l'échange des données des voies métaboliques sont SBML (*Systems Biology Markup Language*) ou BioPAX (*Biological Pathway Exchange*; Strömbäck & Lambrix, 2005). Le format KGML de KEGG n'utilisant pas ces standards, cela rend les scripts

permettant l'extraction des voies métaboliques uniquement applicables aux données de KEGG.

## 3.2 Analyse critique de la méthode

À la section 2.4 (p. 40), nous avons présenté la méthode et l'avons appliquée à des données de la littérature provenant d'une étude de mGWAS afin de valider notre hypothèse de travail tout en démontrant l'utilité de cette méthode. Cette section présentera une analyse critique de la méthode ainsi que des résultats obtenus par son application aux données du mGWAS par Shin *et al*. Cette section décrit les caractéristiques importantes de cette méthode ayant un impact sur les résultats obtenus lors de son application.

En premier lieu, il convient de souligner que seulement 24 des 299 associations gène-métabolite rapportées par Shin *et al.* ont été cartographiées sur la carte globale du métabolisme de KEGG. Nous avons néanmoins pu obtenir des résultats qui confirment l'hypothèse de travail formulée sur la relation entre gènes et métabolites associés.

### 3.2.1 Application de la méthode aux données de mGWAS

Les études mGWAS offrent un grand avantage, soit celui de combiner un très grand nombre de données génomiques et métabolomiques, permettant ainsi d'évaluer l'association entre les éléments du génome et du métabolome, mais aussi de mieux comprendre l'influence des gènes sur les niveaux des métabolites, considérés comme des phénotypes intermédiaires. À ce jour, il n'existe pas d'autres jeux de données à grande échelle faisant le lien direct entre la séquence génomique et les niveaux de métabolites chez des individus sains.

En somme, l'intérêt initial de cette étude résidait au niveau de la validation de la méthode. Par ailleurs, les développements de cette méthode ont du même coup permis l'évaluation des associations de Shin *et al*. et, plus globalement, le *package* permettra d'évaluer d'autres mGWAS.

### 3.2.2 Développement de la méthode

Nous avons décidé d'utiliser les voies métaboliques comme contexte biologique connu, car ces voies contiennent de l'information sur les métabolites ainsi que sur les gènes qui encodent les protéines ou enzymes qui catalysent les réactions métaboliques.

La section suivante présentera les considérations justifiant le choix de la base de données et de la carte métabolique comparativement aux voies métaboliques individuelles.

#### 3.2.2.1 Choix de la base de données et de la carte métabolique

Nous avons utilisé la base de données KEGG (http://www.genome.jp/kegg/) comme référence car elle s'avère la ressource la plus utilisée pour ses voies métaboliques (Bader, Cary, & Sander, 2006; Weber, Winder, Larcombe, Dunn, & Viant, 2015). KEGG contient aussi des fonctions de téléchargement, ce qui a facilité le développement de la méthode, ainsi que des outils de visualisation très utiles pour fins de vérification.

Tel que mentionné dans l'introduction de ce mémoire (section 1.4, p. 19), la définition et l'organisation des voies métaboliques varient considérablement selon la base de données utilisée, et ce même à l'intérieur d'une même base de données. Bien que d'autres voies métaboliques puissent être utilisées par le *package*, nous avons choisi d'utiliser la carte globale de KEGG pour cette application, car elle offre une couverture plus globale du fait qu'elle contient plusieurs voies métaboliques connectées les unes aux autres. Cette carte inclut un grand nombre de voies métaboliques intermédiaires classiques, dont le cycle de Krebs et la glycolyse, ainsi que les réactions les reliant.

Il convient de préciser que le nombre d'associations gène-métabolite cartographiées est limité par l'information comprise sur la carte métabolique utilisée. Par exemple, si cette carte est petite, seulement peu d'associations seront cartographiées, car il faut que le gène et son métabolite soient présents dans cette carte. Ainsi, le choix de la carte métabolique, en particulier sa taille (soit le nombre de réactions et de voies métaboliques la constituant), influencera grandement le nombre d'associations à partir desquelles il est possible de calculer des distances. Par exemple, si nous avions choisi une seule voie métabolique comme le cycle de Krebs, très peu d'associations auraient pu être cartographiées. En outre, même en utilisant

une grande carte métabolique, il est aussi attendu que des associations soient manquantes, puisque les connaissances actuelles du métabolisme humain sont encore incomplètes.

Plus particulièrement, les sections suivantes présentent les particularités et les difficultés rencontrées lors de l'identification et de la cartographie des gènes et des métabolites provenant des données de mGWAS de Shin *et al.*

*Identification et cartographie des gènes de Shin et al.*

Pour l'identification de gènes, il convient de souligner que les études mGWAS rapportent des associations entre loci et métabolites, et non entre gènes et métabolites. Par conséquent, la méthode utilisée par les auteurs afin d'annoter ces loci en gènes aura un impact sur les gènes à identifier. Pour l'étude par Shin *et al.,* cette annotation du SNP indexe (*lead*) d'un locus en gène est basée sur les connaissances actuelles de la fonction des protéines encodées par ces gènes. Cette méthode sélectionne les gènes codant pour des protéines à proximité du SNP, soit à moins de 500 kb sur le génome. On s'attend donc à ce que l'identification des gènes soit complète, puisque tous les loci sont annotés en gènes connus et devraient être compris dans la base de données choisie.

Pour la cartographie, la méthode d'annotation tente de prédire le gène causal selon les différents gènes à proximité. Cette prédiction est effectuée en utilisant le métabolite associé, les connaissances de la fonction de gènes rapportés dans la littérature, ainsi que les bases de données enzymatiques et métaboliques. De plus, il a été montré qu'une partie des loci qui ont été associés à des métabolites dans les études mGWAS peuvent être fonctionnellement reliés à des enzymes (Kastenmüller et al., 2015). Puisque l'annotation de ces loci en gènes est effectuée sur la base des connaissances biochimiques des fonctions des gènes et de son (ses) métabolite(s) associé(s), il est ainsi attendu qu'une grande partie des gènes soient annotés en gènes codant pour des enzymes. On s'attend donc à avoir une cartographie d'une grande partie de ces gènes dans une carte du métabolisme comprenant principalement des enzymes.

