Université de Montréal
Faculté des études supérieures

Ce mémoire est intitulé :

Développement d'un outil bio-informatique pour l'annotation des associations entre gènes et métabolites basée sur les voies métaboliques

Présenté par
Sandra Therrien-Laperrière

Département de Biochimie et de Médecine Moléculaire
Faculté de Médecine

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

Novembre 2017

© Sandra Therrien-Laperrière

Université de Montréal

Faculté des études supérieures

Ce mémoire est intitulé :

Développement d'un outil bio-informatique pour l'annotation des associations entre gènes et métabolites basée sur les voies métaboliques

Présenté par

Sandra Therrien-Laperrière

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

Julie Hussin
Membre du jury

Résumé

La métabolomique permet l'étude de l'ensemble des métabolites (ex : lipides, sucres, acides aminés) par le biais d'une variété d'outils analytiques et de protocoles expérimentaux qui engendre des coûts importants. Actuellement aucun laboratoire ne peut analyser l'ensemble des métabolites. C'est pourquoi, il est crucial de pouvoir prédire des classes de métabolites pertinentes à analyser en lien avec le phénotype étudié. Toutefois, il n'existe actuellement pas d'outil bio-informatique idéale pour accomplir cette tâche. Dans le cadre de ce projet, l'objectif était de développer un outil bio-informatique afin de prédire les métabolites pertinents à analyser en se basant sur la connaissance seule de l'architecture génétique du phénotype étudié. Afin d'atteindre notre objectif, nous avons posé l'hypothèse que les gènes encodant des enzymes catalysant des réactions métaboliques, modulent la concentration des métabolites à leur proximité dans les voies métaboliques. Cette hypothèse a été testée en calculant le court chemin réactionnel (*srd - sorthest reactional path*) entre les gènes (*SNPs* annotés à leur gène putatif) et les métabolites faisant partie des associations statistiques provenant du jeu de données *mGWAS* de Shin *et al.* en les cartographiant sur le réseau métabolique de la base de données *KEGG*. Des 79 associations impliquant un gène encodant une enzyme, 49 ont été annotées par une valeur *srd,* dont la valeur médiane est de 1. C'est-à-dire, qu'il existe une réaction entre le gène et son métabolite associé, ce qui indique que l'expression du gène peut avoir une influence importante sur la concentration du métabolite. L'annotation majoritaire de courte valeur *srd* pour les associations statistiques de Shin *et al.* démontre la pertinence de cette métrique pour définir un profil métabolique à analyser en fonction de l'architecture génétique. En revanche, le manque au niveau de la couverture de l'annotation de l'ensemble des associations pourrait être amélioré en appliquant la méthode avec d'autres bases de données, notamment, *Recon2*. En somme, *PathQuant* avec ses futurs développements représente un outil intéressant pour la prédiction d'un profil métabolique à analyser en fonction de l'architecture génétique d'un phénotype donné, en plus de préciser notre compréhension du contrôle des gènes sur le métabolisme.

Mots-clés : Métabolomique, Génomique, Réseaux métaboliques, *KEGG*, *mGWAS*, théorie des graphes

Abstract

Metabolomic enables the investigation of metabolites belonging to different chemical classes (ex: lipids, sugars, amino acids) which requires various methodologies and analytical tools. The current bottleneck is the impossibility to investigate every metabolite classes within one metabolic study or using one protocol. Thus, it is crucial to develop methods and tools to predict metabolites or metabolite classes to analyze for a given phenotype. The aim of this study was to develop a bioinformatic tool to prioritize metabolites to analyze based on the genomic architecture of a given phenotype. To achieve our goal, we hypothesized that genes encoding enzymes catalyzing metabolic reactions have an impact on the metabolite levels that are near them in metabolic pathways. We developed a method to compute the shortest reactional distance (srd) between a gene and a metabolite mapped on the metabolic pathways of the KEGG database. To test our method, we applied it to a dataset of statistical associations between genes (SNPs annotated to their putative gene) and metabolites reported by the mGWAS study of Shin et al. We mapped and annotated an srd value for 49 of the 79 associations involving a gene encoding an enzyme and a metabolite of that dataset with a median value of 1. Meaning there is only one reaction separating the gene from the associated metabolite. This indicates the genes could have a significant impact on metabolite levels. On the other hand, the lack of coverage of the associations could be improved by applying the method to other databases, in particular, Recon2. In conclusion, PathQuant and its future developments represent a relevant tool to predict a metabolic profil to analyze based on the genomic architecture of a given phenotype, in addition, it can improve the understanding of the genes control on metabolism.

Keywords: Metabolomic, Genomic, Metabolic network, KEGG, mGWAS, Graph theory

Table des matières

Liste des Tableaux

Liste des Figures

Liste des Documents Spéciaux

Tableaux supplémentaires de l'article :

Supplementary Table 1. Gene Identifications in KEGG

Supplementary Table 2. Metabolites Identifications in KEGG

Supplementary Table 3. Associations Identification in KEGG

Supplementary Table 4. PathQuant output of Shin et al. dataset

Supplementary Table 5. PathQuant output for every genes in 500 kb region of SNP of associations with an srd annotation

Supplementary Table 6. Srd annotation for associations annotated to an IEMS disease

Liste des Sigles et Abréviations

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

ACP : Analyse par composante principale

ADP : Adénosine DiPhosphate

ADN : Acide désoxyribonucléique

API : Application Program Interface

ARNm : Acide Ribonucléique messager

ASPG : asparaginase

ATP : Adénosine TriPhosphate

BRENDA : The Comprehensive Enzyme Information System

COSMOS : Coordinatino of Standards in MetabolomicS

DT2 : Diabète de Type 2

EHMN : Edinburgh Human Metabolic Network

EMBL-EBI : Institut Européenne de Bio-Informatique

eQTL : expression quantitative trait loci

ETFDH : electron-tranfering-flavoprotein dehydrogenase

GC : Gaz Chromatographie

GEM : Genome-scale metabolic model

GIM : Genitically Influenced Metabotype

GGM : Gaussian Graphical Modelling

GSEA : Gene Set Enrichment Analysis

GWAS : Genome-Wide Association Study

GO : Genome Ontology

HapMap : Haplotype Map

HMR : Human Metabolic Atlas

HUGO : HUGO Gene Nomenclature Committee

HumanCyc : Encyclopedia of Human Genes And Metabolism

IDs : Identifiers

IEMs : *Inborn Errors of Metabolism*

InChI : *IUPAC International Chemical Identifier*

INRA : Institut national de recherche agronomique

KEGG : *Kyoto Encyclopedia of Genes and Genomes*

KORA : *Kooperative Gesundheitsforschung in der Region Augsburg*

KGML : KEGG Markup Language

LC : Liquid Chromatography

LMSD : *LIPID MAPS Structure Database*

MARCH8 : *E3 ubiquitin-protein ligase MARCH 1/8*

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

mQTL : *metabolic quantitative trait loci*

NA : *Not Applicable*

NHGRI : Institut National de Recherche sur le Génome Humain

MS : *Mass spectrometry*

MSEA : *Metabolite Set Enrichment Analysis*

Overview : *KEGG Global and overview map of metabolic pathways*

PPM1K : *protein phosphatase 1K*

RMN : Résonance magnétique nucléaire

SLC33A1 : solute carrier family 33 member 1

SNP : *Signle Nucleotide Polymorphism*

SMILES : *Simplified Molecular-Input Line-Entry System*

SMPDB : *The small molecule pathway database*

srd : *shortest reactional distance*

WGCNA : *Weighted Correlation Network Analysis*

XML : *eXtensible Markup Language*

Remerciements

J’aimerais tout d’abord remercier ma directrice, Dr Christine Des Rosiers, pour m’avoir donnée la chance de réaliser ma maîtrise dans son laboratoire. Son support, ses encouragements, ses idées, son énorme générosité, sa rigueur scientifique, son ouverture d’esprit et son implication continuelle tout au long de mon parcours ont été des éléments clés pour mon succès et pour l’aboutissement de ce travail. Le laboratoire du Dr Des Rosiers est caractérisé par l’ambiance positive qui y règne et qui se traduit par une grande coopération ainsi qu’un excellent travail d’équipe.

Ensuite, j’aimerais remercier mes deux co-directeurs Guillaume Lettre et John David Rioux pour leur support, leurs conseils ainsi que pour leur contribution à mon projet. Merci également pour la correction de mon mémoire.

Dans un troisième temps, j’aimerais sincèrement remercier Matthieu Ruiz pour le temps qu’il a investi dans mon projet par son implication continuelle autant pour la correction de ce mémoire que pour la préparation de mes présentations orales aux congrès ASBMB à Chicago ainsi que R à Québec. Sa patience, ses encouragements, ses conseils et son expérience ont également été cruciaux à ma réussite.

Merci à l’ensemble des membres du laboratoire pour leur écoute, leurs conseils et leur support tout au long des ces deux dernières années. Merci particulièrement à Isabelle Robillard et Valérie Houde pour son aide lors de la conversion des identifiants des métabolites. J’aimerais également remercier Lise Coderre pour son soutien, ses conseils, les longues conversations ainsi que pour la correction de se mémoire.

Un grand merci également à Fabien Jourdan de m’avoir accueilli au sein de son laboratoire de l’INRA à Toulouse, France, lors du stage de deux semaines. Je remercie également spécialement Clément Frainay pour ses précieux conseils et son aide tout au long de ce stage.

J’aimerais également remercier Élaine Meunier pour son efficacité administrative sans pareil ainsi que pour conseils et recommandations.

Finalement, j’aimerais remercier ma famille, mon mari ainsi que ma belle-famille pour le support et les encouragements tout au long de mon parcours.

"The consequences of our actions are so complicated, so diverse that predicting the future is a very difficult business indeed"

- J.K Rowling

Avant-propos

La complexité de la biologie humaine est illustrée par le bagage génétique de chacun, par la diversité des mécanismes physiologique, moléculaire et cellulaire ainsi que part l'ensemble des facteurs environnementaux ayant un impact sur notre santé. Les maladies multi-factorielles ou complexes, telles que les maladies auto-immunes ou les cancers, reflètent cette complexité et nécessitent l'utilisation d'approches variées pour permettre la compréhension des mécanismes sous-jacents qui mènent au développement et à la prolifération de ces phénotypes. Ceci est primordiale pour le développement de meilleures méthodes diagnostiques, de traitements personnalisés et méthodes préventives. L'expansion fulgurante des technologies "omiques", principalement, la génomique, l'épigénomique, la transcriptomique, la protéomique et la métabolomique, a engendré une riche production d'information pour préciser la caractérisation des phénotypes. Cependant, malgré la grande quantité de données produites par chacune des ces disciplines, elles représentent chacune seulement partiellement le phénotype. En conséquence, le développement d'approches systémiques afin de combiner les différentes approches des sciences "omiques" sont actuellement activement favorisées.

Les approches systémiques ont pour but d'étudier tout les facteurs pouvant être impliqués dans les maladies incluant les facteurs environnementaux. Ces approches sont actuellement la base de nombreux projets de médecine personnalisée. Notamment, ces derniers visent à jouer un rôle important pour la prédiction de la réponse ou la non-réponse d'un patient à un traitement, pour la précision du diagnostique et pour la prévention. En plus des limitations individuelles de chaque discipline, la combinaison de certaines d'entre elles crée plusieurs nouveaux défis de taille

concernant, notamment, le développement de méthodes pour l'analyse des données mais aussi pour l'interprétation biologique des résultats, ce qui a été l'intérêt principal de ce projet.

Précisément, l'objectif de ce travail était de développer une approche bio-informatique pour permettre l'identification d'un profil métabolique pertinent à analyser en fonction de l'architecture génétique d'un phénotype. Ce travail a été entamé dans le cadre d'une étude multidisciplinaire concernant la maladie de Crohn pour laquelle 163 loci ont été identifiés et pour laquelle il est nécessaire de définir un profil métabolique à analyser. Pour développer une approche prédictive, il est en premier lieu nécessaire d'approfondir notre compréhension des liens biologiques entre les gènes et les métabolites. Ainsi les 163 loci liés à la maladie de Crohn pourront être utilisé dans une étape subséquente. Dans le cadre de ce travail, nous avons donc développé une approche permettant de caractériser et d'annoter les données provenant d'un type d'étude récemment développée qui combine la génomique à la métabolomique, nommé étude d'association pan-génomique avec la métabolomique (mGWAS), qui rapporte des associations statistiques entre locus(annoté au gène putatif) et métabolite. Ces dernières ont un potentiel immense pour comprendre le rôle des gènes et leurs contrôle sur le métabolisme. L'objectif principal de ce mémoire était de développer une méthode systématique permettant une annotation des associations par le calcul du court chemin réactionnel (*srd*), basé sur les voies métaboliques de la base de données *KEGG,* entre un gène et un métabolite. Un article original énonçant les résultats de la méthode développé a été rédigé.

Ce manuscrit a été soumis à *BMC Bioinformatics (BioMed Central)* sous le nom de « *Critical assessment of the shortest path metric for annotation of gene-metabolite pairs revealed by*

mGWAS ». La première version de cet article a été rédigé par une ancienne étudiante à la maîtrise, Sarah Cherkaoui. L'article avait été soumis et refusé par le journal de *Genome Medecine*. Depuis, l'apport de mon travail, c'est-à-dire le développement de la librairie R, PathQuant (https://github.com/sandraTL/pathquant), l'analyse approfondie de la métrique *srd* pour l'annotation des associations gène-métabolite discutés dans l'article ainsi que la compréhension de la portée biologique des résultats ont apportés une profondeur au manuscrit.

Cet ouvrage sera divisé en trois chapitres, le premier étant un relevé de littérature pour permettre a un sujet non-expert de comprendre l'intérêt et les résultats de la méthode développée, le deuxième présentera l'article rédigé et le troisième sera une discussion de la méthode ainsi que des résultats obtenues.

1 Relevé de Littérature

1.1 De la génomique à la métabolomique

Le dogme central de la biologie moléculaire énonce que l'information génétique se transmet de manière linéaire et unidirectionnelle au travers des mécanismes cellulaires ce qui conditionne le phénotype d'un individu [1]. Ce dernier est défini par l'ensemble des caractéristiques physiques (macroscopiques) et physiologiques (microscopiques) propres à un individu. L'importance du lien entre génotype (information génétique propre à un individu) et phénotype a été mise en valeur par la caractérisation de l'architecture génétique des maladies monogéniques qui sont causées par une mutation sur un gène. Toutefois, que ce soit pour les maladies monogéniques ou pour les maladies complexes (cancers, diabètes etc.), l'identification de mutations génétiques ne mène pas nécessairement à une compréhension de leurs impacts au niveau physiologique, moléculaire et cellulaire [2]. Ceci est dû à l'influence de l'environnement ainsi qu'à la présence de plusieurs strates ou niveaux cellulaires entre le génotype et le phénotype, entre autres, l'épigénome, le transcriptome, le protéome et le métabolome (définis dans la section suivante 1.1.1 ; figure 1.1). Le dernier niveau, le métabolome, représente l'ensemble des petites molécules et offre une lecture précise du phénotype au niveau microscopique. La particularité du métabolome est qu'il varie rapidement (de l'ordre de la milliseconde ou seconde) en fonction de l'état physiologique et de l'environnement. De manière plus générale, on considère que le profil des métabolites d'un individu dépend à la fois du génotype ainsi que de l'impact de l'environnement et représente « son empreinte métabolique », autrement dit, le métabotype. Ainsi, ce dernier représente un intermédiaire de choix pour comprendre les rôles du génome et de

l'environnement sur le phénotype. Cependant, dans le cadre d'approches où l'on veut combiner plusieurs niveaux cellulaires (approches systémiques), de nombreux défis restent à être relevés, entre autres, la gestion des données hétérogènes ainsi que l'interprétation biologique des résultats par les approches bio-informatiques.

Dans ce travail, une emphase a été mise sur l'interprétation et l'annotation des données provenant d'études combinant les données de génomique et les données de métabolomique. Ainsi, les sujets suivants seront discutés dans cette section : le dogme central de la biologie moléculaire et ses avancées, les aspects importants de la génomique (permet l'étude du génome) et de la métabolomique (permet l'étude du métabolome) qui sont nécessaires à la compréhension de la méthode développée.

1.1.1 Le dogme central de la biologie moléculaire et son évolution

La découverte du génome et de sa fonction au sein de la machinerie cellulaire est due à la redécouverte des travaux de Gregor Mendel au début des années 1900 sur l'hérédité des traits physiques d'une génération à l'autre. Les découvertes de Mendel combinées à l'identification des chromosomes en 1875 par M. Strasburger ont été la base des connaissances actuelles de la génétique. Les travaux de ce dernier ont été suivis de la découverte de l'acide désoxyribonucléique (ADN) comme étant le support de l'hérédité par Oswaldo T. Avery [3], dont les recherches ont été inspirées par Frédérique Griffith [4]. L'ADN est une macromolécule dont la structure se constitue d'un squelette sucre (ribose)/phosphate et d'un assemblage variable de paires de bases azotées ou nucléotides qui sont au nombre de quatre : l'adénine (A), la cytosine (C), la guanine (G) ou la thymine (T). Cette découverte a été suivie de celle de la structure de la

double hélice de la molécule d'ADN (1953) qui a précédé l'énoncé du dogme central de la biologie moléculaire sur le mécanisme de transfert de l'information entre les trois principales strates (ADN, acides ribonucléiques messagers - ARNm, protéines) considérées à l'époque par Francis Crick et James Watson [5]. En effet, le dogme énonce que la séquence nucléotidique de la molécule d'ADN est transcrite en séquence nucléotidique d'ARNm. Cette dernière est finalement traduite en séquence d'acides aminés (protéines). Francis Crick faisait alors mention de l'importance de l'ordre séquentiel et unidirectionnel du transfert de l'information dans la cellule (ADN ⟶ ARNm ⟶ protéine) [1]. L'hypothèse énoncée était soutenue spécifiquement par la perte d'information lors de la traduction de l'ARNm en protéines, ceci dû notamment à une redondance du code génétique qui répertorie 64 combinaisons possibles de triplets nucléotidiques, appelés codons, chacun codant pour l'un des 20 acides aminés. Ainsi, chaque acide aminé est codé par un ou plusieurs triplets de nucléotides, empêchant une traduction inverse précise. Parallèlement au transfert d'information énoncé par le dogme, il existe de nombreuses interactions inter-strates [6, 7], par exemple, le rôle des protéines dans la régulation de l'expression des gènes et l'impact des petites molécules sur les strates supérieures [8] ce qui contredit la notion d'unidirectionnalité du transfert d'information génétique. Par exemple certaines protéines, comme les facteurs de transcription, ou encore les histones vont avoir un impact sur l'expression des gènes. De plus, la complexité du transfert de l'information par la cellule a aussi été mise de l'avant par la découverte de la protéine transcriptase inverse qui permet une transcription de l'ARN d'un virus à l'ADN de l'hôte [9]. Plusieurs virus utilise ce mécanisme de propagation tel que le virus de l'immunodéficience humaine (VIH) [10, 11] et le virus T-lymphotropique (HTLV) [10]. Parallèlement, la découverte des germes pathogènes

composés entièrement de protéines, nommés prions, a démontré qu'il est possible qu'une infection se propage de protéines en protéines exclusivement, un exemple bien connu est celui de l'encéphalopathie spongiforme bovine (TSE), communément appelé maladie de la vache folle [9].