*Identification et cartographie des métabolites de Shin et al.*

Pour les études métabolomiques, le consortium COSMOS (COordination of Standards in MetabolomicS; Salek, Steinbeck, Viant, Goodacre, & Dunn, 2013) recommande aux

chercheurs de rapporter le nom commun des métabolites ainsi qu'un identifiant structurel afin d'éviter toute ambigüité. De fait, l'identification des métabolites sur la base de leur nom commun n'est pas fiable, car il existe plusieurs synonymes et variantes d'épellation pour un même métabolite (Hettne et al., 2009). Les identifiants structurels, tels que SMILES (S*implified Molecular-Input Line-Entry System*) et InChI (IUPAC International Chemical Identifier), sont obtenus par une méthode d'assignation d'identifiant généré selon la structure chimique, laquelle est unique à chaque métabolite (http://www.iupac.org/home/publications/e-resources/inchi.htm). Toutefois, peu d'études ont utilisé cette annotation à ce jour. Par conséquent, l'utilisation des données métabolomiques de la littérature peut être entravée par la difficulté à déterminer de façon automatique l'identité des métabolites rapportés.

L'article de Shin *et al.* rapporte seulement les noms communs des métabolites et des identifiants de la base de données de Metabolon Inc. (Durham, NC, USA). Or, cette compagnie privée, qui a effectué l'analyse métabolomique, ne donne pas un libre accès aux identifiants de sa base de données. Nous avons donc eu recours à une chimiste qui a manuellement identifié les métabolites selon leur nom commun dans la base de données KEGG, en s'assurant de la correspondance avec une autre base de données, soit HMDB.

Finalement, les analyses métabolomiques peuvent aussi inclure des signaux inconnus (nommés *unknowns*) pour lesquels aucun métabolite n'a été attribué à ce jour. Il va sans dire que ces signaux ne seront pas identifiés dans la base de données.

La section suivante présentera les détails de la couverture obtenue à partir des données de Shin *et al.* en utilisant la base de données KEGG (référé dans le texte comme l'identification) et sa carte globale (référé dans le texte comme la cartographie).

#### 3.2.2.2 **Résultats obtenus avec KEGG**

Lors de l'identification des gènes et des métabolites rapportés dans l'étude de Shin *et al*. dans KEGG, on a pu remarquer une disparité entre le nombre de gènes et de métabolites identifiés. En effet, tous les 132 gènes de la liste ont été identifiés. Cette identification complète s'explique par le fait que KEGG contient tous les gènes de la ressource RefSeq du NCBI (http://www.ncbi.nlm.nih.gov/refseq/). Il était donc attendu que les gènes de l'étude de Shin *et al.* aient un identifiant KEGG. Au contraire, seulement 69% des métabolites annotés

(229 sur 332) ont pu être identifiés (Table 1; section 2.7, p. 55). Ce faible pourcentage s'explique en partie par le fait que les bases de données comme KEGG incluent principalement des métabolites intracellulaires alors que les métabolites rapportés dans l'étude de Shin *et al.* ont été mesurés dans le plasma; ils sont donc extracellulaires. Les métabolites intracellulaires, tels que les acyl-CoA, ne sont pas transportés hors de la cellule et ne sont donc pas ou peu présents dans le sang. Pour cette raison, nous avons proposé une règle avec les acylcarnitines afin de combler cet écart, ce qui a permis d'identifier 10 composés de cette classe. En outre, nous avons aussi observé un biais quant à la diversité des classes de métabolites retrouvés ou non dans KEGG. À titre d'exemple, les classes de métabolites qui ne sont pas bien représentées dans KEGG sont surtout les peptides (15%) et les lipides (65%). À l'avenir, il faudra créer plus de règles afin d'augmenter la couverture des métabolites, un point qui sera discuté ultérieurement à la section 3.3.1.1 (p. 72). Ainsi, seulement 129 des 299 associations ont été identifiées dans KEGG.

En ce qui concerne la cartographie des gènes et des métabolites sur la carte globale de KEGG, même si tous les gènes ont été identifiés, cela n'impliquait pas nécessairement qu'ils encodaient pour des protéines ayant un impact sur le métabolisme. De fait, 49 des 132 gènes ont été cartographiés sur la carte globale de KEGG et sont donc des enzymes catalysant des réactions du métabolisme intermédiaire. Quant aux métabolites, seulement 98 des 229 identifiés ont été cartographiés, ce qui est faible sachant que 529 métabolites ont été mesurés et 197 étaient non-identifiés (*unknown*). La faible couverture des métabolites est encore plus apparente pour certaines classes de composés telles que les lipides, qui est la classe la plus représentée (130 des 333 métabolites annotés) alors que seulement 22% ont pu être cartographiés (Table 2; Section 2.7, p. 56). Or, il est reconnu que les lipides jouent un rôle essentiel dans le métabolisme au niveau de la cellule et de l'organisme, soit entre autres dans la formation de l'énergie par la β-oxydation et le stockage de l'énergie sous forme de triglycérides, des voies qui sont impliquées dans le métabolisme intermédiaire (Santos & Schulze, 2012). On peut donc affirmer que la couverture du métabolisme lipidique dans la carte globale de KEGG est faible comparativement à ce qui était attendu. D'autres classes de métabolites telles que les xénobiotiques (11%) et les peptides (0%) sont aussi très peu ou pas cartographiées, mais ceci était attendu puisque ces métabolites ne jouent pas un rôle

prépondérant dans le métabolisme intermédiaire. En résumé, on peut conclure qu'en général, la couverture de KEGG et de sa carte métabolique globale n'est pas optimale, et que lors d'analyses métabolomiques chez l'homme, il existe une grande disparité entre les métabolites mesurés dans des échantillons biologiques accessibles pour le prélèvement comme le sang ou l'urine, lesquels représentent le milieu extracellulaire, et les métabolites répertoriés dans les voies métaboliques de KEGG. De plus, on remarque qu'entre l'identification et la cartographie, une grande partie des gènes et des métabolites sont perdus. On peut donc supposer que la couverture pourrait être améliorée si les autres voies de KEGG étaient ajoutées.

En somme, 24 des 129 associations identifiées ont pu être cartographiées sur la carte globale.

*Les avantages de KEGG en comparaison avec les autres bases de données existantes*

Afin d'évaluer l'impact du choix de la base de données, nous nous sommes demandé s'il y avait des différences dans l'information répertoriée dans KEGG comparativement aux autres bases de données de voies métaboliques couramment utilisées. À cet égard, l'étude par Stobbe *et al.* (Stobbe, Houten, Jansen, van Kampen, & Moerland, 2011) avait effectué une comparaison approfondie des 5 bases de données métaboliques humaines les plus connues : KEGG, Edinburg Human Metabolic Newtork (EHMN; Ma et al., 2007), HumanCyc (Trupp et al., 2010), Recon1 (Duarte et al., 2007) et Reactome (Joshi-Tope et al., 2005).