Depuis l'établissement du dogme, la complexité du transfert de l'information par la cellule a également été illustrée par l'identification de nouvelles strates en plus de celles classiquement décrites comme le transcriptome (ensemble des transcrits ou ARNm) et le protéome (ensemble des protéines) (Figure 1.1). Parmi les nouvelles strates, deux niveaux cellulaires essentiels augmentant la complexité du transfert de l'information génétique et faisant évoluer l'énoncé initial du dogme ont été ajoutés, nommément: l'épigénome [12, 13] et le métabolome [14-16]. En effet, l'épigénome, étudié par l'épigénétique (figure 1.1) [17], représente l'ensemble des modifications qui module la lecture d'une séquence d'ADN, et donc l'activité des gènes, sans en modifier la séquence nucléotidique. Ces variations jouent un rôle essentiel sur le contrôle de l'expression des gènes soit en activant soit en inhibant l'expression d'un gène donné. En ce qui concerne le métabolome, étudié par la métabolomique (figure 1.1) [18], cela représente l'ensemble des petites molécules ayant un poids inférieur à 1500 Da (métabolites), lesquels reflètent les produits de la machinerie cellulaire ainsi que l'impact de l'environnement sur l'organisme et donc le phénotype cellulaire.

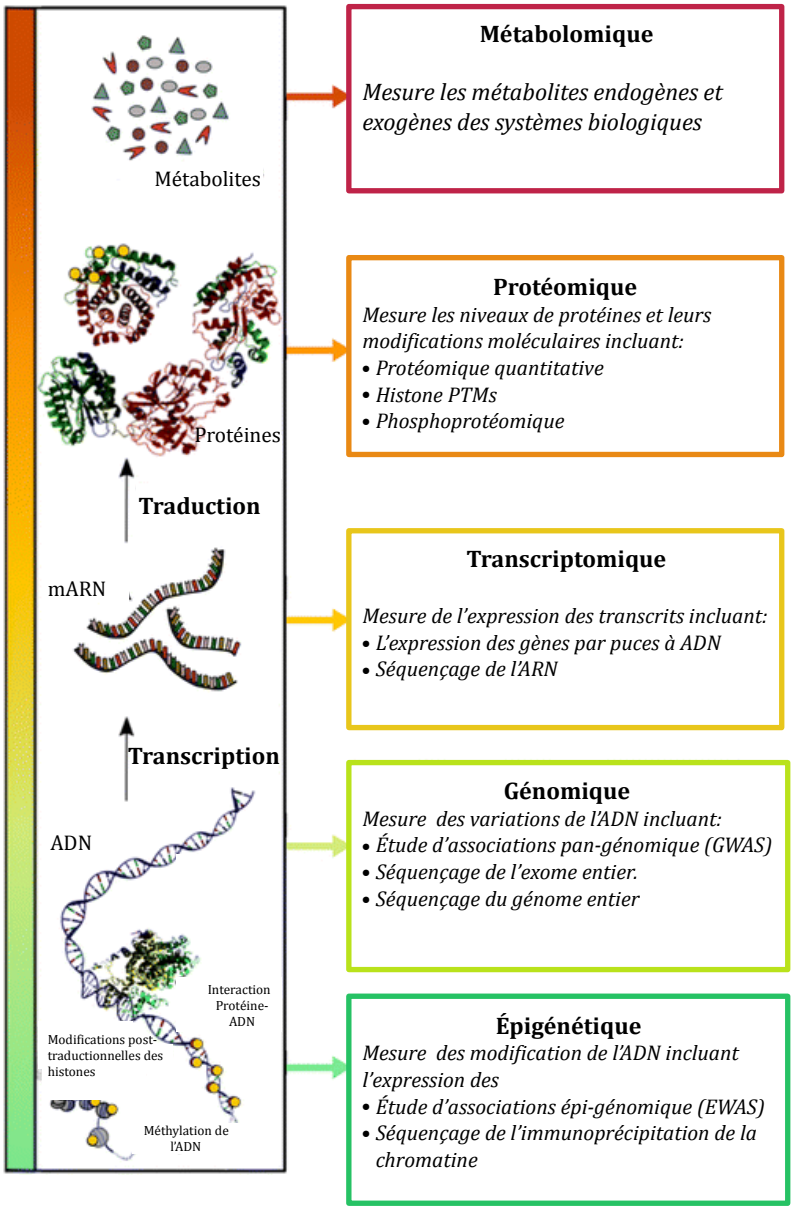

Figure 1.1: Liens entre les différentes disciplines « omiques » du génome au phénotype.

Tirée et adaptée de [19]. Cette image décrit les strates principales entre la source d'information, le génome, et les phénotypes observés. Tel que montré, la mesure du métabolome est la dernière avant d'atteindre le phénome.

1.1.2 La génomique

La génomique a été la première technologie “omique” à être développée. Cette science étudie l’ensemble des gènes (le génome) présent dans la molécule d’ADN, autrement dit l’ensemble du matériel génétique représentatif d’un individu (figure 1.1 A). Chez l’humain, le génome est constitué d’environ 3 milliards de paires de bases nucléotidiques (pb) et d’environ 25 000 gènes qui sont les vecteurs de l’hérédité. La singularité de chaque individu est définit par des milliers de variations génétiques, ou polymorphismes (*SNPs - single nucleotide polymorphisms)*. L’objectif principal de la génomique est de caractériser la structure des gènes et du génome en général, son organisation (génomique structurale), d’identifier les *SNPs* et les ressemblances (hérédité) entre les gènes des individus ainsi que le rôle, la fonction et la régulation de chaque gène (génomique fonctionnelle). Les *SNPs* étant la cause ou la conséquence de maladies sont principalement identifié par les études d’association pan-génomique (*GWAS - genome-wide association studies*), discutées à la section 1.1.2.2 [20]. Cependant, un des défis majeurs réside en la compréhension de l’impact biologique de ces *SNPs* sur le phénotype. Un bref historique des avancées de la génomique, les *GWAS*, l’annotation des *SNPs* ainsi que les limitations de cette technologie seront donc discutés dans les sous-sections suivantes.

1.1.2.1 Avancées en génomique

Les 25 dernières années ont donné lieu au développement du séquençage de première génération qui a permis d’obtenir la séquence du génome de plusieurs espèces ayant des molécules d’ADN de petite taille [21]. Toutefois, la taille du plus grand génome séquencé était 25 fois plus petite que celle du génome humain. C’est pourquoi, 20 groupes de recherche ont dû unir leurs compétences pour finaliser le projet de séquençage du génome humain (*Human Genome Project*)

dont la version initiale a été publiée en 2001 [21, 22], puis achevée en 2003 [23]. Malgré le succès et l'importance de cette étape déterminante, le coût excessivement élevé lié au séquençage du génome humain, estimé à environ de 550 millions de dollars canadiens, a grandement limité l'utilisation de cette technologie par la communauté scientifique. En revanche, le développement du séquençage de deuxième génération, au début du XXI^{ème} siècle, a grandement amélioré l'accessibilité du séquençage en réduisant les coûts [24]. Ceci, en raison d'une rapidité nettement supérieure par le séquençage en parallèle des échantillons ainsi que la lecture des séquences en temps réel [24]. De la même manière, cette nouvelle technologie a permis la simplification de la préparation des échantillons ainsi qu'une diminution de la quantité de données générées [24]. Cette avancée majeure a été suivi d'une augmentation exponentielle du séquençage de génomes [25]. Notamment, quelques groupes ont répertorié des séquences consensus du génome humain en plus d'identifier 2.4 millions de *SNPs* communs (présents chez plus de 5% de la population humaine), nommément: 1) le projet *HapMap (Haplotype Map),* dont le but est de définir les groupes de gènes ou des séquences conservés dans la population [26] 2) le projet des 1000 génomes (*1000 Genome Project*) dont le but est d'identifier l'ensemble des *SNPs* communs [27], ainsi que récemment 3) le Consortium de Référence de l'Haplotype (*Haplotyple Reference Consortium*) [28] dont le but est de créer une référence contenant des milliers d'haplotypes, c'est-à-dire un ensemble de variations génétiques qui ont tendance à être hérités, représentant l'ensemble de la population mondiale.

Le séquençage de deuxième génération permet d'identifier plusieurs types de variations génétiques, notamment: les *SNPs*, les variations structurelles et les variations épigénétiques [29]. Malgré les progrès fulgurants de cette technologie, l'identification des variations structurelles et

des variations épigénétiques est complexe contrairement à l'identification des *SNPs* [29]. Plusieurs types de *SNPs* peuvent être identifiés dans la séquence génomique : 1) substitution (remplacement d'un acide nucléique en un autre), 2) insertion (l'insertion d'un acide nucléique dans la séquence génomique) et 3) suppression (suppression d'un acide nucléique de la séquence génomique). Indépendamment du type de mutation, la conséquence au niveau protéique varie, notamment : 1) silencieuse (qui ne modifie pas l'acide aminé traduit) 2) faux-sens (qui induit un changement d'acide aminé et peut altérer la fonction de la protéine) et 3) non-sens (qui introduit un codon STOP). Les *SNPs* sont rapportés par les études de *GWAS,* notamment car elles permettent une étude systématique et non biaisée de l'ensemble du génome [20].

1.1.2.2 Les études d'associations pan-génomiques (*GWAS*)

Depuis la publication du premier *GWAS* réussi avec succès en 2005 [30], le nombre de publications de ces études a drastiquement augmenté [29]. Notamment, en date du 1er septembre 2016, 2 518 *GWAS* étaient répertoriées dans la littérature ainsi que dans le catalogue des *GWAS* (*GWAS Catalog)* [31] créé par l'Institut National de Recherche sur le Génome Humain (*NHGRI*) et l'Institut Européenne de Bio-Informatique (*EMBL-EBI*). À ce jour, le catalogue des *GWAS* répertorie plus de 51 000 associations entre un *SNP* et un trait donné. L'objectif des *GWAS* est de caractériser l'architecture génétique en identifiant les *SNPs* des séquences génomiques associés à un trait commun d'un groupe d'individus. Il existe deux types de traits: les traits quantitatifs (ex : la taille, le poids, le taux de glucose dans le sang) et les traits qualitatifs (ex : diabétique ou non). Les *GWAS* sont très performants et utilisent des puces de génotypage pouvant contenir plus de 2.4 millions *SNPs* communs pour identifier ceux pouvant être associés au trait étudié. Cependant, les *SNPs* peu fréquents (présents entre 1% et 5% dans la population) ainsi que les *SNPs* rares

(présents dans moins de 1% de la population) ne sont de manière générale pas détectés par les *GWAS*. Une association entre un *SNP* et un trait est considérée comme étant statistiquement significative si la fréquence du trait dans la population varie en fonction du *SNP* testé [32]. Pour les maladies monogéniques, une ou plusieurs mutations identifiées sur un même gène indiquent la cause directe du développement de ces phénotypes. En revanche, pour les maladies complexes, les *SNPs* identifiés ne définissent pas la cause mais représentent une prédisposition au développement de ces phénotypes.

1.1.2.3 L'annotation fonctionnelle des *SNPs*

L'identification exacte du gène dont la fonction est altérée par un *SNP* étant associé à un phénotype est primordiale pour la compréhension du rôle d'une mutation sur l'homéostasie des cellules menant au phénotype observé. Un *SNP* peut se retrouver soit 1) dans une partie exonique (partie codante d'un gène qui représente environ 3% du génome humain), 2) dans une partie intergénique (partie non-codante entre deux gène qui représente environ 75%) et 3) dans une partie intronique (partie non-codante entre deux séquences codantes même gène), partie non codante du génome où l'on retrouve, par exemple, des séquences de régulation de l'expression génique (~97%). Généralement, un *SNP* qui se retrouve dans une partie exonique du génome sera annoté au gène correspondant. Ceci dit, l'identification du gène ne mène pas nécessairement à une compréhension de l'impact fonctionnel de la mutation [33]. D'autre part, l'annotation des *SNPs* se trouvant dans les parties non codantes du génome est plus complexe. Il existe des outils proposant une liste de gènes candidats potentiels tels que *HaploReg* [34] et *regulomeDB* [35] qui offrent des informations sur l'expression des gènes et leur régulation. Toutefois, les choix doivent être manuellement analysés. Généralement, une évaluation de la littérature scientifique des gènes

candidats est effectuée, si ces recherches sont infructueuses, une annotation du gène étant le plus proche physiquement du *SNP* peut être faite. Dans l'ensemble, cette étape cruciale n'est pas triviale et nécessite toujours le développement de nouvelles techniques pour la faciliter.

1.1.2.4 Limitations et perspectives des *GWAS*

Les études *GWAS* ont jusqu'à présent apporté une richesse incroyable pour la caractérisation de l'architecture génétique des maladies monogéniques et complexes [31]. Malgré tout, deux limitations doivent être considérées: 1) la compréhension de l'impact des *SNPs* sur les strates des niveaux inférieurs et sur le phénotype et 2) la limitation des *SNPs* communs pour expliquer l'héritabilité des maladies complexes. Un exemple démontrant la première limitation est la maladie monogénique du syndrome de Leigh de type canadien français (LSFC). Cette dernière est caractérisée au niveau biochimique par une dysfonction de la chaine chaîne respiratoire mitochondriale causée par une mutation du gène LRPPRC [36]. Cependant, les conséquences moléculaires et cellulaires, et donc phénotypiques, de cette mutation sont encore mal connues dans LSFC. En outre, cette pathologie se caractérise par la survenue impromptue de crises d'acidose lactique souvent fatales, dont les causes ne sont pas définies. Ainsi malgré la connaissance de la cause génétique, le phénotype moléculaire associé à la maladie est encore incompris [37, 38]. En ce qui concerne la deuxième limitation, les mutations identifiées comme étant liées aux maladies complexes expliquent généralement seulement un faible pourcentage de l'héritabilité du trait [2]. Au-delà de l'influence non-négligeable de l'environnement, une grande partie des *SNPs* identifiés se retrouvent dans les parties intergéniques du génome, ce qui explique en partie de la difficulté pour la compréhension de l'impact de ces *SNPs* au niveau fonctionnel. La troisième limitation est expliquée par le fait que lors d'une étude GWAS, plusieurs *SNPs*

statistiquement associés au phénotype étudié sont trouvés dans une même région génomique, toutefois, identifier le ou les variants causaux est complexe et plusieurs stratégies de fine-mapping ont été mises en place pour pallier à ce problème [39]. Un des exemples le plus illustratif des limitations 2) et 3) est celui de la maladie complexe du diabète de type 2 (DT2). En 2012, près de 50 *SNPs* étaient identifiés comme ayant un impact sur le développement ou la prolifération du phénotype. Ceci dit, l'estimation de l'héritabilité du DT2 expliquée par ces 50 *SNPs* était seulement de 5% à 10%. Aujourd'hui, 100 mutations sont connues cependant, elles expliquent un pourcentage seulement légèrement plus élevé d'héritabilité du DT2 [39]. L'hypothèse actuelle suppose que l'architecture génétique de l'ensemble des mutations menant aux maladies complexes est hétérogène; composée de *SNPs* communs ayant un effet faible, de *SNPs* peu fréquents ainsi que de *SNPs* rares. Cependant, comme mentionné ci-haut, ces deux derniers ne sont pas détectés par les GWAS classiques. En revanche, plusieurs efforts ont été menés dans les dernières années pour le développement de techniques permettant leur détection en utilisant le séquençage de deuxième génération [33]. Il va sans dire qu'une meilleure compréhension de l'architecture génétique de ces phénotypes est indispensable pour une compréhension précise au niveau mécanistique. Pour atteindre cet objectif, la combinaison de l'étude du génome à l'étude d'autres niveaux cellulaires est nécessaire, en particulier le métabolome (figure 1.1) pour sa proximité avec le phénotype [40].

1.1.3 La métabolomique

La métabolomique étudie l'ensemble des métabolites issus des transformations biochimiques du métabolisme ce qui forme le métabolome [14-16]. Contrairement à la génomique, la métabolomique est la plus jeune des technologies "omiques". Cette technologie permet de

caractériser le métabolome d'un individu et de définir son métabotype dans une condition donnée [41]. Plusieurs matrices biologiques différentes telles que le sérum, le plasma, l'urine, les selles ainsi que divers tissus biologiques peuvent être utilisés et permettent la détection de milliers de métabolites. Toutefois, le succès des études de métabolomique dépend grandement de la précision de l'approche expérimentale, ce qui inclus : la conception du protocole expérimental, la préparation des échantillons, les méthodes analytiques et l'analyse de la grande quantité de données obtenues, autrement appelées mégadonnées [16, 42]. Étant donné la sensibilité et la variabilité rapide du métabolome en réponse à de nombreux facteurs (appelés facteurs confondants), le développement de l'approche expérimentale doit se faire avec une extrême précaution. En outre la température est un élément clé à considérer et conditionne l'absolu nécessité d'un prélèvement dans des conditions précises (4 degrés celsius) pour arrêter toute réaction biochimique au moment *t* du prélèvement. Également, il est important de bien établir la question ou l'hypothèse de recherche. Par exemple, si l'on veut s'intéresser à l'impact d'un état pathologique sur le métabolome, il est important de contrôler tout facteur confondant qui pourrait également avoir une influence non négligeable sur le métabolome comme l'âge, le sexe, l'état nutritionnel etc. [42]. En outre, l'analyse et l'interprétation biologique des megadonnées générées par la métabolomique engendrent également des défis considérables [43]. Dans cette section, la diversité du métabolome, les études de métabolomique, l'analyse de résultats ainsi que les avancées et les défis actuels de cette technologie seront discutés.

1.1.3.1 La diversité des métabolites

La particularité et la complexité que représente la métabolomique s'explique en partie par la diversité des propriétés chimiques des métabolites. Plusieurs bases de données visent à répertorier

des ensembles précis ou l'ensemble des métabolites toutes espèces confondues, telles que: *The Toxin and Toxin Target Database (T3DB)* [44]*, the Food Component Database (FooDB), the database and ontology of Chemical Entities of Biological Interest (ChEBI)* [45, 46]*, the Kyoto Encyclopedia of Genes and Genomes (KEGG)* [47] et *the public repository for biological properties of small molecules (PubChem)* [48]. Tandis que d'autres bases de données visent à répertorier le métabolome d'organismes spécifiques, telles que: *the Human Metabolome Database (HMDB)* [49] et *the Small Molecule Pathway Database (SMPDB)* [50] pour l'humain. Contenant près de 75 000 métabolites identifiés, dont 25 000 ont été mesurés à partir d'échantillons sanguins, l'*HMDB* est la base de données la plus imposante pour l'humain [49]. Près de 30 000 des 75 000 métabolites sont endogènes, c'est-à-dire produits par l'organisme, et se divisent en deux catégories principales soit les hydrosolubles (ex. : acides aminés, sucres, nucléotides) et les liposolubles (ex. : lipides). Autrement, plus de 45 000 métabolites répertoriés par l'*HMDB* sont des métabolites exogènes, provenant de l'alimentation, des médicaments, de l'environnement ou autre. En revanche, l'annotation complète du métabolome d'aucune espèce n'a été accomplie jusqu'à présent [16] et il est raisonnable de penser qu'il existe des milliers de métabolites toujours inconnus. Une limitation majeure à la caractérisation de l'ensembles des métabolites est due à leur diversité chimique [51] car il n'existe pas un seul protocole expérimental et un seul instrument pouvant analyser l'ensemble des métabolites en une analyse unique, mais aussi, à la complexité des méthodes expérimentales et analytiques pour identifier les métabolites inconnus [52, 53].

1.1.3.2 Les études de métabolomique

Les études de métabolomique se divise en deux catégories principales, précisément, ciblées et non-ciblées [42]. Les études ciblées visent à analyser et quantifier un ensemble prédéterminé de métabolites connus ayant des propriétés chimiques semblables tandis que, l'approche non-ciblée vise à mesurer le plus grand nombre de métabolites pouvant être détectés dans un échantillon donné sans à priori sur les métabolites recherchés. Cette approche est utilisée plutôt dans une perspective de découverte et on la considère comme une approche génératrice d'hypothèses. Quoiqu'il en soit, dans les deux cas, deux types d'instruments analytiques sont majoritairement utilisés: la résonance magnétique nucléaire (RMN) [54] et la spectrométrie de masse (*MS*) [55].

1.1.3.2.1 Les instruments analytiques

La RMN [54] analyse les métabolites par l'exploitation des propriétés magnétiques des noyaux atomiques des molécules lorsqu'ils sont soumis à un champ magnétique. Les avantages majoritaires de la RMN sont: 1) peu de préparation des échantillons, 2) la non-destruction des échantillons et 3) la possibilité d'identifier la structure des composés. En revanche cette méthode s'accompagne des limitations suivantes: 1) la détection des métabolites à concentration élevée seulement (de l'ordre du micromolaire), 2) le besoin d'un grand volume d'échantillon et 3) la détection d'un faible nombre de métabolites. En ce qui concerne la *MS* [55], cette dernière permet l'analyse de métabolites préalablement ionisés selon leur masse, plus précisément masse/charge (m/z) et leur temps de rétention lorsque la *MS* est couplée à une colonne de chromatographie. La *MS* peut être utilisée seule, sans séparation chromatographique préalable (« shotgun »), ou couplée à de la chromatographie gazeuse (*GC-MS*) ou liquide (*LC-MS*), dépendamment des composés analysés et de leurs propriétés physico-chimiques [56]. Les avantages de la *MS* sont :

1) la détection de métabolites jusqu'à des concentrations très faibles (ordre du pico-molaire), 2) la détection de milliers des métabolites (non-ciblé), 3) l'analyse dans un faible volume d'échantillon. Tandis que ses désavantages sont : 1) une préparation d'échantillons complexe et 2) la destruction des échantillons lors de l'analyse. En outre, dans le cas d'une approche non-ciblée, les composés analysés sont au préalable caractérisés seulement sur la base de leur m/z et leur temps de rétention sur la colonne. Il est donc indispensable de procéder à des étapes supplémentaires pour identifier les métabolites d'intérêts. Toutefois cette identification n'est pas triviale, ce qui rend la procédure post-analytique plus complexe [57, 58]**.**

1.1.3.2.2 L'analyse des données

L'analyse des données provenant des études ciblées est relativement simple comparativement à l'étude des mégadonnées provenant des études non-ciblées qui est complexe indépendamment de la plateforme utilisée [43]. Cette complexité sera illustrée au travers de la brève description du traitement des mégadonnées provenant de la *MS*. Particulièrement, sur les spectres de masse*,* les métabolites sont caractérisés par leur ratio m/z, leur temps de rétention et l'intensité des pics reflétant la concentration des métabolites. L'analyse des données nécessite globalement les étapes suivantes : 1) l'évaluation de la qualité des données, 2) l'harmonisation des données, 3) la détection des pics et l'alignement des spectres 4) l'identification des métabolites connus [52] et 5) l'identification des métabolite inconnus (optionnel). Notamment, la complexité de l'analyse provient des étapes 3 et 5. Le défi crucial de l'étape de la détection des pics (3) réside dans la grande quantité de pics et le chevauchement de plusieurs d'entre eux dans le spectre. En revanche, l'alignement des spectres générés pour un ensemble d'échantillon cause aussi des problèmes au niveau algorithmique puisqu'il y a souvent des décalages entre le temps de

rétention (axe des x) pour un même pic d'un spectre à l'autre. Plus le nombre de métabolites analysé est grand, plus grand est le nombre de pics et donc la complexité de ce problème augmente rapidement. Comme mentionné dans les sections précédentes, l'étape 5 est également complexe et nécessite des études complémentaires. En revanche, pour comparer deux conditions expérimentales, il existe diverses approches statistiques. Plus précisément on retrouve les approches univariées, en particulier le test de t ou les approches multivariées telles que l'analyse de composante principale (ACP). Une différence d'intensité du signal entre deux conditions traduit une variation de la concentration du métabolite [43, 59]. Lorsque des métabolites différant d'une condition à l'autre sont identifiés, l'interprétation biologique peut être faite par une revue de la littérature, par l'intermédiaire des bases de données ou par des analyses d'enrichissement des métabolites majoritairement dans les voies métaboliques mais également dans les voies de signalisation (*MSEA*) [60]. Il existe des outils permettant d'accomplir l'ensemble des étapes nécessaires à l'analyse des données, entre autres, *MetaboAnalyst* [61] est utilisable en ligne.

1.1.3.3 Limites et perspectives

L'objectif ultime de la métabolomique est de caractériser le métabolome de l'ensemble des espèces ainsi que d'associer un métabotype à un phénotype donné. Malgré les avancées fulgurantes de cette discipline, la caractérisation du métabolome est limitée par la technologie ainsi que par le traitement des données. Ceci dit, l'étape la plus limitante à ce jour est l'identification des composés inconnus [57], notamment parce que la technologie et les protocoles expérimentaux ne permettent pas d'identifier tous les métabolites. De plus l'identification des métabolites est coûteuse en temps et argent, ainsi peu de chercheurs s'aventurent dans cette étape ce qui a pour conséquence de ralentir considérablement

l'identification de nouveaux composés. Une autre limitation coûteuse en temps est le traitement des mégadonnées, pour les études non-ciblées [52]. Néanmoins, l'évolution des techniques expérimentales, de la technologie et du traitement des mégadonnées ont permis l'identification de plusieurs milliers de composés en l'espace d'une dizaine d'années seulement, chez l'humain, en passant de ~2000 métabolites à ~75 000 métabolites [62].

Par ailleurs, un défi commun à l'ensemble des disciplines en biologie est la compréhension de la pertinence des différences observées d'un point de vue cellulaire et moléculaire mais également de faire le lien entre le génotype et le phénotype. Dans ce contexte, l'approche systémique qui vise à combiner les données venant de plusieurs niveaux cellulaires apporte une richesse indispensable et complémentaire à l'étude individuelle des strates cellulaires afin d'améliorer notre compréhension globale de l'ensemble des comportements de nos cellules.

1.2 L'approche systémique : intégration des "omiques"

L'approche systémique se veut une approche intégrative ayant pour objectif de refléter avec la meilleure précision possible le phénotype d'un individu ou d'un trait donné [8]. L'importance de l'approche systémique est d'ores et déjà considérable en dépit de la jeunesse de cette discipline [63]. Cependant, plusieurs défis s'ajoutent à ceux de chaque "omique" et restent à résoudre, notamment, la gestion de l'immense quantité de données hétérogènes ainsi que l'interprétation biologique qui résulte d'une telle approche. Dans cette section, seront discutés les éléments clés caractérisant l'approche systémique ainsi que les stratégies d'intégrations des données et les limitations actuelles.

1.2.1 L'approche systémique

À ce jour, il n'est généralement pas réalisable d'utiliser la totalité des technologies "omiques" pour étudier un même phénotype. C'est pourquoi, la conception des études systémiques dépend de la question biologique posée, du phénotype étudié, des technologies disponibles et des ressources financières. En conséquence, les approches systémiques sont diversifiées et combinent au minimum deux "omiques" [64-67] ou plus [68-71]. Des centaines d'études ont implémenté une approche systémique, par exemple, une étude ayant combiné la génomique, la transcriptomique et la protéomique pour caractériser les réponses cellulaires impliquées dans la réparation des lésions de l'ADN de différentes maladies chez l'humain [69]. Ces mécanismes sont extrêmement importants pour la compréhension d'une multitude de maladies. De même, les études ayant combinées la génomique à la métabolomique ont apporté une meilleure compréhension et caractérisation du contrôle du génome sur le métabolisme [72-74]. Une fois les données "omiques" acquises, l'analyse se divise en deux approches [70] discutées à la section suivante, notamment: 1) les méthodes basées sur les données [75, 76] et 2) les méthodes basées sur les connaissances [77, 78].

1.2.2 Stratégies pour l'intégration des omiques

Méthodes basées sur les données

Ces approches sont utilisées pour détecter les liens biologiques intra et/ou inter données "omiques". Ces approches statistiques sont particulièrement intéressantes lorsqu'il y a un manque de connaissances à propos des mégadonnées acquises pour l'espèce étudiée. Un exemple d'approche statistique est l'analyse pondérée du réseau de corrélation des gènes (*WGCNA*) [79] qui permet de regrouper les gènes ayant de propriétés similaires. Cette méthode permet

également d'ajouter aux groupes de gènes fortement corrélés de l'information supplémentaire à propos d'autres "omiques" tel que la protéomique mais également des données cliniques. Les méthodes de corrélation sont généralement facilement applicables, cependant, elles peuvent être limitantes lorsque les réseaux de corrélation formés sont trop denses. Le modèle graphique gaussien (*Gaussian graphical modelling* - *GGM*) basé sur des corrélations partielles remédie à ce problème par sa capacité à dissocier les associations directes et indirectes. Ce dernier a d'ailleurs été utilisé pour analyser les données provenant de la métabolomique [75] et pour les données provenant de la transcriptomique [76, 80].

Méthodes basées sur les connaissances

Plusieurs méthodes d'analyse des données "omiques" ont été développées pour utiliser les connaissances répertoriées dans les bases de données. Par exemple, la méthode mentionnée à la section 1.1.3.2.2, *MSEA* [81], a pour but d'identifier les voies réactionnelles contenant plusieurs métabolites selon le phénotype étudié. Une méthode similaire pour les gènes existe, *GSEA* (*Genome Set Enrichment Analysis*) [82], pour identifier les voies réactionnelles où il y a plusieurs gènes liés à un phénotype donné. Par ailleurs, quelques bases de données permettent une visualisation des données dans les voies réactionnelles telles que *KEGG* [47] et Reactome [83]. Ceci dit, plusieurs reconstructions du métabolisme à l'échelle du génome ont été développées pour apporter une compréhension plus globale des résultats provenant de ces méthodes. Ces reconstructions ont jusqu'à présent été bâties à partir de l'immense quantité de connaissances disponibles pour des espèces grandement étudiées, appelés organismes modèles, tels que *Saccharomyces cerevisiae* [84], *Caenorhabditis elegans* [85], *Mus musculus* [86] ainsi que pour l'humain [87]. Ces reconstructions contiennent au minimum les données de génétique et de

métabolomique. Toutefois les données de transcriptomique, d'épigénétique et de protéomique ou autres peuvent y être ajoutées [71]. De plus, ces modèles peuvent être manipulés pour construire des modèles cellulaires et tissulaires spécifiques en utilisant les données "omiques" récoltées [88-90]. Malgré l'incomplétude de ces modèles, un avantage colossal est qu'il est possible d'analyser la complexité des interactions inter et intra "omiques". Toutefois, en raison de la complexité et de la diversité des méthodologies, cet avantage représente également un grand défi pour l'approche systémique [7].

1.2.3 Limitations et perspectives

La jeunesse de cette discipline, la complexité et la diversité des approches entrainent directement plusieurs limitations. Évidemment, les limitations inhérentes à chacune des disciplines "omiques" sont également des limitations des approches systémiques les utilisant. On peut notamment penser aux données manquantes dans les données de génomique que l'on doit imputer à l'aide d'un génome de référence ainsi que dans les résultats d'études de métabolomique où parfois, certains métabolites ne sont pas retrouvés dans l'ensemble des échantillons analysés. Ceci dit, les méthodes d'analyses basées sur les données sont limitées par le fait qu'elles ne tiennent pas compte des connaissances actuelles. En ce qui concerne les méthodes basées sur les connaissances, elles sont majoritairement limitées par l'incomplétude des modèles utilisés. Enfin, l'interprétation biologique des résultats provenant des deux méthodes représente un défi considérable pour les chercheurs par la diversité et la grande quantité de données. Ainsi, le développement d'outils facilitant cette tâche est nécessaire pour rendre l'interprétation biologique optimale et la plus fiable possible.

1.3 Les études d'associations pan-génomiques combinées à la métabolomique (*mGWAS*)

Dans le cadre de ce projet, une attention particulière est donnée à la combinaison des données de génomique et de métabolomique car elle permet de combiner l'étape initiale du transfert de l'information génétique ainsi que la dernière étape qui est modulée par les gènes et l'environnement : le métabolome. De cette manière, la combinaison entre la génomique et la métabolomique est très intéressante car elle promet une compréhension globale de la machinerie cellulaire en plus de permettre une meilleure caractérisation du contrôle génétique sur le métabolisme appelé: métabotype influencé par le génome (*GIMs*) [40, 91]. Le potentiel de cette technique a été démontré par plusieurs études qui ont aidé à la caractérisation du *GIM* de centaines de traits métabolique [73]. Notamment, deux études d'une plus grande envergure ont été produites: 1) par Shin et al. en 2014 [72] et 2) par Telenti et al. en 2017 [74] rapportant chacune plusieurs centaines d'associations entre *SNPs* et métabolites. Cette section présentera la méthodologie des études *mGWAS*, puis les stratégies utilisées pour l'interprétation biologique des associations et finalement les limitations et les perspectives.

1.3.1 La méthodologie

Une étude *mGWAS* [64] débute par l'acquisition des données de génomique et de métabolomique de manière distincte. Comme discuté à la section 1.1.3.2, l'importance de la conception du protocole pour l'acquisition des données de métabolomique est cruciale pour le succès de l'étude. Une fois les données acquises, le principe d'un *mGWAS* est de réaliser un *GWAS* pour chaque métabolite mesuré (identifié ou non), ceux-ci sont les traits quantitatifs. Ainsi, ces études rapportent une liste d'associations statistiques entre *SNPs* et métabolites. Une étape

supplémentaire de ces études est d'accomplir un *GWAS* pour chaque ratio métabolique, ce qui forme des associations statistiques *SNP*-(métabolite 1/métabolite 2). Cette étape est cruciale puisqu'il a été démontré que la quantification de la concentration de chaque métabolite de manière individuelle ne reflète pas nécessairement l'impact d'un phénotype [92]. Ceci s'explique par les interactions entre les voies réactionnelles qui causent une modulation complexe des concentrations des métabolites [92]. De plus, l'utilisation des ratios entre métabolites est couramment utilisée pour approximer l'activité réactionnelle des enzymes. Un des premiers exemples de l'utilisation des ratios a été l'utilisation du ratio phénylalanine/tyrosine pour le diagnostic de la maladie monogénique phénylcétonurie [93]. De plus, la métrique *p-gain* [92] a été introduite pour tester les ratios métaboliques dans les études *mGWAS*.