En premier lieu, il est important de mentionner que les reconstructions métaboliques de EHMN et Recon1 sont différentes de KEGG d'un point de vue conceptuel. Le principe sous-jacent consiste à construire un réseau complet à ces reconstructions et d'extraire, pour un organisme choisi, toutes les réactions provenant de base de données de voies métaboliques. Par exemple, Recon1 est basée en partie sur les réactions chez l'humain identifiées dans KEGG. Puis, ces réactions sont analysées dans la perspective d'un réseau complet. Cela revient à collecter de l'information métabolique d'un organisme, puis compiler cette information dans un modèle mathématique, en considérant les propriétés des réactions telles que leurs produits/substrats/cofacteurs, leur directionnalité (le sens des réactions si elles sont irréversibles), leur localisation cellulaire, leur stœchiométrie, etc (Thiele & Palsson, 2010).

Ces reconstructions permettent l'utilisation de connaissances biologiques en format mathématique afin de répondre à une grande variété de questions scientifiques dont l'analyse des données « omiques », selon des états physiologiques, des contraintes chimiques ou génétiques (Oberhardt, Palsson, & Papin, 2009).

La comparaison de ces 5 bases de données par Stobbe *et al.* a été effectuée sur l'ensemble des voies métaboliques présentes chez l'humain et retrouvées dans chaque base de données, et pas seulement sur la carte globale. Nous nous sommes particulièrement intéressés à la comparaison au niveau des gènes et des métabolites de ces bases de données comparativement à KEGG. Le consensus est le nombre de données communes entre les 2 bases de données. Le consensus le plus grand de KEGG était de 61% avec Recon1 pour les gènes, et de 40% avec EHMN pour les métabolites. En comparant l'ensemble des organismes, une autre étude a montré que KEGG contient significativement plus de composés que la base de données Biocyc (Altman et al., 2013). Néanmoins, Biocyc contient davantage de réactions que KEGG.

Une explication possible pour les différences observées entre ces bases de données est que le nombre d'étapes intermédiaires utilisées afin de décrire une réaction varie selon la base de données (Stobbe et al., 2011). Certaines bases de données expliquent des réactions en une étape, tandis que d'autres les séparent en plusieurs réactions. Par exemple, la conversion du citrate en isocitrate (EC 4.2.1.3) est définie en une étape par Recon1, alors qu'il y en a deux dans HumanCyc et KEGG, où le cis-aconitate est décrit comme intermédiaire.

En outre, un avantage de KEGG comparativement aux autres bases de données est qu'elle contient de l'information sur les paires de substrats et produits, nommées *reactant pair* (RPAIR). Les RPAIRs ont été définies comme des paires de métabolites, un substrat et un produit, d'une réaction ayant des atomes ou des groupes d'atomes en commun (Kotera, Okuno, Hattori, Goto, & Kanehisa, 2004). Pour chaque paire de substrat et produit, le système assigne une classification RPAIR en utilisant une méthode d'alignement de structure chimique (Kotera et al., 2004). Les différentes classes de RPAIR retrouvées sont les suivantes : *main, trans, cofac, ligase,* et *leave* (Faust, Croes, & van Helden, 2009). De ces différentes classes, on définit les métabolites principaux (*main*) comme les acteurs principaux des réactions, lesquels sont modifiés dans une série de réactions d'une voie métabolique donnée. Ces métabolites

forment le « squelette » (Faust et al., 2009) des voies métaboliques comparativement aux autres métabolites qui agissent comme co-substrats ou cofacteurs (nommés *cofac*). Nous avons donc utilisé KEGG car elle facilitait la reconstruction de réseaux biochimiques par l'utilisation de ses métabolites principaux. Cette observation avait aussi été montrée par Faust *et al.* (Faust et al., 2009). De plus, cette information sur les métabolites principaux n'est pas retrouvée dans d'autres bases de données telles que EHMN, Recon1 (Stobbe, Jansen, Moerland, & van Kampen, 2014) et HumanCyc (Faust et al., 2009). Ainsi, pour ces bases de données, il aurait fallu développer des stratégies afin d'identifier les métabolites principaux, ce qui peut être fastidieux (Stobbe et al., 2014).

Selon certains auteurs, les méthodes bio-informatiques basées sur les bases de données métaboliques sont limitées par les connaissances reliées à l'organisme d'intérêt (Grapov, Fahrmann, et al., 2015). Pour la méthode qui utilise la base de données KEGG chez l'humain, on a pu remarquer que les connaissances répertoriées pour certaines classes de métabolites, dont les lipides, étaient limitées. En revanche, d'autres bases de données, telles que EHMN, mettent l'emphase sur le métabolisme des lipides afin d'obtenir une plus grande couverture lipidique (Hao et al., 2010).

En résumé, on peut affirmer que comparativement à KEGG, l'information retrouvée dans les autres bases de données n'est pas la même, et qu'il est donc attendu que le choix de la base de données pour la méthode ait un impact sur les résultats obtenus. Néanmoins, KEGG avait l'avantage d'être un bon modèle pour le développement de la méthode en raison de sa carte globale du métabolisme et de sa définition des métabolites principaux pour chaque réaction.

#### 3.2.2.3 **Choix de mesure pour l'évaluation des associations gène-métabolite**

La présente section discutera du choix de la mesure utilisée pour la méthode, soit la distance, laquelle représente le nombre de réactions entre les gènes et les métabolites impliqués dans une voie métabolique donnée. Ce calcul de distance permet une évaluation objective, quantitative et rapide des associations des mGWAS, d'autant plus que la quantité de données à analyser est sans cesse en croissance. Le fait d'utiliser une valeur quantitative nous a aussi permis de développer un test statistique permettant de répondre à la question : les

paires de gène-métabolite associées sont-elles significativement plus rapprochées que les paires non associées dans la carte du globale du métabolisme ?

Par ailleurs, il a été rapporté que les métabolites ou ratio de métabolites associés à un gène étaient souvent le produit ou le substrat de la réaction catalysée par l'enzyme encodée par ce gène (Kastenmüller et al., 2015) ou d'une réaction à proximité, soit en amont ou en aval (Adamski, 2012). Le principe de compter les étapes, soit le nombre de réactions entre les métabolites et les gènes existe mais à ce jour, ceci a été effectué de façon manuelle et seulement dans les cas où les métabolites étaient à une courte distance.

Dans le cas des gènes qui encodent des enzymes qui catalysent plusieurs réactions, plusieurs distances de ces différentes réactions peuvent être calculées jusqu'au métabolite associé. En supposant que la distance la plus courte reflèterait une fonction plus vraisemblable, il est alors possible d'identifier, à travers plusieurs réactions, celle contrôlant les niveaux du métabolite associé.