1.3.2 Interprétation et analyse des résultats

Jusqu'à présent, deux méthodes ont été utilisées pour analyser les associations entre *SNPs*-métabolites ou *SNPs*-ratios résultant des études *mGWAS* [94]: le *GGM* [75] et *Minning the Unknown* [95]. Plus précisément, ces approches sont axées sur les données générées et non sur les connaissances de la littérature scientifique et les bases de données. Les deux méthodes sont décrites brièvement ci-dessous.

1.3.2.1 Le modèle de graphe gaussien - *GGM*

La méthode *GGM* [75] produit un graphe représentant les voies métaboliques basé sur les corrélations partielles entre les métabolites mesurés. Il a été démontré que les corrélations partielles obtenues par cette technique représentent majoritairement des métabolites appartenant à la même classe chimique ou étant présent dans les mêmes voies métaboliques [75]. Dans le cadre de l'étude par Shin *et al.* [72], un *GGM* a été construit pour évaluer les relations entre les

métabolites mesurés. À ce graphe, ont été ajoutés, les gènes (*SNPs* annotés à leurs gènes putatifs) en les liants à leur(s) métabolite(s) associé(s) [96]. Dans le cadre de l'étude par Shin *et al.,* ce graphe a été construit dans le but de visualiser l'interaction entre les gènes et les métabolites *in vivo*. Malgré son intérêt, cette méthode ne tire pas avantage des connaissances actuellement présentes dans les modèles cellulaires multi-omiques existants.

1.3.2.2 *Mining the unknown*

Mining the unknown [95] tire avantage du *GGM* puisque ce dernier permet de regrouper les métabolites ayant des propriétés biochimiques semblables. Ainsi, *Mining the unknown* construit un *GGM* en utilisant l'ensemble des métabolites mesurés (connus et inconnus) ce qui permet d'émettre des hypothèses sur l'identité biochimique des molécules inconnues selon les propriétés des molécules connues auxquelles elles sont fortement corrélées. Tout en reconnaissant l'importance de cette méthode pour l'identification de métabolites inconnus, elle ne permet pas une meilleure compréhension des associations entre gènes et métabolites rapportées par les *mGWAS*.

1.3.3 Les limitations et perspectives

Le potentiel des résultats provenant des études *mGWAS* est immense, ceci dit, les outils utilisés pour comprendre les processus biologiques mis en avant par les associations rapportées ne mènent pas à une compréhension des associations d'un point de vue cellulaire, moléculaire ou physiologique. Partant de ce fait, il est évident qu'il y a un besoin au niveau du développement des méthodes d'analyses basées sur les connaissances actuelles. En outre, les limitations propres à la génomique et à la métabolomique font également parties des limitations inhérentes aux mGWAS. Notamment, les associations formées à partir de métabolites inconnus ne peuvent pas

être analysées à ce jour. De plus, l'annotation des gènes affectés par les *SNPs* détectés peut être erronée ce qui demande une vigilance lors de l'analyse pour assurer une compréhension juste au niveau biologique des associations rapportées par les *mGWAS*.

1.4 Bases de données des voies métaboliques

La découverte et l'étude de plusieurs enzymes ont motivé D. Nicholson et S. Dagley, en 1955, à éditer le premier livre rassemblant les 20 réactions enzymatiques, autrement dit réactions métaboliques, connues à cette époque [97]. Depuis lors, plusieurs domaines d'étude dont l'enzymologie ou la biologie moléculaire ont contribué à la découverte et à la compréhension d'un nombre considérable de réactions métaboliques. L'ensemble de ces dernières est aujourd'hui rapporté dans les livres de biochimie ainsi que dans les bases de données qui répertorient, à ce jour, plusieurs milliers de métabolites, d'enzymes, de réactions métaboliques ainsi que de voies métaboliques (ensemble de réactions métaboliques). Il existe deux types de bases de données répertoriant les voies métaboliques, celles qui les répertorient de manière individuelle et celles qui répertorient des reconstructions du métabolisme où l'ensemble des voies sont interconnectées. Les principales bases de données répertoriant les voies métaboliques individuelles sont *SMPDB* [50], *HumanCyc* (*Encyclopedia of Human Genes And Metabolism*) [98], *Reactome* [83], LMSD (*LIPID MAPS Structure Database*) [99], *BRENDA* (*The Comprehensive Enzyme Information System*) [100] et *WikiPathways* [101]. Ces dernières ont des utilités multiples, notamment pour la visualisation et l'analyse des données expérimentales. En ce qui concerne les bases de données répertoriant les reconstructions du métabolisme, on retrouve *KEGG* (*Kyoto Encyclopedia of Genes and Genomes*) [47], *Recon 2* [87], *EHMN* (*Edinburgh Human Metabolic Network*) [102] et *HMR* (*Human Metabolic Atlas*) [103], qui sont actuellement

les plus utilisées par la communauté scientifique. Ces dernières sont majoritairement mises à profit dans le cadre d'analyses de prédiction de phénotype et/ou de réponses aux perturbations biologiques par des analyses de flux ainsi que pour l'analyse et l'interprétation des mégadonnées. Certaines des bases de données (*Reactome, HumanCyc, Recon 2, SMPDB, HMR, EHMN*) répertorient seulement des voies métaboliques spécifiques à l'humain tandis que les autres (*KEGG, LMSD, WikiPathways, BRENDA*) répertorient des voies métaboliques issues de plusieurs espèces, incluant l'humain. Dans cette section, les réactions enzymatiques et les voies métaboliques, les protocoles de reconstruction des voies métaboliques, les bases de données ainsi que les limitations reliées à l'interprétation de résultats d'analyse combinant l'utilisation de plusieurs bases de données seront abordés.

1.4.1 Réactions enzymatiques et voies métaboliques

Une réaction enzymatique permet la transformation d'un métabolite, nommé substrat, en un autre métabolite, nommé produit. Ces réactions sont initiées par une enzyme, encodée par un gène, qui joue le rôle de catalyseur, ce qui a pour but d'accélérer le processus réactionnel. Par souci de précision, le terme réaction métabolique sera employé lorsqu'il est question d'une réaction enzymatique ayant une seule étape réactionnelle de substrat(s) à un produit(s). Tandis que, le terme voie métabolique sera utilisé lorsque l'on parle d'un enchainement de plusieurs réactions où le produit de l'une est le substrat de l'autre. Parfois, plusieurs enzymes sont nécessaires à la catalyse d'une réaction métabolique, ces dernières forment un complexe enzymatique. De plus, les enzymes ont souvent besoin de cofacteurs (ex : NAD+, Mg2+, H2O, CO_2), lesquels sont indispensables à la catalyse des réactions. Les enzymes sont répertoriées par un système de classification qui se base sur le type de réaction enzymatique catalysée afin de leur attribuer un

identifiant unique, soit le *Enzyme Commission Number* (*EC*) [104]. Il existe également quelques réactions spontanées qui se produisent sans l'aide d'enzyme, par exemple, les réactions d'oxydation des lipides poly-insaturés [105]. Ainsi, le métabolisme est défini par l'ensemble des voies métaboliques formées par les réactions enzymatiques et les réactions spontanées qui se produisent dans un organisme.

1.4.2 Méthodes de reconstruction du métabolisme

Il existe actuellement deux approches pour générer la reconstruction du métabolisme d'une espèce, en particulier, l'approche ascendante et l'approche descendante.

Approche ascendante

L'approche ascendante se base sur les connaissances présentes dans la littérature scientifique pour construire les voies métaboliques en utilisant les gènes, les enzymes, les métabolites et les réactions répertoriées. Cette approche permet d'obtenir une reconstruction qui reflète au mieux les connaissances actuelles du métabolisme de l'espèce en question. Néanmoins, les connaissances dans la littérature scientifique sont fragmentaires, ce qui implique que les bases de données le sont également. Une information manquante dans les reconstructions du métabolisme est identifiée lorsqu'un métabolite ne peut être consommé ou produit par aucune réaction créant ainsi des trous dans les voies métaboliques. Les trous ont un impact non négligeable sur les résultats d'analyses obtenus avec ces bases de données [106]. Une méthode automatique *GapFind* [107] a été développée afin d'identifier les endroits où il y a de l'information manquante dans les voies métaboliques. Plus précisément, *GapFind* sélectionne les métabolites présents dans le réseau mais non produits par une réaction enzymatique. Pour pallier ce déficit, quelques méthodes ont été développées afin de proposer des réactions et/ou des enzymes qui

pourraient combler les données manquantes. Par exemple, il existe une méthode qui utilise les reconstructions du métabolisme de plusieurs espèces pour émettre des hypothèses quant à la présence de certaines réactions chez une espèce indépendante [107]. En outre, certaines méthodes utilisent l'homologie entre les séquences génomiques des différentes espèces pour proposer des enzymes et des réactions manquantes. En raison de sa robustesse en matière de contenu biologique, l'approche ascendante a été favoriser par la majorité des bases de données et en particulier : *MetaCyc* (*Metabolic Pathways From all Domains of Life*) [108], *KEGG, Reactome, Recon 2, WikiPathways, et LMSD*.

Approche descendante

L'approche descendante se base sur les données obtenues par les technologies "omiques" pour inférer une reconstruction globale du métabolisme puis effectuera, en second lieu, une vérification manuelle de son contenu. Quelques méthodes algorithmiques ont été développées, par exemple, le *GGM* expliqué à la section 1.3.2.1 pour modéliser les voies métaboliques à partir de jeux de mégadonnées provenant d'études de métabolomique [75]. Dans un autre registre, la base de données *HumanCyc* a été bâtie avec l'outil *PathoLogic*, faisant partie de la suite d'outils *Pathway Tools* (http://bioinformatics.ai.sri.com/ptools/), qui utilise l'homologie des séquences génomiques et l'annotation fonctionnelle de *Gene Ontology* (*GO)* [109] pour inférer les voies métaboliques qui devraient être présentes chez l'espèce choisie, l'humain dans ce cas-ci, en se basant sur les voies métaboliques de la base de données multi-organismes *MetaCyc* [98]. L'outil *PathoLogi*c implémente aussi un algorithme pour la gestion de l'information manquante.

1.4.3 Bases de données

Plusieurs des bases de données existantes répertoriant l'ensemble des voies métaboliques connues ainsi que les reconstructions du métabolisme ont été mentionnées au travers des sections précédentes. Toutefois, plusieurs facteurs communs doivent être pris en compte lors de la conception de ces bases de données, soit : 1) la qualité de l'information utilisée, 2) la représentation de l'information manquante ou incertaine, 3) la représentation des complexes enzymatiques, 4) l'intégration des données génétiques, 5) le système d'annotation et d'identification des métabolites, des gènes et des réactions, 6) la distinction entre les différents compartiments cellulaires, 7) l'annotation du sens des réactions et 8) l'inclusion des cofacteurs dans l'annotation des réactions. Chaque base de données a été développée avec un but particulier qui a orienté les choix fait quant aux contraintes utilisées dans l'annotation de l'information selon les facteurs mentionnés ci-haut. En conséquence, chaque base de données a des forces et des faiblesses qui conditionnent l'usage que l'utilisateur en fera.

Le développement de notre méthode a nécessité l'utilisation d'une reconstruction du métabolisme de l'humain. La base de données *KEGG* a été préférentiellement utilisée pour la première phase de développement pour la qualité de son contenu, la simplicité et l'efficacité de son *API* (*application program interface*) ainsi que pour son interface visuel. Au début de ce projet, la base de données *Recon* 2 n'était pas disponible, cependant depuis sa sortie, *Recon* 2 est devenu la référence en matière de reconstruction du métabolisme humain. En effet, la reconstruction du métabolisme de *Recon 2* est plus complète que celle de *KEGG*, particulièrement en matière de lipides, c'est pourquoi nous l'avons sélectionné dans un deuxième temps pour faire des essais

préliminaires avec notre méthode. Ici, seront discutés plus en détails des deux bases de données *KEGG* et *Recon 2*.

1.4.3.1 *KEGG*

KEGG est une base de données contenant des reconstructions du métabolisme de l'ensemble des espèces ayant une séquence génomique complète, incluant celle de l'humain, en plus de quelques organismes ayant des séquences incomplètes. La première version de *KEGG*, publiée en 1999, contenait majoritairement des reconstructions de voies métaboliques. Les voies de signalisation ainsi que d'autres processus cellulaires ont été ajouté progressivement par la suite. *KEGG* a été bâti pour satisfaire les objectifs suivants : 1) construire les voies de signalisations et les voies métaboliques pour mettre en liaison les données génétiques et métaboliques, 2) établir un catalogue répertoriant les gènes avec une annotation constante par rapport à leurs rôles dans les voies construites, 3) établir un catalogue répertoriant les métabolites avec une annotation constante par rapport à leurs rôles dans les voies métaboliques et de signalisation et 4) offrir des outils additionnels pour l'interprétation des mégadonnées, la prédiction des systèmes biologiques et pour aider dans la conception d'approches expérimentales [47].

La reconstruction du métabolisme de *KEGG*, appelé *Metabolic pathways* de la section *Global and overview maps* (vue d'ensemble du métabolisme), qui contient la majorité des réactions métaboliques connues pour chaque espèce. En outre, son outil de visualisation offre une carte qui superpose l'ensemble des reconstructions du métabolisme qu'elle répertorie. Il est ensuite possible de mettre en évidence la reconstruction d'un organisme particulier, par exemple la reconstruction du métabolisme de l'humain répertorié par *KEGG* (figure 1.2), où les 977 réactions (arêtes) annotées chez l'humain et les 943 métabolites (sommets) de ces réactions sont

colorées. De plus, les réactions sont classées selon les voies du métabolisme suivantes : le métabolisme des nucléotides, des glucides, des lipides, énergétique, des acides aminés, de la biosynthèse des glycanes, des cofacteurs et des vitamines, de la biosynthèse des métabolites secondaires, des polycétides et terpénoïdes ainsi que de la biodégradation des xénobiotiques (figure 1.2). Parallèlement, *KEGG* répertorie également plusieurs voies métaboliques individuelles, un exemple chez l'humain est la voie de la biosynthèse des acides gras insaturés (http://rest.kegg.jp/get/hsa01040/image), qui ne fait pas partie de la reconstruction du métabolisme.

L'annotation des réactions métaboliques que propose *KEGG* ne spécifie pas les cofacteurs et les complexes enzymatiques nécessaires à la production d'une réaction, ce qui se traduit par une diminution de la complexité des voies métaboliques et qui facilite l'analyse mathématique de modélisation du métabolisme (section 1.5.2). En plus, l'annotation de la direction dans laquelle se produisent les réactions n'est pas constante, *i.e* une réaction présente dans deux voies métaboliques distinctes peut avoir une direction dans une et être bidirectionnelle dans l'autre. Lorsque l'information est manquante, les réactions sont annotées par défaut comme étant bidirectionnelles. Les voies métaboliques et de signalisation répertoriées par *KEGG* ont été construites en utilisant une approche ascendante et elles ne tiennent pas compte de la compartimentation intracellulaire. Bien qu'un avantage de *KEGG* soit la valeur biologique de son contenu, les contraintes strictes quant à son contenu, peuvent conduire à une perte d'information pertinente. Notamment, KEGG annote une enzyme à une espèce, par exemple l'humain, seulement si le gène qui encode cette enzyme est connu de cette espèce. Autrement dit, même si la preuve expérimentale de la présence d'une réaction et/ou d'une enzyme existe, si le gène encodant cette enzyme est inconnu ou incertain, cette réaction métabolique ne sera pas

répertoriée comme étant présente chez l'humain. En revanche, elle pourrait être annotée à une autre espèce présente dans *KEGG*. De plus, l'information manquante pourrait également être présente dans d'autres bases de données qui appliquent des règles différentes. La visualisation de la reconstruction du métabolisme humain (figure 1.2) permet de générer des hypothèses quant aux réactions annotées aux autres espèces (parties grises) qui pourraient se retrouver chez l'humain. Un autre outil intéressant est le *KEGG Mapper* [110] qui permet de cartographier les gènes et les métabolites d'intérêts sur une voie biologique choisie.