#### 3.2.2.4 **Les avantages de notre méthode en comparaison avec les méthodes existantes**

À la section 2.5.5 (p. 46), nous avons brièvement comparé notre méthode avec deux autres qui, au meilleur de notre connaissance, étaient les seules disponibles pour l'analyse des données de mGWAS dans un contexte biologique. D'emblée, il est important de souligner que l'objectif des deux méthodes, soit *GGM* et *mining the unknown,* est différent de celui de notre méthode, ce qui rend toute comparaison plus complexe. Toutefois, pour les fins de cette discussion, nous aborderons seulement *GGM*, qui était celle utilisée dans l'article de Shin *et al*., puisque *mining the unknown* a pour but d'identifier les métabolites « inconnus », lesquels n'ont pas été considérés dans notre méthode. La prémisse sous-jacente à *GGM* est que la corrélation entre les métabolites permet de reconstruire des voies métaboliques. Dans Shin *et al.*, le réseau GGM est d'abord généré grâce aux corrélations des données obtenues par l'analyse métabolomique. Puis, les gènes associés aux métabolites sont ajoutés au réseau. De cette façon, les auteurs ont obtenu un réseau (nommé « atlas ») récapitulant les relations entre les métabolites et les gènes associés. Cette approche est donc basée sur le regroupement de l'ensemble des associations et des connaissances extraites grâce au GGM afin de reconstruire l'activité métabolique. En outre, cet atlas fournit de l'information sur les voies métaboliques

provenant de KEGG. Toutefois, cette information est qualitative, et seulement le nom d'une voie métabolique dont font partie les métabolites, et non les gènes, est affiché. La cartographie des métabolites n'étant pas disponible, on ne connaît donc pas la localisation des métabolites dans des voies métaboliques. On peut donc affirmer que comparativement à notre méthode, GGM est non seulement utilisé de façon différente, mais il est aussi basé sur des principes différents tout en n'étant pas quantitatif.

En conséquence, la méthode que nous avons développée permet d'évaluer les associations de façon individuelle dans des voies métaboliques, ce qui n'était pas possible avec les méthodes précédemment développées.

## 3.3 Perspectives futures

À court terme, les perspectives futures consistent à parachever le développement du *package* de cette méthode, qui sera ajouté à l'article lors de sa soumission. La section suivante propose des avenues de travaux futurs afin d'améliorer la performance de la méthode développée ainsi que d'autres applications potentielles de cette méthode au-delà de l'analyse des données mGWAS.

### 3.3.1 Développements futurs

#### 3.3.1.1 Augmenter la couverture

Une des limites de la méthode proposée est que plusieurs associations gène-métabolite rapportées dans l'étude de Shin *et al.* n'ont pas été cartographiées sur la carte métabolique (seulement 24 des 129 associations rapportées identifiées dans KEGG). Ici, nous proposons des approches afin d'augmenter le nombre d'associations cartographiées.

Tel que mentionné plus tôt, il a été montré que certains loci associés à des métabolites dans les mGWAS pouvaient aussi être fonctionnellement liés à un transporteur (Kastenmüller et al., 2015). Or, les transporteurs ne sont pas inclus dans la base de données KEGG. Ainsi, pour augmenter la couverture génétique de notre méthode, il serait intéressant d'y ajouter les transporteurs. Par exemple, dans l'étude de Shin *et al*., SLC7A5, qui encode pour un transporteur d'acides aminés, est associé à la kynurénine, un métabolite issu de la

transformation de l'acide aminé essentiel tryptophane, mais n'avait pas été cartographié par notre méthode.

Alors que chaque base de données a des avantages et des inconvénients, il serait intéressant d'implémenter les principes de la méthode développée dans d'autres bases de données utilisant le principe des reconstructions métaboliques (Sévin et al., 2015). À cet égard, nous proposons la dernière version de Recon (Recon2) (Thiele et al., 2013), car cette base de données offre la représentation du métabolisme humain obtenue grâce à des outils de modélisation. Recon2 est basé sur Recon1 qui a été modifiée pour ajouter EHMN ainsi que les réactions d'autres bases de données et de reconstructions métaboliques. En utilisant Recon2, on s'attend à un large réseau métabolique, ce qui permettrait d'obtenir une meilleure couverture comparativement à la carte globale de KEGG. De plus, nous nous attendons à avoir davantage d'associations cartographiées sur Recon2, car en plus des métabolites extracellulaires dont les acylcarnitines (Sahoo, Franzson, Jonsson, & Thiele, 2012), absents de la carte globale dans KEGG, elle contient également les réactions de transport. Par exemple, une comparaison entre Recon2 et la carte globale de KEGG révèle que Recon2 contient davantage de réactions (1789 contre 1237), de gènes (7440 contre 2407) et de métabolites (2626 contre 1640).

En outre, Recon2 permettrait également une amélioration de l'aspect topologique du réseau, puisqu'il s'agit d'une représentation plus réaliste du métabolisme étant donné que cette base de données inclut les différents compartiments cellulaires. De plus, Recon2 est construit de façon à combler les réactions manquantes (nommées trous ou *gaps*) de son réseau afin que celui-ci soit plus complet. En effet, ces étapes sont effectuées par plusieurs itérations où des recherches ciblées de la littérature sont menées afin d'identifier les réactions manquantes (Thiele & Palsson, 2010). En diminuant le nombre de trous, cela permettrait de réduire le nombre de distances infinies obtenues avec notre méthode.

Bien que Recon2 pourrait offrir une meilleure couverture, KEGG offre l'avantage d'être plus largement répandue et d'une simplicité d'utilisation plus adaptée au développement initial de la méthode. De fait, KEGG contient des outils de visualisation qui ont permis la validation initiale des distances calculées ainsi qu'une définition des réactions principales, ce qui n'étaient pas été possible avec Recon2. Il serait par la suite intéressant de comparer les

résultats obtenus utilisant Recon2 avec ceux obtenus avec KEGG, ce qui permettrait aussi une double validation des distances obtenues.

#### 3.3.1.2 Intégrer la topologie des réactions métaboliques

De la même manière, il serait pertinent d'ajouter plus d'information sur la position des métabolites associés aux gènes, c'est-à-dire s'ils se trouvent en amont ou en aval de la réaction catalysée par l'enzyme encodée par ce gène. Cela expliquerait mieux l'effet du variant du gène sur les métabolites. De plus, il conviendrait d'ajouter aussi la directionnalité des réactions, lesquelles peuvent être réversibles ou irréversibles. Cette directionnalité peut être prédite à partir des données thermodynamiques ou des connaissances acquises sur les réactions d'intérêt (Duarte et al., 2007).