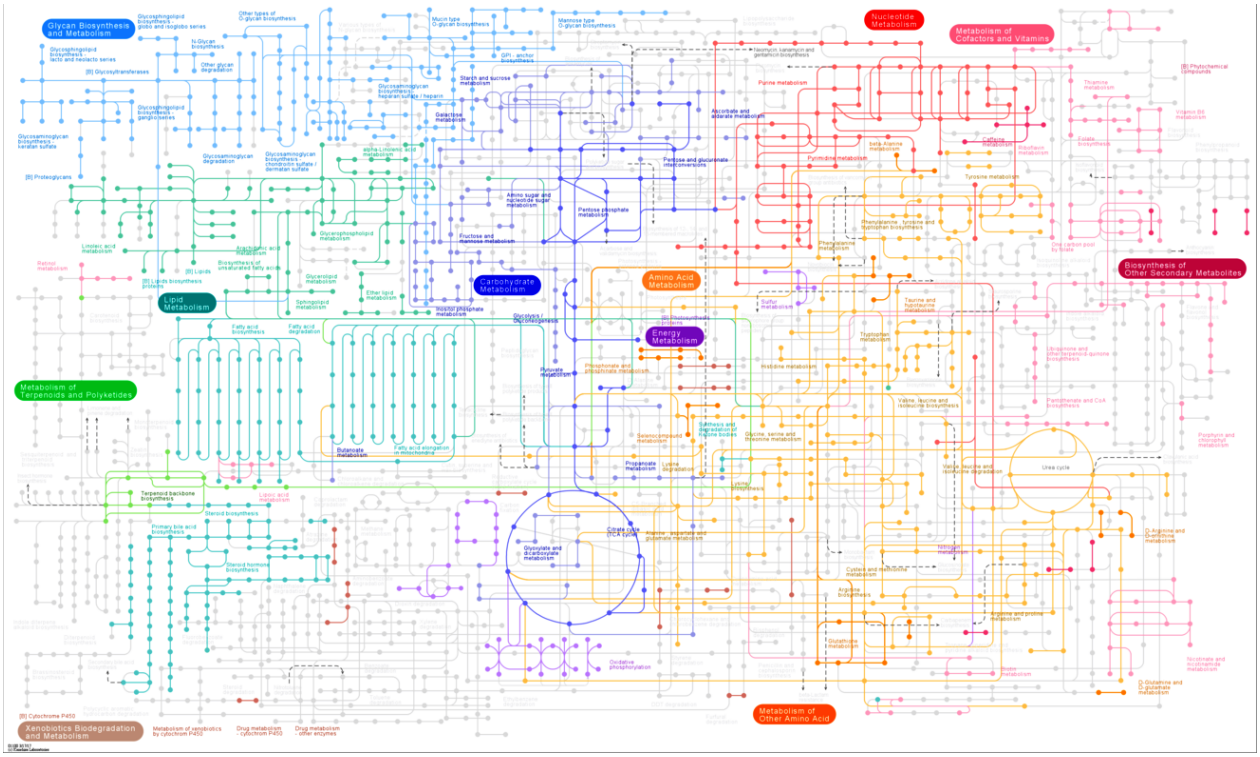

Figure 1.2: Cette carte est une représentation du métabolisme de l'humain de KEGG appelée l'*overview.*
Dans cette représentation graphique, les couleurs différencient les voies métaboliques, les arêtes représentent les gènes (enzymes, réactions) et les sommets représentent les métabolites. Les arêtes et sommets de couleur grise ne font pas partie du métabolisme de l'humain.

L'extraction des données de *KEGG* ce fait aisément à l'aide de l'*API KEGGREST* (http://www.kegg.jp/kegg/rest/keggapi.html) et les fichiers sont encodés sous le format *KGML (KEGG Markup Language)* qui est une variante des fichiers *XML (eXtensible Markup Language)*.

1.4.3.2 *Recon* 2

Recon 2 est une base de données qui répertorie la reconstruction du métabolisme de l'humain la plus complète à ce jour. Le développement de *Recon 2* est issu d'un effort de collaboration au sein de la communauté scientifique ayant contribué à l'annotation et à la compréhension du métabolisme humain principalement, mais aussi d'autres espèces [87, 111]. *Recon 2* est la deuxième version de la base de données *Recon 1* [112]. Cette dernière a été construite avec l'objectif de réunir la majorité des connaissances sur le métabolisme humain pour devenir la référence dans le domaine. Un autre objectif principal de *Recon 2* est de faciliter la modélisation dynamique *in silico* du métabolisme afin de simuler celui de la cellule. C'est pourquoi, une grande importance a été porté au balancement stœchiométrique des réactions métaboliques répertoriées [113-115]. Une autre de ses caractéristiques principales est qu'il est possible de simuler le métabolisme de différents types de cellules dépendamment des gènes qu'elles expriment. Ceci est possible grâce à l'utilisation de certaines données "omiques" tels que la transcriptomique et la protéomique [88].

Initialement, *Recon 1* a été construite avec comme point de départ la base de données *KEGG*, l'annotation *E.C.* des enzymes, la reconstruction du métabolisme de la bactérie *E. coli* ainsi qu'un important travail de vérification manuelle de plus de 1500 articles de la littérature. Au contraire de *KEGG*, la reconstruction de *Recon 1* met l'emphase sur la compartimentation intracellulaire des voies métaboliques. Le métabolisme ainsi que la séquence génomique d'*E. coli* sont

extrêmement bien définis, c'est pourquoi il a servi de référence pour l'élaboration des compartiments cellulaires dans la reconstruction du métabolisme humain. De plus, *Recon 2* spécifie les cofacteurs nécessaires au bon fonctionnement des réactions biochimiques, les complexes enzymatiques ainsi que la direction des réactions. Ces informations sont pertinentes car elles permettent d'avoir une compréhension réelle du métabolisme, cependant elles apportent une complexité supplémentaire pour l'analyse des résultats, ce qui sera discuté à la section 1.5. *Recon 2* a récemment publié une version visuelle et interactive de leur reconstruction du métabolisme, appelé *ReconMap* [116]. En créant un compte personnel, un utilisateur peut, entre autres, cartographier une liste de plusieurs gènes, métabolites ou réactions afin de comprendre leurs interactions dans le métabolisme.

La version actuelle, *Recon 2.2* [117], est plus vaste que celle de *KEGG* et répertorie chez l'humain 5324 métabolites, 7785 réactions ainsi que 1675 gènes associés. L'encodage utilisé pour cataloguer *Recon 2* est le langage *SBML* [118] qui est majoritairement utilisé par la communauté scientifique.

1.4.4 Comparaison et consensus des bases de données

Il a récemment été démontré que le contenu en matière de gènes, d'enzymes, de métabolites et de réactions diverge grandement entre les cinq bases de données suivantes: *Reactome, KEGG, Recon 1, EHMN et HumanCyc* [119]. Plus précisément, l'ensemble des enzymes répertoriées conjointement dans ces cinq bases de données est de 1410, pourtant seulement 18% des enzymes sont communes aux cinq bases de données [24]. De plus, 448 enzymes ne sont retrouvées que dans une des cinq bases de données. En outre, pour les gènes, les métabolites et les réactions, les pourcentages de contenu commun sont moindres que celui des enzymes avec 13%, 9% et 3%

respectivement. Parallèlement, le contenu commun par voie métabolique au travers de ces bases de données est également relativement faible. Pour illustrer cette problématique, le cycle de Krebs est utilisé comme exemple puisqu'il a été largement étudié par la communauté scientifique. Les pourcentages de contenu en commun entre les bases de données pour cette voie réactionnelle sont de 36% pour les gènes, 30% pour les enzymes, 18% pour les métabolites et 17% pour les réactions [24] ce qui est extrêmement faible pour une voie métabolique autant étudié. De plus, récemment, une version consensus du cycle de Krebs a été élaboré en comparant les représentations de ce dernier venant de 10 bases de données ainsi que de plusieurs livres de biochimie [120]. Cette étude a démontré le manque de similitude entre la représentation du cycle de Krebs pour ces bases de données puisqu'elles avaient seulement trois réactions en commun. Une partie de la différence entre les sources s'explique par l'inclusion ou l'exclusion de certaines réactions dans le cycle de Krebs. A titre d'exemple, *KEGG* inclut la conversion du pyruvate en acétyl-CoA dans le cycle de Krebs alors que *Recon 1* et *EHMN* annotent cette réaction dans la glycolyse. Une autre explication est la différence entre l'annotation du nombre d'étapes intermédiaires pour la conversion d'un métabolite en un autre. Par exemple, la conversion du citrate en isocitrate est annotée sans réaction intermédiaire dans *Recon 1* et *Reactome*, avec une réaction intermédiaire dans *HumanCyc* et *KEGG* et elle est annoté de deux façons dans *EHMN* notamment une version sans étapes intermédiaires et une version avec deux étapes intermédiaires [24]. Finalement, la différence peut être expliquée par une inconsistance entre le nombre de substrats et de produits pour une même réaction. Prenons par exemple la réaction catalysée par l'enzyme glucokinase qui transforme le glucose en glucose-6-phosphate. Cette réaction utilise un groupement phosphate de l'ATP (co-substrat), qui devient de l'ADP (co-produit), pour l'ajouter

au glucose. Les bases de données qui incluent les cofacteurs dans les voies métaboliques annotent l'ATP et l'ADP comme étant un substrat et un produit respectivement [119], ce qui crée des contradictions entre les bases de données. En conséquence, il faut être très prudent lorsque l'on veut comparer des résultats d'analyses obtenus avec plusieurs bases de données en tenant compte de la différence de chacune dans leur manière de répertorier les voies métaboliques.

1.5 La théorie des graphes appliquée aux réseaux métaboliques

Un graphe [121] est une entité mathématique définie par un ensemble de points, nommés sommets, ces derniers sont reliés par des traits, nommés arêtes. Les sommets représentent un ou plusieurs types d'informations tandis que les arêtes représentent le lien entre deux sommets. Par exemple, le réseau routier d'une ville peut être représenté mathématiquement par un graphe, où les sommets représentent les intersections et les arêtes représentent les rues qui relient les intersections. De plus, il est possible d'ajouter de l'information sur les arêtes, pour le réseau routier spécifiquement, l'ajout de la direction des rues (sens unique ou bidirectionnel) est essentiel et formerait un graphe orienté. En outre, l'ajout d'informations quant à la densité de la circulation en temps réel (fluide, ralenti ou congestion) est possible, comme le fait le système de géolocalisation *Google Maps*. En pratique, l'ajout de la densité de la circulation routière se fait par l'ajout d'un nombre représentant le temps nécessaire pour parcourir la distance que l'on nomme poids. Finalement, une application bien connue des graphes du réseau routier est de calculer le court chemin en partant du point A au point B.

La théorie des graphes [122] est la discipline mathématique qui étudie leurs propriétés. Cette discipline est largement étudiée et utilisée au sein de nombreuses sphères de notre vie quotidienne, notamment dans les réseaux sociaux, les télécommunications, l'intelligence

artificielle, les applications de géolocalisation ainsi que dans les réseaux biologiques. Récemment, la croissance exponentielle des données récoltées grâce aux différentes technologies "omiques" a fait naître l'intérêt de représenter les données sous forme de graphe pour en extraire l'information pertinente par l'application des méthodes d'analyse provenant de la théorie des graphes. Notamment, les graphes sont utilisés pour représenter une grande variété de données biologiques, voici quelques exemples: les réseaux d'interactions protéine-protéine [123, 124], les réseaux d'interactions électriques entre les différentes zones du cortex cérébral [125], les réseaux de régulation de l'expression des gènes [76] ainsi que les réseaux de voies métaboliques [122]. Les applications reliées à l'analyse des réseaux biologiques sont diverses comme la prédiction de cibles thérapeutiques ou encore la prédiction de la fonction des gènes ou des protéines. Notamment, pour ce projet, la reconstruction du métabolisme humain de *KEGG* a été utilisée ainsi que la métrique du calcul du court chemin. Dans cette section, les différents types de graphes utilisés pour la représentation des réseaux métaboliques seront discutés, ce qui sera suivi des métriques servant à la caractérisation de la topologie des graphes en mettant l'emphase sur la métrique du court chemin, et les différents algorithmes existant pour le calculer et finalement, nous discuterons de l'interprétation biologique résultant de ces analyses.

1.5.1 Type de graphes pour la représentation des réseaux métaboliques

L'information peut être représentée différemment dépendamment du type de graphe. En ce qui concerne les réseaux métaboliques, ces derniers sont majoritairement représentés par trois types de graphe : l'hypergraphe, le graphe biparti et le graphe de composés (*compound graph*) [122, 126]. Pour illustrer les différents types de graphes ainsi que l'impact sur le calcul de court chemin, la réaction catalysée par l'enzyme glucokinase sera utilisée (figure 1.3). La réaction de

phosphorylation catalysée par la glucokinase implique quatre métabolites; les deux métabolites *principaux,* conversion du glucose en glucose-6-phosphate et de l'ATP en ADP pour les métabolites secondaires enzymatiques. Dépendamment de la base de données utilisée, les réactions enzymatiques peuvent être annotées par le nom de la réaction, l'identifiant de la réaction, l'identifiant de l'enzyme qui catalyse la réaction ou encore par le nom ou l'identifiant du gène qui encode l'enzyme. Par souci de simplicité, le terme réaction sera utilisé dans ce texte. Il faut également noter que tous les types de graphes peuvent avoir des arêtes orientées, nommées arcs, et avoir un poids sur leurs arrêtes.

L'hypergraphe

L'hypergraphe est celui offrant la représentation la plus réaliste des voies métaboliques car il a la particularité de permettre aux arêtes de relier plus de deux sommets, comme illustré à la figure 1.3 A. Les sommets de l'hypergraphe représentent les métabolites alors que les arêtes représentent les réactions. Par exemple, la réaction catalysée par la glucokinase implique deux substrats (glucose et ATP) et produits (glucose-6-phosphate et ADP) étant reliée par une même arête. Il est important de noter que la distinction entre les métabolites principaux et les métabolites secondaires n'est pas conservée dans l'annotation de l'arête ce qui est problématique pour le calcul d'un court chemin (discuté à la section 1.5.3). Malgré la qualité du modèle de l'hypergraphe pour représenter les réseaux métaboliques, il est très peu utilisé pour les représenter dû à la complexité des algorithmes qui doivent gérer les hyperliens [126].

Graphe biparti

Le graphe biparti (figure 1.3 B) est un graphe où les sommets représentent 2 ou plusieurs types d'informations. Pour les réseaux métaboliques, les sommets représentent les métabolites et les réactions.

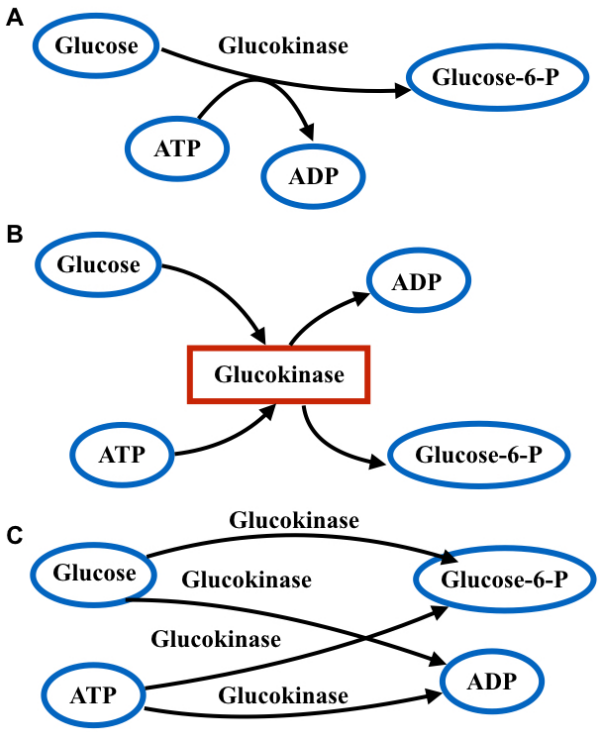

Figure 1.3: Les différents types de graphes utilisés pour représenter les réseaux métaboliques : exemple de la glucokinase.
A représentation graphique de l'hypergraphe, **B** représentation graphique du graphe biparti et **C** représentation du graphe de composés. Tiré de "*Computational methods to identify metabolic sub-networks based on metabolomic profiles*" par C. Frainay et F. Jourdan [126].

Chaque arête lie un métabolite et une réaction seulement si ce métabolite est un substrat ou un produit de cette réaction. Ainsi, la réaction catalysée par la glucokinase est représentée par cinq sommets et quatre arêtes. Par ailleurs, la complexité du calcul du court chemin est semblable à celle du graphe de composés.

Graphe de composés

Le graphe de composés est le graphe le plus couramment utilisé pour représenter les réseaux métaboliques. Dans un graphe de composés, les sommets représentent les métabolites et les arêtes représentent les réactions, une arête existe entre deux métabolites seulement s'ils forment une paire de substrat/produit appartenant à une réaction métabolique. La réaction catalysée par la glucokinase est représentée par quatre arêtes (figure 1.3 C). Malgré la grande utilisation du graphe biparti et du graphe de composés pour leur performance au niveau algorithmique, ils représentent moins fidèlement le réseau métabolique. Notamment, dans le graphe de composés, la même enzyme (glucokinase) se retrouve sur plusieurs arêtes (figure 1.3 C). Néanmoins, le graphe biparti et le graphe de composés sont très utiles lorsque l'utilisateur comprend et incorpore ses limitations lors de l'analyse.

1.5.2 Caractérisation des réseaux métaboliques

Un graphe peut être caractérisé par plusieurs métriques dont les suivants :

Le court chemin moyen : le court chemin est défini par $d_G(x,y)$, où x et y représentent deux sommets appartenant au graphe (G). La moyenne est obtenue en calculant $d_G(x,y)$ pour chaque paire de sommets et divisée par le nombre total de paires.

Le diamètre : le court chemin maximal du graphe G.

Coefficient de connexion : Le coefficient de connexion du sommet *x* est la proportion des arêtes existantes sur l'ensemble des arêtes possibles entre les sommets voisins de *x*. Cette métrique permet d'identifier les sommets très connectés d'un graphe.

Une caractéristique pertinente aux graphe représentant les voies métaboliques est le degré des sommets car il est possible d'identifier les métabolites impliqués dans plusieurs réactions qui sont généralement des cofacteurs.

Le degré : le degré du sommet *x* est le nombre d'arêtes du sommet. Cette métrique permet d'identifier les sommets très connectés versus les sommets très peu connectés.

À ce jour, la caractérisation de la topologie des graphes représentant les réseaux métaboliques n'est pas unanime, ce qui n'est pas surprenant étant donné notre connaissance partielle du métabolome. Toutefois, une comparaison du réseau métabolique de 43 organismes a démontré que la topologie des réseaux métaboliques répond à la distribution d'une loi de puissance, ce qui est une caractéristique importante des réseaux sans échelle (*scale-free*) [127]. Plus précisément, la proportion de noeuds de degré *k* est proportionnelle à $k^{-\gamma}$ pour *k* grand, où γ est se situe généralement entre 2 et 3. En revanche, une étude subséquente plus poussée de ces 43 réseaux métaboliques a démontré que les coefficients de regroupement de ces réseaux sont plus élevés que le coefficient de regroupement attendu pour les réseaux sans échelle [128]. Ceci suggère une bonne modularité, autrement dit, un faible nombre d'interactions entre les différentes voies métaboliques qui sont représentées par les modules.