Finalement, lors de l'analyse des résultats de Shin *et al.*, nous avons décidé de séparer les ratios de métabolites rapportés en deux associations distinctes. Les auteurs suggèrent que les ratios reflètent les propriétés cinétiques des enzymes, ce qui a par la suite été remis en cause par Mootha *et al.* (Mootha & Hirschhorn, 2010). Il serait donc intéressant d'intégrer les ratios à la méthode afin de vérifier si ceux-ci sont en effet à proximité de la même réaction. De plus, ceci permettrait de déterminer la position des deux métabolites du ratio, l'un par rapport à l'autre, à savoir s'ils sont tous les deux du même côté d'une réaction donnée ou de part et d'autre.

### 3.3.2 Applications futures

La méthode présentée dans ce mémoire a été optimisée pour l'analyse des données d'association entre gène et métabolite rapportées dans les études mGWAS. Ces données étaient les seules disponibles dans la littérature nous permettant de tester notre hypothèse. Toutefois, on peut envisager d'appliquer la méthode à d'autres jeux de données afin d'étudier le lien entre métabolite et gène, outre les SNPs, dans le contexte de voies métaboliques.

#### 3.3.2.1 Prédire les métabolites à mesurer à partir d'une liste de gènes

L'application aux données mGWAS nous a permis de mieux comprendre et définir les associations gène-métabolite. Toutefois, l'objectif initial de ce projet était de prédire quels

métabolites devraient être ciblés lors d'analyse métabolomique en donnant comme entrée une liste de gènes associés au risque de développer des maladies. Pour ce faire, il faudrait trouver une distance « universelle » qui permettrait une priorisation des métabolites d'intérêt à tester. Une façon d'y parvenir serait d'utiliser les données de mGWAS comme ensemble d'apprentissage afin d'obtenir cette distance. Cet ensemble contient de l'information sur les gènes et les métabolites d'intérêt (les associations), ce qui permettrait de proposer une distance qui considérerait la sensibilité et la spécificité des métabolites trouvés. Par la suite, cette distance pourrait être appliquée sur les données où seuls les gènes sont connus, et où il faut prédire les métabolites à mesurer, ce qui constituait l'objectif initial de ce travail. Pour les gènes sélectionnés à partir des 163 loci associés aux maladies inflammatoires de l'intestin, cette distance pourrait être utilisée afin de proposer des métabolites à mesurer lors de l'analyse métabolomique.

À l'avenir, l'établissement d'une distance permettant la prédiction de métabolites sera utile afin d'évaluer l'impact des gènes associés à une maladie et permettra, de façon plus générale, une meilleure compréhension de la maladie et de son diagnostic. De plus, une récente étude (Guo et al., 2015) a souligné l'utilité de combiner données métabolomiques et génomiques en médecine personnalisée afin d'améliorer l'interprétation médicale des risques de développer des maladies pour chaque individu d'une cohorte clinique. Pour certains patients, les auteurs de cette étude ont identifié des anomalies métaboliques comme signes précoces de développement d'un diabète, associées au dysfonctionnement du foie et en lien avec des perturbations de l'homéostasie du microbiome de l'intestin. La méthode utilisée étant basée sur la priorisation des gènes grâce à l'analyse des voies métaboliques issues des données métabolomiques, il serait donc possible d'utiliser le calcul de distance afin d'affiner cette priorisation. Les résultats de cette étude démontrent que la métabolomique peut être une approche efficace pour complémenter le séquençage de nouvelle génération afin prédire le risque de développer des maladies.

# Conclusion

Ce mémoire présente une méthode basée sur une approche bio-informatique utilisant la base de données métabolique la plus communément utilisée, KEGG. Cette méthode permet de lier les gènes et les métabolites en calculant le nombre de réactions les séparant sur les voies métaboliques de KEGG. De cette façon, on obtient une valeur quantitative, soit la distance entre un gène et un métabolite. Notre hypothèse était que les métabolites d'intérêt pour un gène donné seraient des substrats/produits se trouvant à proximité des réactions catalysées par l'enzyme encodée par ce gène. Afin de tester cette hypothèse, nous avons orienté le développement de la méthode pour l'analyse des données mGWAS. Plus particulièrement, afin de valider la méthode, nous avons utilisé les données génomiques et métabolomiques de l'étude mGWAS de Shin *et al.* effectuée chez l'homme. De plus, nous avons inclus une analyse statistique, ce qui a permis de tester si les distances entre les gènes et les métabolites associés dans cette étude étaient significativement plus petites que celles obtenues entre des gènes et des métabolites sélectionnés de façon aléatoire. Le test étant statistiquement significatif, nous avons pu confirmer l'hypothèse de proximité entre gènes et métabolites. Néanmoins, seulement 24 des 299 associations ont été cartographiées sur la carte globale du métabolisme de KEGG, soulignant ainsi la faible couverture génétique et plus particulièrement métabolique de cette carte et de la base de données. Bien que nous ayons intégré une règle faisant le lien entre certains métabolites extracellulaires mesurés et leurs équivalents intracellulaires rapportés dans KEGG, 57% (98/229) des métabolites n'ont pas pu être identifiés dans KEGG, limitant ainsi le nombre d'associations cartographiées.

Notre méthode est la première à utiliser la cartographie des voies métaboliques pour l'analyse systématique des associations gène-métabolite rapportées dans les études mGWAS. À ce jour, l'interprétation de ces association s'effectuait par la fouille manuelle de la littérature et en utilisant les quelques méthodes bio-informatiques disponibles. La méthode que nous avons développée permet l'interprétation de façon systématique et quantitative des associations gène-métabolite révélées pas les études mGWAS, lesquelles sont ainsi évaluées rapidement pour leur pertinence au réseau métabolique d'intérêt. À l'avenir, bien que la méthode offre des perspectives prometteuses pour l'interprétation objective des mGWAS et

qu'elle pourrait être utilisée avec d'autres voies métaboliques de KEGG, il conviendrait d'en augmenter la couverture métabolique en utilisant d'autres bases de données. Nous proposons Recon2 car elle offre une approche par reconstruction métabolique, et contient plus d'information génétique et métabolique que la carte globale de KEGG. En outre, bien que toutes ces bases de données représentent les connaissances actuelles du métabolisme, il est attendu que les développements futurs de ces bases de données permettront de diminuer l'écart entre le nombre de métabolites répertoriés et mesurés dans un système biologique donné. Plus généralement, la méthode développée et son application aux données d'une étude mGWAS ont permis une meilleure compréhension de la pertinence biochimique des associations gène-métabolite rapportées, mais également de l'influence des gènes sur le métabolisme. Enfin, cette méthode pourrait aussi servir à prédire quels métabolites devraient être ciblés lors d'analyses métabolomiques à partir d'une liste de gènes associés à un phénotype donné.

# Bibliographie


Adamski, J. (2012). Genome-wide association studies with metabolomics. *Genome Medicine*, *4*, 34.

Adamski, J., & Suhre, K. (2013). Metabolomics platforms for genome wide association studies--linking the genome to the metabolome. *Current Opinion in Biotechnology*, *24*, 39–47.

Adourian, A., Jennings, E., Balasubramanian, R., Hines, W. M., Damian, D., Plasterer, T. N., … Schuppe-Koistinen, I. (2008). Correlation network analysis for data integration and biomarker selection. *Molecular bioSystems*, *4*, 249–59.

Allan, J. D., Cusworth, D. C., Dent, C. E., & Wilson, V. K. (1958). A disease, probably hereditary characterised by severe mental deficiency and a constant gross abnormality of aminoacid metabolism. *Lancet*, *1*, 182–187.

Altman, T., Travers, M., Kothari, A., Caspi, R., & Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, *14*, 112.

Bader, G. D., Cary, M. P., & Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Research*, *34*, D504–6.

Baker, M. (2011). Metabolomics: from small molecules to big ideas. *Nature Methods*, *8*, 117–121.

Barallobre-Barreiro, J., Chung, Y.-L., & Mayr, M. (2013). Proteomics and metabolomics for mechanistic insights and biomarker discovery in cardiovascular disease. *Revista Española de Cardiología (English Edition)*, *66*, 657–61.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics*, *40*, 340–5.

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74.

Cacciatore, S., & Loda, M. (2015). Innovation in metabolomics to improve personalized healthcare. *Annals of the New York Academy of Sciences*, *1346*, 57–62.

Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., … Karp, P. D. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, *40*, D742–53.

Cavill, R., Kamburov, A., Ellis, J. K., Athersuch, T. J., Blagrove, M. S. C., Herwig, R., … Keun, H. C. (2011). Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Computational Biology*, *7*, e1001113.

Chan, E. K. F., Rowe, H. C., Hansen, B. G., & Kliebenstein, D. J. (2010). The complex genetic architecture of the metabolome. *PLoS Genetics*, *6*, e1001198.

Clegg, A. B., & Shepherd, A. J. (2008). Text mining. *Methods in Molecular Biology*, *453*, 471–91.

Cottret, L., Wildridge, D., Vinson, F., Barrett, M. P., Charles, H., Sagot, M.-F., & Jourdan, F. (2010). MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Research*, *38*, W132–7.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*, 561–3.

de Carvalho, L. P. S., Zhao, H., Dickinson, C. E., Arango, N. M., Lima, C. D., Fischer, S. M., … Rhee, K. Y. (2010). Activity-based metabolomic profiling of enzymatic function: identification of Rv1248c as a mycobacterial 2-hydroxy-3-oxoadipate synthase. *Chemistry & Biology*, *17*, 323–32.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, *26*, 51–78.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., … Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, *104*, 1777–82.

Dumas, M.-E. (2012). Metabolome 2.0: quantitative genetics and network biology of metabolic phenotypes. *Molecular bioSystems*, *8*, 2494–502.

Dunn, O. J. (1959). Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, *30*, 192–197.

Ewald, J. C., Matt, T., & Zamboni, N. (2013). The integrated response of primary metabolites to gene deletions and the environment. *Molecular bioSystems*, *9*, 440–6.

Faust, K., Croes, D., & van Helden, J. (2009). Metabolic pathfinding using RPAIR annotation. *Journal of Molecular Biology*, *388*, 390–414.

Fiehn, O. (2002). Metabolomics--the link between genotypes and phenotypes. *Plant Molecular Biology*, *48*, 155–71.

Fiehn, O., Barupal, D. K., & Kind, T. (2011). Extending biochemical databases by metabolomic surveys. *The Journal of Biological Chemistry*, *286*, 23637–43.

Friolet, R., Hoppeler, H., & Krähenbühl, S. (1994). Relationship between the coenzyme A and the carnitine pools in human skeletal muscle at rest and after exhaustive exercise under normoxic and acutely hypoxic conditions. *The Journal of Clinical Investigation*, *94*, 1490–5.

Galperin, M. Y., & Koonin, E. V. (2004). "Conserved hypothetical" proteins: prioritization of targets for experimental study. *Nucleic Acids Research*, *32*, 5452–63.

Garrod, A. E. (1923). *Inborn Errors of Metabolism*. Oxford University Press.

German, J. B., Hammock, B. D., & Watkins, S. M. (2005). Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, *1*, 3–9.

Gieger, C., Geistlinger, L., Altmaier, E., Hrabé de Angelis, M., Kronenberg, F., Meitinger, T., … Suhre, K. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genetics*, *4*, e1000282.

Gligorijević, V., & Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society*, *12*, 20150571–.

Goodacre, R. (2005). Metabolomics – the way forward. *Metabolomics*, *1*, 1–2.

Grapov, D., Fahrmann, J., & Wanichthanarak, K. (2015). Genomic, Proteomic, and Metabolomic Data Integration Strategies. *Biomarker Insights*, *10*, 1–5.

Grapov, D., Wanichthanarak, K., & Fiehn, O. (2015). MetaMapR: pathway independent metabolomic network analysis incorporating unknowns. *Bioinformatics*, *31*, 2757–60.

Gu, P., & Chen, H. (2014). Modern bioinformatics meets traditional Chinese medicine. *Briefings in Bioinformatics*, *15*, 984–1003.

Guo, L., Milburn, M. V., Ryals, J. A., Lonergan, S. C., Mitchell, M. W., Wulff, J. E., … Caskey, C. T. (2015). Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. *Proceedings of the National Academy of Sciences*, *112*, 201508425.

Hao, T., Ma, H.-W., Zhao, X.-M., & Goryanin, I. (2010). Compartmentalization of the Edinburgh Human Metabolic Network. *BMC Bioinformatics*, *11*, 393.

Hettne, K. M., Stierum, R. H., Schuemie, M. J., Hendriksen, P. J. M., Schijvenaars, B. J. A., Mulligen, E. M. van, … Kors, J. A. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, *25*, 2983–91.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, *6*, 95–108.

Hood, L., Heath, J. R., Phelps, M. E., & Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science*, *306*, 640–3.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*, e1000529.

Inouye, M., & Abraham, G. (2013). Look, no hands! Spectral biomarkers from genetic association studies. *Genome Medicine*, *5*, 14.

International HapMap Consortium. (2003). The International HapMap Project. *Nature*, *426*, 789–96.

Irizarry, R. A., Wang, C., Zhou, Y., & Speed, T. P. (2009). Gene set enrichment analysis made simple. *Statistical Methods in Medical Research*, *18*, 565–75.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., … Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, *33*, D428–32.

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., … Cho, J. H. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, *491*, 119–24.