1.5.3 Les algorithmes de calcul du court chemin

Les algorithmes discutés dans cette section s'appliquent aux graphes bipartis et aux graphes de composés. Le choix de l'algorithme utilisé dépend de la direction des arêtes ainsi que du poids des arêtes. Ainsi, pour un réseau métabolique comme celui de la base de données *KEGG,* où les arêtes ne sont pas orientées et n'ont pas de poids, l'algorithme *Dijkstra* [129] ou l'algorithme de recherche en largeur *(Breadth-first search)* [129] sont appropriés pour calculer le *srd (shortest reactional path*). De plus, cette reconstruction ne contient pas de cofacteurs, ainsi les chemins réactionnels trouvés ont une pertinence en termes de suite de réactions enzymatiques. Par contre, lorsque les graphes sont bâtis à partir de réseaux métaboliques contenant des cofacteurs comme celui de *Recon 2*, la complexité du calcul du court chemin est augmentée. Ceci, parce qu'il faut éviter les chemins n'ayant pas de pertinence biologique, c'est-à-dire, des chemins réactionnels qui passeraient par des cofacteurs qui sont partagés par un grand nombre de réactions enzymatiques. De plus, ces bases de données n'identifient généralement pas distinctement les métabolites principaux des cofacteurs, donc, il n'est pas possible de seulement les retirer du graphe ou de les ignorer lors du calcul. Une solution intéressante pour résoudre ce problème est d'implémenter un algorithme qui choisit son chemin en fonction de la conservation du nombre de carbone entre le métabolite actuel et le prochain [126]. Pour cet algorithme, un poids entre 0 et 1 est ajouté sur chaque arête en fonction du nombre de carbone du substrat et du produit. Par exemple, une arête liant un substrat et un produit ayant un nombre similaire de carbone aurait un poids près de 0 tandis qu'une arête reliant un ou deux cofacteurs aurait un poids près de 1. Ainsi, les algorithmes de *Dijkstra* et de recherche en largeur peuvent être adaptés pour identifier le

chemin le plus court qui minimise le poids sur les arêtes. Ces deux algorithmes peuvent également gérer les contraintes d'orientation des arêtes (arcs).

1.5.4 L'interprétation des résultats dans un contexte biologique

L'interprétation des résultats obtenus avec le calcul du court chemin dépend grandement des caractéristiques (cofacteur, orienté) du réseau métabolique utilisé. Généralement, lorsqu'un court chemin est identifié entre deux métabolites ou entre une réaction et un métabolite, cela suggère qu'il existe un lien biologique entre eux. Évidemment, la validation manuelle des chemins trouvés est préférable. La sélection d'un seuil pour définir un court chemin dans un réseau est subjective et dépend du type de graphe ainsi que de ses caractéristiques. Par exemple, pour un graphe où le court chemin moyen est de 10, un *srd* significativement court pourrait être entre 0 et 5. En revanche, un *srd* plus grand que cinq pourrait également être pertinent par exemple si le gène et le métabolite concernés font partie d'une longue voie réactionnelle de biosynthèse. Également, il est possible qu'il n'existe aucun chemin entre un gène et un métabolite donné (*srd* = infini), ce qui peut signifier qu'il n'existe pas de lien biologique entre ce gène et ce métabolite ou que le lien biologique est manquant dans le réseau, ceci serait dans ce cas dû à l'incomplétude des réseaux métaboliques que l'on utilise et peut permettre la génération d'hypothèse.

Objectif et Hypothèse

L'objectif de ce travail était de développer une approche bio-informatique basée sur l'architecture génétique du phénotype étudié afin de prédire un ensemble de métabolites pertinent à analyser pour permettre une meilleure caractérisation de ce dernier. Ce travail a été entamé dans le cadre d'une étude multidisciplinaire concernant la maladie de Crohn combinant des analyses de protéomique, d'immunologie, d'histologie, de métabolomique et de génomique fonctionnelle par lesquelles 163 loci la caractérisant ont été identifiés [130]. L'objectif global de la caractérisation de ce phénotype au niveau moléculaire, cellulaire et physiologique est d'améliorer la prédiction de réponse ou non-réponse des patients aux différents traitements pharmacologiques existants. Concernant la métabolomique, il est absolument nécessaire de cibler un ensemble de métabolites pertinents à analyser considérant, notamment, le fait qu'il n'est pas possible d'analyser l'ensemble des classes de métabolites avec un seul outil analytique et un seul protocole. Au meilleur de nos connaissances, il n'existe pas d'outil idéal pour réaliser ce travail.

Pour développer un outil prédictif nous avons tout d'abord émis l'hypothèse que les gènes encodant des enzymes catalysant des réactions métaboliques ont un impact sur la concentration des métabolites à proximité de ces réactions dans les réseaux métaboliques. Nous avons testé cette hypothèse en développant une méthode permettant de cartographier les gènes et les métabolites sur les voies métaboliques de la base de données *KEGG* afin de calculer le court chemin réactionnel (*srd*) entre eux. Nous avons appliqué notre méthode aux associations statistiques entre *SNP* (annoté à leur gène putatif) et métabolite rapporté par l'étude *mGWAS* réalisée Shin *et al.* afin d'évaluer la pertinence de la métrique *srd* pour annoter et comprendre le lien biologique entre les gènes et les métabolites. L'étape de prédiction d'un ensemble de

métabolites basé sur 163 loci connus pourra être réalisée sur la base des connaissances acquises lors de ce travail.

2. Critical assessment of the shortest path metric for annotation of gene-metabolite pairs revealed by mGWAS

2.1 Avant-propos

La méthode présentée dans cette article a été élaborée par Sarah Cherkaoui, une ancienne étudiante à la maîtrise du laboratoire dirigé par Dr Christine Des Rosier. Ma contribution à ce projet se définit par le développement de la librairie *PathQuant*, par l'utilisation de l'ensemble des cartes répertoriant des voies métaboliques de *KEGG* ainsi que par l'analyse approfondie des résultats obtenus avec notre méthode.

Critical assessment of the shortest path metric for annotation of gene-metabolite pairs revealed by mGWAS

Sarah Cherkaoui [1,2,8 *] & Sandra Therrien-Laperriere[1,2 *], Matthieu Ruiz1[2,3], Kwanjeera Wanichthanarak[5], Dmitry Grapov[6], Clément Frainay[7], Gabrielle Boucher[2], The iGenoMed Consortium, Fabien Jourdan[7], Guillaume Lettre[2,4], John David Rioux[1,2,4], Christine Des Rosiers[1,2,3]

[1] Département de Biochimie et de Médecine Moléculaire, 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] West Coast Metabolomics Center, Genome Center, University of California Davis, Davis, California, USA

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

[7] Toxalim, Université de Toulouse, INRA, Université de Toulouse 3 Paul Sabatier, Toulouse, France

[8] Current address: Institute of Molecular Systems Biology, ETH. Zurich, Switzerland

*****Equally contributing authors

2.2 Abstract

Background & Aim: Recent studies combining metabolomic and genome-wide associations (GWAS), named mGWAS, have provided valuable insight into our understanding of the genetic control of metabolite levels. mGWAS report associations between single nucleotide polymorphisms (SNPs) (annotated to a putative gene) and metabolites, but the biological interpretation of theses associations still remains challenging. The objective of this study was to provide a critical assessment of KEGG database to report relevant biological information to explain mGWAS data using the shortest path metric and pathway mapping. ***Methods & Results*:** For this, we have used a bioinformatics package that we developed, PathQuant, which enables a robust, systematic and automated computation of the shortest reactional distance (srd) values between any list of pairs of genes (encoding for an enzyme) and metabolites mapped onto KEGG metabolic pathway maps, while keeping their original and well curated topology. Using data from Shin et al. (2014), which is one of the largest mGWAS, we found that most (73%) of the mapped reported associated pairs between a gene, which encoded for a metabolic enzyme, and a metabolite had a shortest reactional distance (srd) value ≤5. This indicated, as expected, a close biochemical relationship between a gene encoding enzyme and its associated metabolite. There were, however, other statistically significant associated pairs that had longer or infinite srd values. Manual curation of these data highlighted some limitations and annotation errors of KEGG and the utility of PathQuant in identifying them. **Conclusion:** PathQuant computes srd values between any list of pairs of genes (encoding for an enzyme) and metabolites (statistically associated or not) mapped into KEGG. This metric provides an objective and quantitative annotation of these gene-metabolite pairs within the current biochemical knowledge. Future

development of PathQuant includes addition of other metabolic pathway databases such as Recon2, as well as signaling pathways, to enhance coverage and value of the srd metric.

Keywords: Metabolic network – R package - KEGG - mGWAS – network-based analysis – graph theory - metabolites

2.3 Backgroung

The application of high throughput (HT) technologies and approaches, such as genome-wide association studies (GWAS), in large human cohorts has also contributed significantly to the identification of the genetic risk factors for many common diseases. Unraveling the underlying mechanisms remains, however, a challenge. This is currently tackled by combining GWAS data with other omics data, such as transcriptomics, epigenetics and metabolomics. Metabolomics offers a means to systematically measure thousands of low-molecular-weight compounds, in order to provide a global view of a given perturbation whether resulting from a gene mutation or disease onset. While the combination of dataset from any of the aforementioned omic analyses has its value [66, 67, 131], in this work, we focus on data revealed by combining GWAS with metabolomics referred to as mGWAS.

mGWAS report associations between metabolites and single nucleotide polymorphisms (SNPs), referred to as metabolite quantitative trait loci (mQTL), SNPs are subsequently annotated to putative causal genes [132]**.** Although very powerful, mGWAS [94, 133] have further increased the quantity, diversity and complexity of data to be processed and interpreted biologically. To date, few methods have been developed in order to gain insight into the biological interpretation of mGWAS data [94]. These include: 1) Gaussian Graphical Modelling (GGM) [75], which infers connectivity between metabolites from metabolomics data; and 2) “Mining the unknown” [95], which enables the identification of unknown metabolites using mGWAS, GGM and metabolic pathway databases. While insightful, these methods do not provide information based on well curated current biochemical knowledge.

The KEGG database, which is recognized for its high curation level of metabolic pathways for many model organisms, including humans, has been successfully used to perform analysis on individual omics datasets such as Gene Set Enrichment Analysis [15] and Metabolic Set Enrichment Analysis [16]. While KEGG has also been used for manual biological annotation of mGWAS association data [10], to the best of our knowledge, there is no method currently available enabling a systematic, objective and quantitative annotation of these data. In this regard, recent studies have highlighted the applicability and utility of topological data analysis based on graph theory approaches [134], including the shortest path [135] in revealing biological knowledge in the context of complex metabolic networks [136-138] and other biological networks [123, 139-143]. For example, this metric was used to characterize the impact of gene deletion on nearby metabolites within the metabolic network of *E. coli* [137]. It was also used to analyze the relationship between expression quantitative trait loci (eQTL) and metabolomic data, first, within the metabolic network of rat adipose tissues [136] and subsequently within merged metabolic and signaling networks [138].

The objective of this study was to critically assess the utility of the KEGG database for calculation of the shortest reactional distance (srd). This was achieved through pathway mapping in providing an objective and quantitative metric, which would reflect current biochemical knowledge and explain statistically associated gene-metabolite pairs revealed by mGWAS. For this, we developed PathQuant, a R package to enable a robust and systematic computation of the shortest reactional distance (srd) values between any list of gene-metabolite pairs (statistically associated or not) mapped onto metabolic pathways of KEGG, while keeping the original, well

curated, topology of each pathway. To conduct this study, we used the dataset from Shin et al. [72], which is among the most complete mGWAS of human blood [73].

2.4 Method

The method was developed using R programming language [144] and delivered as an R package, PathQuant (for Pathway Quantify), which is freely available at *https://github.com/sandraTL/PathQuant* [145]. Figure 2.1 depicts the package workflow for the following steps:

2.4.1 Input parameters

Gene-metabolite pairs: PathQuant accepts the following input for srd computation: 1) genes; and 2) their associated metabolites in columns, each with their specific KEGG identifiers (IDs). Each row represents a unique gene-metabolite association. Only associations between genes and individual metabolites are taken as entry, hence associations between a gene and a ratio of metabolites have to be separated prior to analysis.

Selected metabolic pathways: PathQuant accepts a list of metabolic pathways, each with their specific KEGG IDs. For this application, we used all human metabolic pathway maps available in KEGG, but focused most of our analyses on the metabolism reconstruction pathway map. The latter encompass 931 reactions and 981 metabolites and is named metabolic pathways (ID: has: 01100) in KEGG (referred thereafter as the 'overview' throughout this article).

2.4.2 Association classification

We classify each association of the input by gene product to four broad categories: enzyme, transporter, other (other proteins, transcription factor and more) and not classified using KEGG Brite database [47]. Since our method annotates pairs through metabolic pathway mapping, we focus our analysis on pairs involving genes which encode enzymes. Metabolites are classified

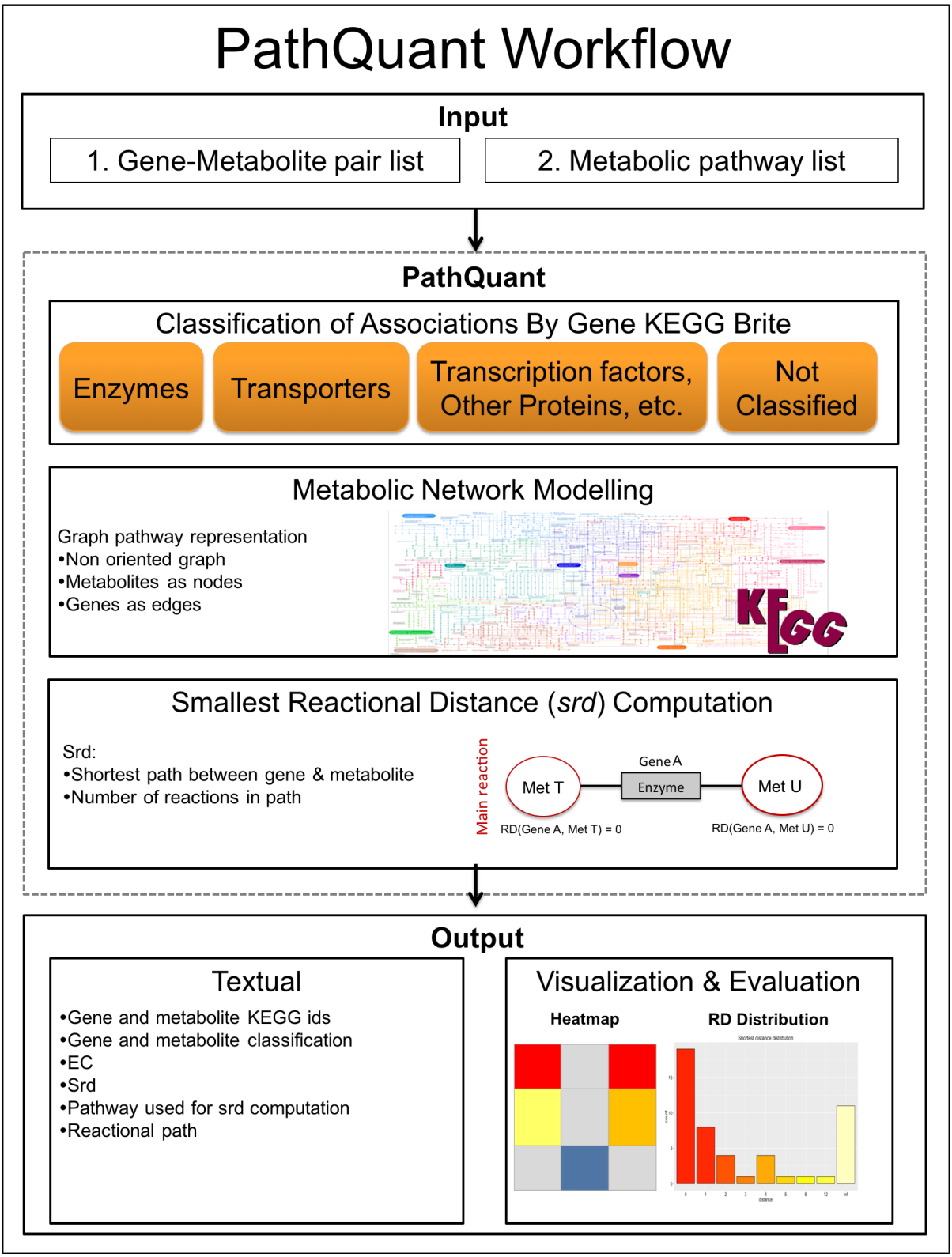

Figure 2.1 Method workflow.

using the chemical class hierarchy of the KEGG Brite database, allowing easy and quick overall understanding of their distribution within metabolic classes.

2.4.3 Metabolic network modelling

The selected KEGG pathway, encoded in KEGG XML file format (KGML), is downloaded using the KEGG API [146], from the most up-to-date KEGG pathways available. The pathway is then converted into a graph of biochemical reactions (also called compound graph [122]) with metabolites, as nodes and genes, mapped to their corresponding encoded enzymes, as edges. The topology of the pathway is captured in the constructed graph. Genes encoding enzymes catalyzing multiple reactions are mapped to multiple edges. The constructed graph represents exactly the metabolic pathways of KEGG which are built mainly with metabolites that are main reactants of a reaction [147], dismissing cofactor metabolites, such as NAD, or common co-substrates/products, such as ATP or H_2O. In fact, cofactors are defined as ubiquitous and are not major components of a transformation series [148]; they are part of many reactions which can cause irrelevant biological paths. Finally, we use a non-oriented graph as the KEGG standards are not consistent in this matter [54]. For example, the gene IDH1 (encoding the isocitrate dehydrogenase enzyme) is present in a few pathway maps of KEGG including the overview and the “Biosynthesis of amino acids”. The enzyme catalyzes the conversion of isocitrate to 2-oxoglutarate, which is annotated as being reversible in the overview and irreversible in the “Biosynthesis of amino acids”. Furthermore, the non-oriented graph gives equal importance to changes in levels of both upstream and downstream metabolites of a given reaction catalyzed by a gene’s encoded enzyme.