Kaddurah-Daouk, R., Baillie, R. A., Zhu, H., Zeng, Z.-B., Wiest, M. M., Nguyen, U. T., … Krauss, R. M. (2011). Enteric microbiome metabolites correlate with response to simvastatin treatment. *PloS One*, *6*, e25482.

Kaddurah-Daouk, R., Kristal, B. S., & Weinshilboum, R. M. (2008). Metabolomics: a global biochemical approach to drug response and disease. *Annual Review of Pharmacology and Toxicology*, *48*, 653–83.

Kamburov, A., Cavill, R., Ebbels, T. M. D., Herwig, R., & Keun, H. C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, *27*, 2917–8.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*, 27–30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, *40*, D109–14.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A., & Krummenacker, M. (1997). EcoCyc: Enyclopedia of Escherichia coli Genes and Metabolism. *Nucleic Acids Research*, *25*, 43–50.

Kastenmüller, G., Raffler, J., Gieger, C., & Suhre, K. (2015). Genetics of human metabolism: an update. *Human Molecular Genetics*, *24*, 93–101.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., … Tsui, L. C. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science*, *245*, 1073–80.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, *295*, 1662–4.

Kotera, M., Okuno, Y., Hattori, M., Goto, S., & Kanehisa, M. (2004). Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *Journal of the American Chemical Society*, *126*, 16487–98.

Krumsiek, J., Suhre, K., Evans, A. M., Mitchell, M. W., Mohney, R. P., Milburn, M. V, … Kastenmüller, G. (2012). Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genetics*, *8*, e1003005.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, *5*, 5–21.

Kuo, T.-C., Tian, T.-F., & Tseng, Y. J. (2013). 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Systems Biology*, *7*, 64–78.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Larrouy-Maumus, G., Biswas, T., Hunt, D. M., Kelly, G., Tsodikov, O. V, & de Carvalho, L. P. S. (2013). Discovery of a glycerol 3-phosphate phosphatase reveals glycerophospholipid polar head recycling in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, *110*, 11320–5.

Lesk, A. (2012). *Introduction to Genomics*. Oxford University Press.

Liesenfeld, D. B., Habermann, N., Owen, R. W., Scalbert, A., & Ulrich, C. M. (2013). Review of mass spectrometry-based metabolomics in cancer research. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *22*, 2182–201.

Lindon, J. C., Nicholson, J. K., Holmes, E., & Everett, J. R. (2000). Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance*, *12*, 289–320.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., & Goryanin, I. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology*, *3*, 135.

Mal, M., Koh, P. K., Cheah, P. Y., & Chan, E. C. Y. (2012). Metabotyping of human colorectal cancer using two-dimensional gas chromatography mass spectrometry. *Analytical and Bioanalytical Chemistry*, *403*, 483–93.

Mamas, M., Dunn, W. B., Neyses, L., & Goodacre, R. (2011). The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of Toxicology*, *85*, 5–17.

Mangravite, L. M., Wilke, R. A., Zhang, J., & Krauss, R. M. (2008). Pharmacogenomics of statin response. *Current Opinion in Molecular Therapeutics*, *10*, 555–61.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics : TIG*, *24*, 133–41.

McClellan, J., & King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, *141*, 210–7.

Medina-Cleghorn, D., & Nomura, D. K. (2014). Exploring metabolic pathways and regulation through functional chemoproteomic and metabolomic platforms. *Chemistry & Biology*, *21*, 1171–84.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*, 31–46.

Miller, M. J., Kennedy, A. D., Eckhart, A. D., Burrage, L. C., Wulff, J. E., Miller, L. A. D., … Elsea, S. H. (2015). Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. *Journal of Inherited Metabolic Disease*, *38*, 1029–39.

Mootha, V. K., & Hirschhorn, J. N. (2010). Inborn variation in metabolism. *Nature Genetics*, *42*, 97–8.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., … Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, *34*, 267–73.

Nachman, M. W. (2001). Single nucleotide polymorphisms and recombination rate in humans. *Trends in Genetics*, *17*, 481–485.

Nam, D., & Kim, S.-Y. (2008). Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, *9*, 189–97.

Oberhardt, M. A., Palsson, B. Ø., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*, 320–35.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *27*, 29–34.

Oliver, S. (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, *16*, 373–378.

Orth, J. D., & Palsson, B. Ø. (2010). Systematizing the generation of missing metabolic knowledge. *Biotechnology and Bioengineering*, *107*, 403–12.

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews. Molecular Cell Biology*, *13*, 263–9.

Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA*, *299*, 1335–44.

Plata, G., Fuhrer, T., Hsiao, T.-L., Sauer, U., & Vitkup, D. (2012). Global probabilistic annotation of metabolic networks enables enzyme discovery. *Nature Chemical Biology*, *8*, 848–54.

Posada-Ayala, M., Zubiri, I., Martin-Lorenzo, M., Sanz-Maroto, A., Molero, D., Gonzalez-Calero, L., … Alvarez-Llamas, G. (2014). Identification of a urine metabolomic signature in patients with advanced-stage chronic kidney disease. *Kidney International*, *85*, 103–11.

Prosser, G. A., Larrouy-Maumus, G., & de Carvalho, L. P. S. (2014). Metabolomic strategies for the identification of new enzyme functions and metabolic pathways. *EMBO Reports*, *15*, 657–69.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–75.

Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., … Oliver, S. G. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, *19*, 45–50.

Ragoussis, J. (2009). Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics*, *10*, 117–33.

Reaves, M. L., Young, B. D., Hosios, A. M., Xu, Y.-F., & Rabinowitz, J. D. (2013). Pyrimidine homeostasis is accomplished by directed overflow metabolism. *Nature*, *500*, 237–41.

Reo, N. V. (2002). NMR-based metabolomics. *Drug and Chemical Toxicology*, *25*, 375–82.

Richards, S. E., Dumas, M.-E., Fonville, J. M., Ebbels, T. M. D., Holmes, E., & Nicholson, J. K. (2010). Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework. *Chemometrics and Intelligent Laboratory Systems*, *104*, 121–131.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews. Genetics*, *16*, 85–97.

Rolin, D. (2012). *Metabolomics Coming of Age with its Technological Diversity*. Academic Press.

Roux, A., Thévenot, E. A., Seguin, F., Olivier, M.-F., & Junot, C. (2014). Impact of collection conditions on the metabolite content of human urine samples as analyzed by liquid chromatography coupled to mass spectrometry and nuclear magnetic resonance spectroscopy. *Metabolomics*, *11*, 1095–1105.

Rueedi, R., Ledda, M., Nicholls, A. W., Salek, R. M., Marques-Vidal, P., Morya, E., … Kutalik, Z. (2014). Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links. *PLoS Genetics*, *10*, e1004132.

Saghatelian, A., & Cravatt, B. F. (2005). Global strategies to integrate the proteome and metabolome. *Current Opinion in Chemical Biology*, *9*, 62–8.

Saghatelian, A., Trauger, S. A., Want, E. J., Hawkins, E. G., Siuzdak, G., & Cravatt, B. F. (2004). Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry*, *43*, 14332–9.

Sahoo, S., Franzson, L., Jonsson, J. J., & Thiele, I. (2012). A compendium of inborn errors of metabolism mapped onto the human metabolic network. *Molecular bioSystems*, *8*, 2545–58.

Salek, R. M., Steinbeck, C., Viant, M. R., Goodacre, R., & Dunn, W. B. (2013). The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience*, *2*, 13–16.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., … Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, *265*, 687–95.

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*, 441–8.

Santos, C. R., & Schulze, A. (2012). Lipid metabolism in cancer. *The FEBS Journal*, *279*, 2610–23.

Saudubray, J.-M., Berghe, G. van den, & Walter, J. H. (2011). *Inborn Metabolic Diseases: Diagnosis and Treatment* (Vol. 16). Springer Science & Business Media.

Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., … Schomburg, D. (2013). BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Research*, *41*, 764–72.

Schulze, A., Lindner, M., Kohlmüller, D., Olgemöller, K., Mayatepek, E., & Hoffmann, G. F. (2003). Expanded newborn screening for inborn errors of metabolism by electrospray ionization-tandem mass spectrometry: results, outcome, and implications. *Pediatrics*, *111*, 1399–406.

Sévin, D. C., Kuehne, A., Zamboni, N., & Sauer, U. (2015). Biological insights through nontargeted metabolomics. *Current Opinion in Biotechnology*, *34*, 1–8.

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., … Soranzo, N. (2014). An atlas of genetic influences on human blood metabolites. *Nature Genetics*, *46*, 543–50.

Shulaev, V. (2006). Metabolomics technology and bioinformatics. *Briefings in Bioinformatics*, *7*, 128–39.

Silverman, E. K., & Loscalzo, J. (2012). Network medicine approaches to the genetics of complex diseases. *Discovery Medicine*, *14*, 143–52.

Skiena, S. S. (2008). *The Algorithm Design Manual*. Springer London.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, *78*, 779–87.

Smith, R., Ventura, D., & Prince, J. T. (2014). Controlling for confounding variables in MS-omics protocol: why modularity matters. *Briefings in Bioinformatics*, *15*, 768–70.

Stobbe, M. D., Houten, S. M., Jansen, G. A., van Kampen, A. H. C., & Moerland, P. D. (2011). Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology*, *5*, 165.

Stobbe, M. D., Jansen, G. A., Moerland, P. D., & van Kampen, A. H. C. (2014). Knowledge representation in metabolic pathway databases. *Briefings in Bioinformatics*, *15*, 455–70.

Strömbäck, L., & Lambrix, P. (2005). Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, *21*, 4401–7.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., & Ebert, B. L. (2005). Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proceedings of the National Academy of Sciences*, *43*, 15545–50.

Sugimoto, M., Kawakami, M., Robert, M., Soga, T., & Tomita, M. (2012). Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. *Current Bioinformatics*, *7*, 96–108.

Suhre, K., & Gieger, C. (2012). Genetic variation in metabolic phenotypes: study designs and applications. *Nature Reviews. Genetics*, *13*, 759–69.

Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohney, R. P., Meredith, D., Wägele, B., … Gieger, C. (2011). Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, *477*, 54–60.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, *22*, 281–5.

Teusink, B., Walsh, M. C., van Dam, K., & Westerhoff, H. V. (1998). The danger of metabolic pathways with turbo design. *Trends in Biochemical Sciences*, *23*, 162–169.

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*, 93–121.

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., … Palsson, B. Ø. (2013). A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, *31*, 419–25.

Tilton, S. C., Matzke, M. M., Sowa, M. B., Stenoien, D. L., Weber, T. J., Morgan, W. F., & Waters, K. M. (2015). Data integration reveals key homeostatic mechanisms following low dose radiation exposure. *Toxicology and Applied Pharmacology*, *285*, 1–11.

Tipton, K., & Boyce, S. (2000). History of the enzyme nomenclature system. *Bioinformatics*, *16*, 34–40.

Tolstoi, L. G., & Smith, C. L. (1999). Human Genome Project and cystic fibrosis--a symbiotic relationship. *Journal of the American Dietetic Association*, *99*, 1421–7.

Trupp, M., Altman, T., Fulcher, C. A., Caspi, R., Krummenacker, M., Paley, S., & Karp, P. D. (2010). Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. *Genome Biology*, *11*, 12.

Trupp, M., Zhu, H., Wikoff, W. R., Baillie, R. A., Zeng, Z.-B., Karp, P. D., … Kaddurah-Daouk, R. (2012). Metabolomics reveals amino acids contribute to variation in response to simvastatin treatment. *PloS One*, *7*, e38386.

Valeri, C., Pozzilli, P., & Leslie, D. (2004). Glucose control in diabetes. *Diabetes/metabolism Research and Reviews*, *20 Suppl 2*, S1–8.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*, 1304–51.

Vorkas, P. A., Isaac, G., Anwar, M. A., Davies, A. H., Want, E. J., Nicholson, J. K., & Holmes, E. (2015). Untargeted UPLC-MS profiling pipeline to expand tissue metabolome coverage: application to cardiovascular disease. *Analytical Chemistry*, *87*, 4184–93.

Weber, R. J. M., Winder, C. L., Larcombe, L. D., Dunn, W. B., & Viant, M. R. (2015). Training needs in metabolomics. *Metabolomics : Official Journal of the Metabolomic Society*, *11*, 784–786.

Westall, R. G. (1960). Argininosuccinic aciduria: identification and reactions of the abnormal metabolite in a newly described form of mental disease, with some preliminary metabolic studies. *The Biochemical Journal*, *77*, 135–144.

Willard, H. F., Ginsburg, G. S., & Geoffrey S. Ginsburg, H. F. W. (2010). *Essentials of Genomic and Personalized Medicine*. Academic Press.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., … Scalbert, A. (2013). HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Research*, *41*, D801–7.

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., … Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Research*, *35*, D521–6.

Xia, J., & Wishart, D. S. (2010). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, *38*, W71–7.

Yousri, N. A., Kastenmüller, G., Gieger, C., Shin, S.-Y., Erte, I., Menni, C., … Suhre, K. (2014). Long term conservation of human metabolic phenotypes and link to heritability. *Metabolomics*, *10*, 1005–1017.