2.4.4 Srd computation

Our method computes the srd, which is defined as the shortest reactional distance path between a given gene and a metabolite. The srd is computed using each metabolic pathway received in input, in which a given pair is mapped. The srd is computed as follows:

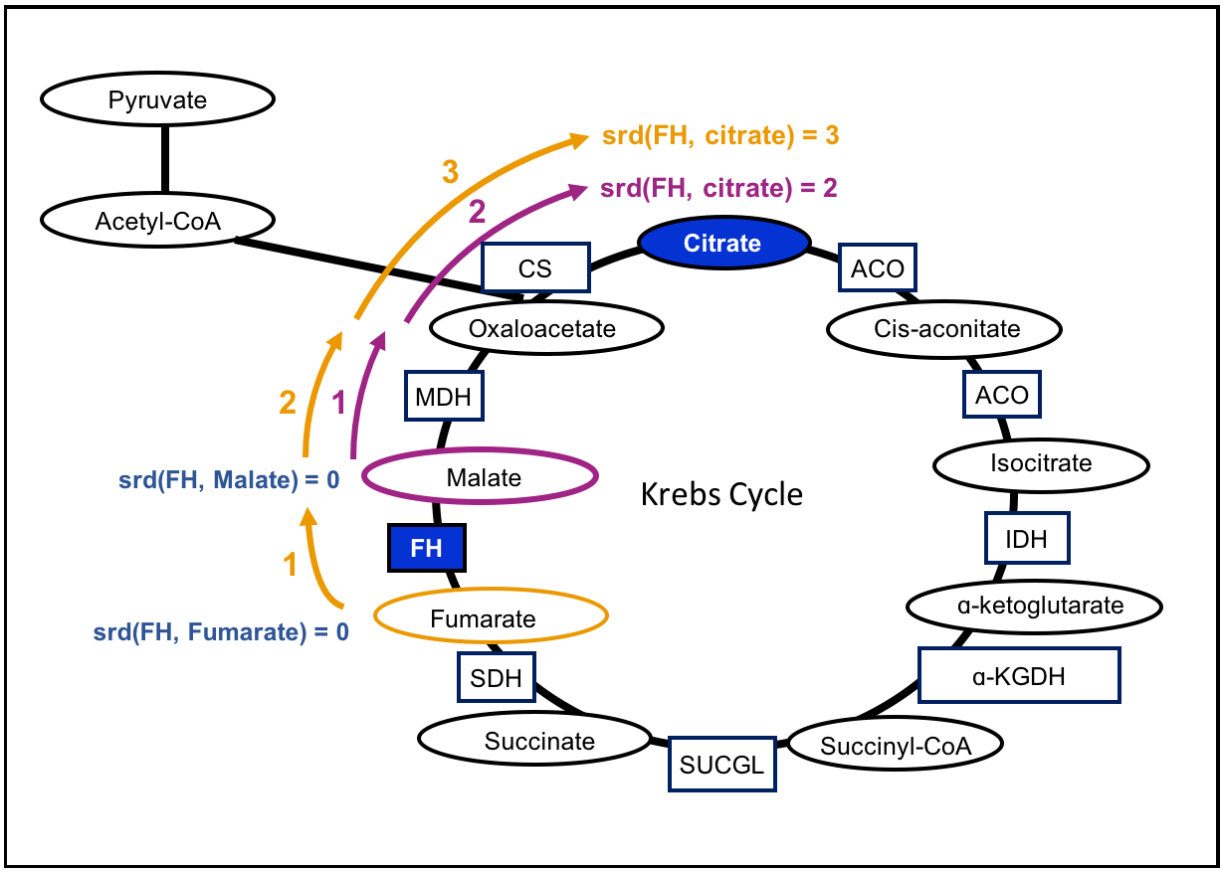

Figure 2.2 Exemple of srd value computation for the gene FH.
This figure depicts the algorithm for the srd computation of a fictive association between the gene FH (in blue) which encodes the enzyme fumarase hydratase and the metabolite citrate (in blue), which are mapped on the Krebs cycle, in which metabolites are represented as nodes and genes mapped to their corresponding enzyme as edges (rectangles). For simplification, some individual genes are not shown but included in their enzymatic complex (for e.g. DLD is both in αKGDH and PDH). The reaction catalyzed by the fumarase hydratase enzyme involves fumarate and malate, therefore the srd is computed from each of them to citrate; 3 and 2 respectively. The smallest srd is 2.

A distance of 0 is assigned to metabolites, which are the main substrates or products of the reaction catalyzed by the enzyme encoded by the selected gene of interest. The srd to all other metabolites are obtained using the breadth-first search algorithm [129]. The algorithm is used to find the srd starting from the substrate and from the product of the mapped gene to the paired metabolite, thereby selecting the smallest srd between these two. Figure 2.2 depicts an example of srd computation for the gene FH, which encodes for the enzyme fumarate hydratase, also called fumarase, and is part of the Krebs cycle. The main reactant pair of this reaction, fumarate and malate, are set at a srd of 0 (colored in red). As example, the computation of the srd for a fictive association between FH and citrate proceeds as follows: 1) the srd is computed between fumarate and citrate (srd = 3), 2) the srd is computed between malate and citrate (srd = 2), 3) the smallest srd between 3 and 2 is selected (srd = 2).

2.4.5 Srd metric analysis

The utility of the computed srd metric for the annotation of gene-metabolite associations reported by mGWAS was assessed using three approaches. (1) All computed srd values obtained for mapped gene-metabolite pairs were manually inspected on the basis of current literature knowledge using KEGG Mapper as a visualization tool [110]. (2) Srd values were compared to the reported SNP (annotated to putative causal gene)-metabolite associations strength in a scatter plot. (3) A randomization test was performed to assess if srd values computed for gene-metabolite associations (referred to as “associated set”) using the overview pathway differ significantly from those obtained by randomly selecting genes within the overview (referred to as “resampled set”). Only the overview was used for this test given that the other maps of KEGG represent specific pathways which contain a limited number of enzymatic reactions. For the selection of genes to be

included in the resampled set, a condition was built to match the number of reactions, with a difference of +/-1, for the replaced gene, except for genes with only 1 reaction, which were replaced by a gene encoding an enzyme catalyzing one reaction. This condition was added because the number of reactions catalyzed by an enzyme encoded by a given gene could bias the calculated srd; for e.g. the odds of having a short srd to a given metabolite is higher for genes encoding enzymes catalyzing multiple reactions. For each resampled set, srd values are computed between the same number of gene-metabolite pairs as the associated set. The statistical significance is obtained by comparing the median srd of the associated set to median srd of resampled sets using the following formula:

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

N_0 denotes the number of resampled sets with median value smaller or equal to the median of the associated sets. N denotes the number of resampled sets. The precision level of the obtained *p*-value will be set according to the number of resampling.

2.4.6 Data outputs and visualization

PathQuant outputs a text file containing gene and metabolite classification, Enzyme Commission number (EC), KEGG Brite, KEGG IDs of used pathways for the srd computation and srd values for all associations. These srd values can also be visualized in a heatmap and a global or multiple distribution plots.

2.5 Results and discussion

To critically assess the utility of KEGG database for srd calculation using pathway mapping of gene-metabolite pairs revealed by mGWAS, we have used data from Shin et al. [72]. The study was conducted on 7824 human (blood samples) and for which 2.1 million SNPs and 529

metabolites (assessed by targeted and untargeted metabolomics) were analyzed. Shin et al. reported 299 SNP-metabolite associations involving 187 unique metabolites and 145 loci, which were annotated by the authors to 132 causal genes (Additional file 1: Tables S1-S3).

2.5.1 Building input data - Identification

Since only gene and metabolite common 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 from LinkDB [149]), for metabolites this step was problematic and ambiguous since metabolites have multiple synonymous names. Thus, the identification for metabolites was achieved manually using KEGG compound database [47] and double checked using the Human Metabolome database (HMDB) [49]. Furthermore, many of the listed metabolites did not have KEGG IDs, among which were acylcarnitine derivatives. These metabolites are principally measured in plasma (or blood samples), but arise from intracellular metabolism of their acyl-CoA counterparts [150]. Since KEGG database encompasses predominantly intracellular enzymatic reactions and their corresponding metabolites, we created a rule to bridge the gap whereby, 16 out of the 21 listed metabolites with names ending with ‘carnitine’ present in KEGG (shown in red in Additional file 1: Table S2) could be replaced by the KEGG ID of their acyl-CoA counterparts (ending with ‘CoA’).

In summary, the 132 genes listed in the study were found in KEGG and given a KEGG ID (see Additional file 1: Table S1). From the 187 unique metabolites that were involved in an association, 136 were identified by the authors; the remaining 51 were listed as unknown [72]. From those 136 identified metabolites, 95 were manually assigned a KEGG ID, among which lipids and amino acids are overrepresented, with 41 and 34 metabolites, respectively (Additional

file 1: Table S2). There was, however, 41 identified metabolites remaining, including 20 lipids, which were not present in KEGG.

As for gene-metabolite associations, Shin et al. reported 299 of them, among which 245 were associations between a single gene and a metabolite, whereas 54 were associations between a gene and a metabolite ratio shown in purple (Additional file 1: Table S3). The latter are pairwise ratios of metabolite concentrations, which were associated with genes in a second genome-wide analysis. For these ratios, it was assumed that the gene was associated with both metabolites; therefore, associations with ratios were separated in two distinct associations. Considering many ratios had one or both metabolites also present in a single association, the total number of single associations remained 299.

2.5.2 Association classification and pathway mapping

The associations were then classified according to the gene product. From the 132 genes, 79 encoded enzymes, the remaining encoded transporters (20) belonging to the Solute Carrier Family (SLC), of which solely 1 encodes an enzyme which is only mapped on signaling pathways, as well as other gene products (16) or unclassified gene product (17). The 79 genes encoding enzymes were involved in 183 associations, out of which, 86 were associated with a metabolite identified in the KEGG database; for the remaining, either the metabolite did not have a KEGG ID (54) or was classified as unknown by the authors (i.e. not chemically identified: 43). We therefore focus the rest of the analysis on the 86 associations where the gene and the metabolite are present in the KEGG database which involves 50 unique genes and 66 unique metabolites.

Table 2.1 Number of genes, metabolites and associations from Shin *et al.* in KEGG

	Gene-metabolite Associations (single)	Genes	Metabolites
In mGWAS Study	299 (54 ratios were separated in 108 associations)	132	187 (136 annotated and 51 unknowns)
Genes encoding enzymes	183	79	Not Applicable
Associations in KEGG	86 (100%)	50 (100%)	66 (100%)
Mapping on KEGG overview	35 (41%)	32 (64%)	36 (55%)
Mapping on all metabolic pathways of KEGG	49 (57%)	46 (92%)	54 (82%)

This table reports: Line 1: The gene-metabolite associations that were reported by Shin et al., the number of uniq genes and uniq metabolites involved in those associations. Line 2: Numbers of genes classified as enzymes using KEGG Brit database and the associations involving those genes. Line 3: Numbers of gene(encoding enzyme)-metabolite associations for which the gene and the metabolite are found in KEGG database, the number of uniq genes and uniq metabolites. These numbers were taken as representing 100% of the dataset used for pathway mapping. Lines 4-5: Numbers (and percentages in parentheses) of genes encoding enzymes, metabolites and associations found in KEGG that were mapped on the overview (Line 4) and all metabolic pathways (Line 5) of KEGG.

Genes and metabolites with KEGG IDs were mapped to the pathway graphs. Out of the 50 genes and 66 metabolites, 32 and 36, respectively, were mapped on the overview (shown in black in Additional file 2: Figure S1). An additional 14 genes and 18 metabolites were mapped on other metabolic pathway maps, for a total mapping of 92% (46) for genes and 82% (54) for metabolites (Table 2.1). There were only 4 genes that were not mapped in any pathways,and were

annotated to the following categories according to KEGG's classification: namely ASPG (asparaginase), ETFDH (electron-transferring-flavoprotein dehydrogenase), MARCH8 (E3 ubiquitin-protein ligase MARCH1/8) and PPM1K (protein phosphatase 1K). The absence of MARCH8 and PPM1K is understandable given that these genes encode for enzymes involved in post-translational modification of proteins via ubiquitinylation and dephosphorylation, respectively; which would not be considered as "main reactions" by KEGG. Similar considerations apply to ETFDH, which encodes for an enzyme of the mitochondrial electron transport chain that is essential for electron transfer from a number of mitochondrial flavin-containing dehydrogenases and uses as substrate a common co-factor metabolite, namely reduced flavin adenine dinucleotide ($FADH_2$). The absence of ASPG is, however, unexpected considering it encodes for the enzyme asparaginase; this will be further discussed in the 'manual curation' section.

A larger number of metabolites (12) were not found on any metabolic pathway maps of which 5 were lipids, 4 were unclassified, 1 was a pesticide and 1 was an anatomical therapeutic chemical (according to KEGG classification). It is noteworthy that although lipids represented 27 of the 66 metabolites, only a limited number of them (10/27) were mapped on the overview, as depicted in Figure S1 (Additional file 2). This illustrates the unequal representation of the various pathway classes in KEGG overview, but also highlights its coverage limitation. As a result, out of the 86 associations with KEGG IDs, 35 were mapped on the overview (Table 1) and an additional 14 on other maps, for a total of 49 associations for which a srd values was computed.

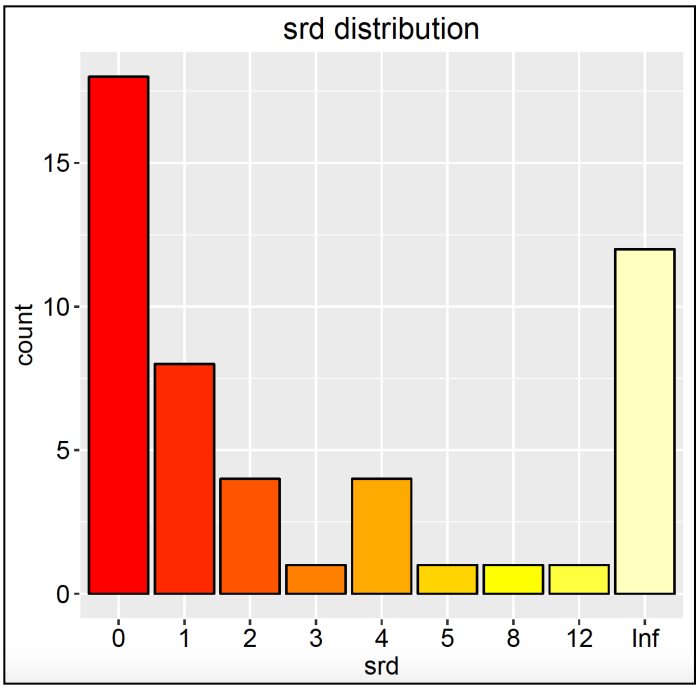

Figure 2.3 Srd distribution for gene-metabolite pairs from Shin et al. using KEGG metabolic pathway maps.
Frequency bars, which represent calculated srds, are shown with colors (from red – closest; to yellow - farthest). Inf = infinite value, which corresponds to no known path between gene and metabolite.

2.5.3 Srd computation and visualization

Calculated srd values by PathQuant between the 35 genes and 42 metabolites, which are part of the 49 associations that were mapped to the KEGG overview and/or other maps, are shown as a distribution plot in Figure 2.3 and reported in Additional file 1: Table S4. In some instances, a given association was mapped to multiple metabolic pathways, thus multiple srd values were computed and the smallest one was chosen. Among the 49 computed srd values for this dataset, 36 were ≤ 5 of which 26 ≤ 1, which indicates close biological relationship between a gene encoding enzyme and its associated metabolite. These results concur with those of Fuhrer et al. [137] showing that single-gene deletions in E. coli are associated to metabolites, which are

within two reaction steps apart, for a substantial fraction of the tested enzymes. There were, however, srd with larger values, namely 8 and 12, and 11 with infinite (inf) values. The meaning of these larger and infinite srd values will be further discussed below in the section "Srd analysis".

Of all mapped associations, 35(71%) were mapped on the overview and are illustrated by a thick black border on the heatmap in Figure 2.4. Fifteen of these associations have a srd of 0, implying that these gene-metabolite pairs, for e.g. PSPH (phosphoserine phosphatase)-L-Serine, are from the same reaction. The heatmap enables also visualization of the relationship between all genes and metabolites of the mapped associations within the pathway. For example, within the 35 associations, some genes are highly connected to many metabolites, such as ACADM (acyl-CoA dehydrogenase) and GLS2 (glutaminase 2) as they are at small srd values (0 < srd <5) from almost all other metabolites. Finally, some genes are isolated and appear to be part of disconnected subgraphs of the overview, such as TDO2 (tryptophan 2,3-dioxygenase).

All the metabolic pathways of KEGG were used to extend further the coverage and knowledge since we were able to map 14 additional associations and compute their srd values (Table 2.1, Additional file 1: Table S4). It is noteworthy that for this, PathQuant uses only individual maps but do not compute srd values by combining 2 or more metabolic pathway maps (i.e. for a given gene and metabolite pairs that mapped on different metabolic pathways). Indeed, we considered that this would break the original topology of KEGG metabolic pathways or/and could possibly create erroneous paths. The importance of using all KEGG metabolic pathway maps individually to compute srd values is highlighted by the following example: the srd for the association TYMP (thymidine phosphorylase)-uridine is 8 in the overview and 1 in the

pyrimidine metabolism map. In the overview, TYMP (EC 2.4.2.4) phosphorylates thymidine to thymine, which is 8 reactions away from uridine, while in the pyrimidine metabolism pathway, TYMP phosphorylates deoxyuridine to uracil which is one reaction away from uridine. While TYMP can phosphorylate both substrates, the latter reaction step is missing in KEGG overview; adding this step in the overview would result in a srd of 1. Hence, although combining individual metabolic maps could potentially be insightful for srd calculation, the aforementioned example advocates for prior careful manual curation of these maps before combining them to ensure their consistency with current biochemical knowledge.

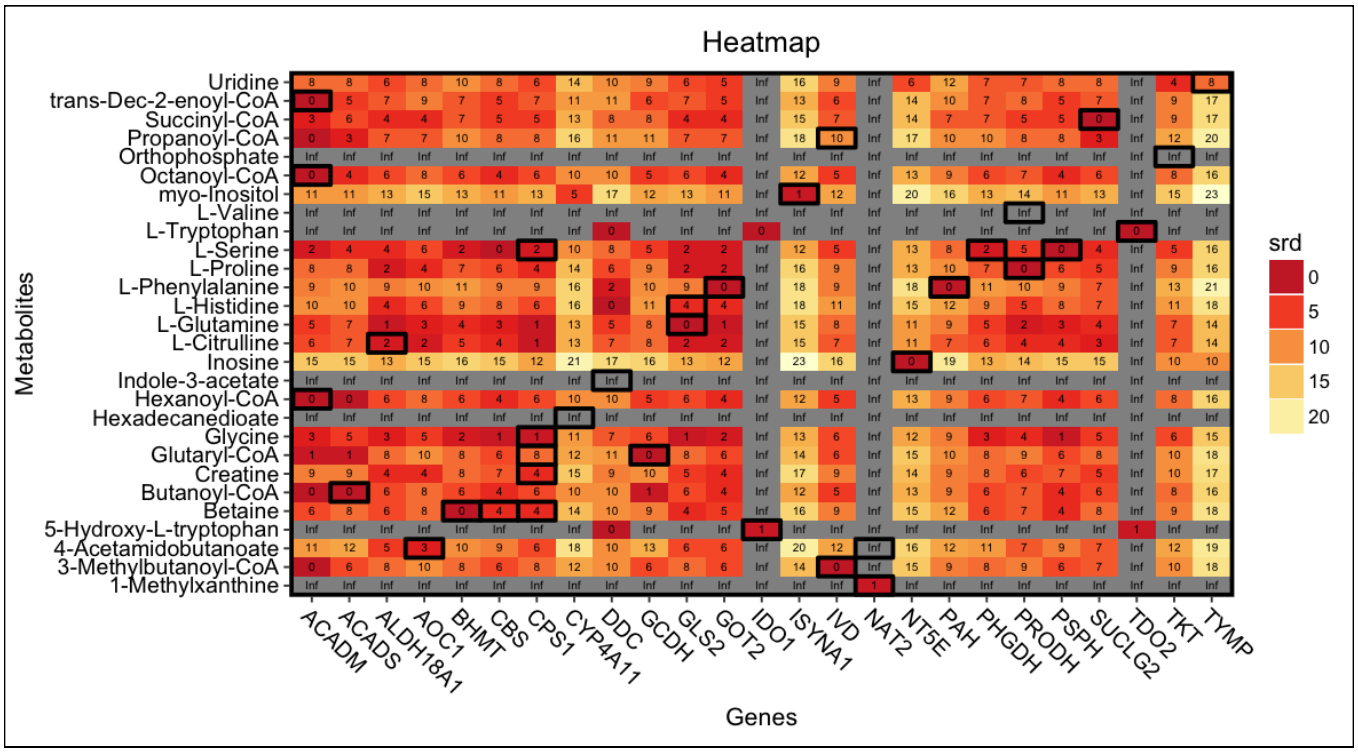

Figure 2.4 Heatmap of srd calculated for gene-metabolite pairs from Shin et al. using the overview.
Columns represent genes and rows represent metabolites. The computed distance is shown in each cell with the corresponding color code (from red – closest; to yellow - farthest). The srd calculated for the 35 gene-metabolite associations are reported in cells with a thick black border.

All the metabolic pathways of KEGG were used to extend further the coverage and knowledge since we were able to map 14 additional associations and compute their srd values (Table 2.1, Additional file 1: Table S4). It is noteworthy that for this, PathQuant uses only individual maps but do not compute srd values by combining 2 or more metabolic pathway maps (i.e. for a given gene and metabolite pairs that mapped on different metabolic pathways). Indeed, we considered that this would break the original topology of KEGG metabolic pathways or/and could possibly create erroneous paths. The importance of using all KEGG metabolic pathway maps individually to compute srd values is highlighted by the following example: the srd for the association TYMP (thymidine phosphorylase)-uridine is 8 in the overview and 1 in the pyrimidine metabolism map. In the overview, TYMP (EC 2.4.2.4) phosphorylates thymidine to thymine, which is 8 reactions away from uridine, while in the pyrimidine metabolism pathway, TYMP phosphorylates deoxyuridine to uracil which is one reaction away from uridine. While TYMP can phosphorylate both substrates, the latter reaction step is missing in KEGG overview; adding this step in the overview would result in a srd of 1. Hence, although combining individual metabolic maps could potentially be insightful for srd calculation, the aforementioned example advocates for prior careful manual curation of these maps before combining them to ensure their consistency with current biochemical knowledge.

Interestingly, 24 of the 49 associations reported by Shin et al. (Additional file 1: Table S6) for which we computed a finite srd value involved a gene implicated in inborn errors of metabolism. Of these 24 associations, 21 had a srd value ≤ 5; only 1 of them namely, PRODH(proline dehydrogenase 1)-L-valine (Table 2.2), had an Inf srd. This limitation of Inf values will be further discussed in the next section. Inherited metabolic diseases, which are

caused by mutations in a single gene resulting in the complete or partial loss of its enzyme function and alterations in related metabolites [151], have often been taken has examples in mGWAS to illustrate inherited variations in human metabolism [152] (see also for review ref. [91]).

Table 2.2 Manual annotation of gene-metabolite associations with Inf srd values using PathQuant

Associations	Manual computed srd	Missing enzyme
GT2-Phenyllactate	1	1.1.1.237 (humanCyc)
PRODH-L-Valine	15	6.3.2.26
KYAT1-Laminaribiose	4	1.1.1.110
FADS1-Docosapentaenoic acid	2	3.1.2.-
FADS1-Arachidonate	1	-*
FADS1-Adrenic acid	2	3.1.2.-
FADS1-Eicosapentaenoic acid	1	-*
FADS1-Icosatrienoic acid	1	3.1.2-
ASPG-asparagine	0	3.5.1.1

For each of the listed gene-metabolite association with an Inf srd value, only one enzyme was missing to enable computation of a finite srd. The table reports manually computed srd values and the identity of missing enzymes that enabled this computation. The symbol * indicates that while no enzyme was missing for this association, two unconnected subgraphs of a pathway was manually linked together to enable srd value computation (Additional file 3: Figure S2). (HumanCyc link: https://humancyc.org/HUMAN/NEW-IMAGE?type=EC-NUMBER&object=EC-1.1.1.237).

2.5.4 Srd metric analysis

The relevance of the computed srd values to explain associations were investigated using three different approaches.

(1) Randomization test on matched samples: The associated set contained 35 associations composed of 25 genes and 28 metabolites. This test was conducted using 1,000 resampled gene-metabolite sets of 35 associations. Each of the 25 genes was replaced by a randomly selected gene on the KEGG overview pathway map. The median value of the srd for each of the 1,000 resampled sets, depicted as a distribution plot in Figure 2.5, varied from 9 to 31, with 2.1% (i.e. 21 resampled sets) having infinite values, compared to the median value of the original associated set which was 1 (shown with a red line). This difference in median distribution between the associated *vs.* resampled sets is statistically significant (P-value $=9.9\times10^{-4}$ using formula (1); with a precision level ≤ 0.01 for 1,000 permutations). Altogether these results support the relevance of the srd metric for objective and quantitative annotation of associated sets of gene encoding enzymes and metabolites following their mapping to KEGG metabolic overview.

(2) Statistical association strength vs. srd value. Figure 2.6 depicts the close relationship between the strength of mapped associations and their computed srd values. As expected, the strongest associations (ex : NAT8 – myo-inositol; P - value $5.02E^{-230}$) have predominantly the smallest srd. There are, however, some strong associations with Inf srd value such as for the pair FADS1 (fatty acid desaturase 1)-Arachidonate. This issue is further discussed below, in the subsection "Manuel annotation".

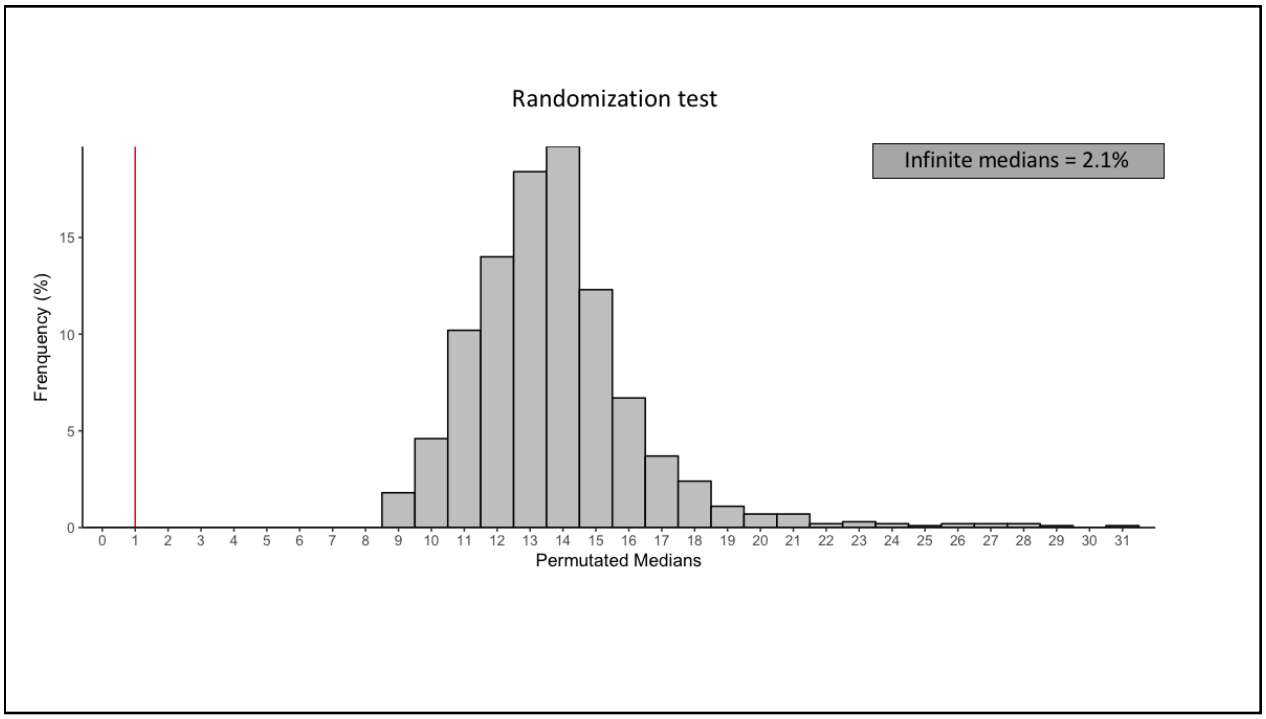

Figure 2.5 Distribution of median srd values for the 1000 randomization gene-metabolite sets.
Frequency bars in grey show the median srd values of the 1000 randomized sets. For comparison, the median value of the associated sets is shown as a red line. The percentage of infinite median srd values for the permutation tests is shown in the insert.

(2) Statistical association strength vs. srd value. Figure 2.6 depicts the close relationship between the strength of mapped associations and their computed srd values. As expected, the strongest associations (ex : NAT8 – myo-inositol; P - value $5.02E^{-230}$) have predominantly the smallest srd. There are, however, some strong associations with Inf srd value such as for the pair FADS1 (fatty acid desaturase 1)-Arachidonate. This issue is further discussed below, in the subsection "Manuel annotation".

(3) Manual annotation: While the computation of srd values using KEGG revealed a close biological relationship for most gene-metabolite associations, the investigation of associations with large and Inf srd values was essential to critically assess the limitations of the KEGG

database in explaining mGWAS gene-metabolite associations. Thus, we manually investigated the 11 associations with Inf values (Figure 2.3, Additional file 1: Table S4) and we were able to manually annotate 8 of them with a hypothetical numerical srd, of which 7 had a value <5 (Table 2) based on the following considerations: 1) Manually curated connection of disconnected subgraphs within a metabolic pathway, and 2) Using the multi-species design of the metabolic pathway offered by KEGG.

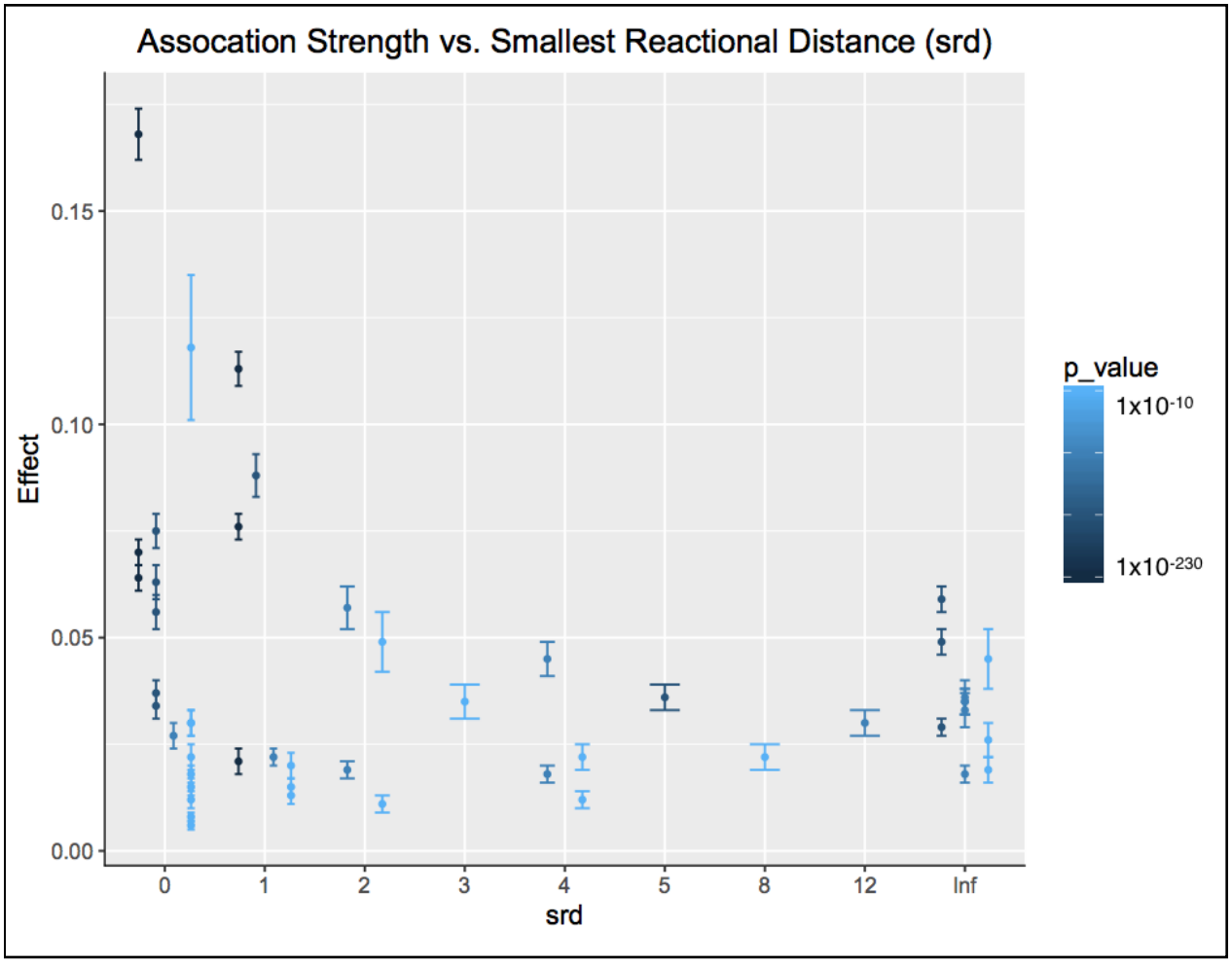

Figure 2.6 Scatter plot showing association strength *vs.* srd values for 49 mapped gene-metabolite associations.
The X axis represents srd values and the Y axis represents for effect ± SE and p-values for gene-metabolite associations reported by Shin et al. (cf. Supplementary Table 5 of Reference [72]) The magnitude of the P-values is proportional to the intensity of the blue color code (smallest: dark blue; to largest: light blue).

For example, one association that had an Inf srd value involved the unmapped gene, ASPG. This gene encodes for the enzyme asparaginase (E.C 3.5.1.1), which catalyzes the hydrolysis of L-asparagine to L-aspartate. Since this enzyme E.C 3.5.1.1 is known to be present in humans and is in the multi-species overview pathway of KEGG; its absence from the human KEGG overview map represents an annotation mistake. Manual annotation of this gene and its reported association to asparagine resulted in a srd value of zero.

Another example is the pair FADS1 (fatty acid desaturase 1)-Arachidonate, which is part of the "Biosynthesis of unsaturated fatty acids pathway". Its Inf srd value can be explained by the highly disconnected graph structure composing this metabolic pathway in particular (Additional file 3: Figure S2). This topological limitation is attributed to the incomplete nature of biological knowledge, which is inherent to pathway-based integration [70]. A manual connection of these subgraphs resulted in a srd value of 1, thereby supporting the strong relationship between the strength of associations between genes and metabolites and their computed srd value (Fig. 2.6).

Finally, although PathQuant was designed to compute srd values using KEGG metabolic pathway maps of a given species, namely human, multi-species pathway maps are available [106]. Using these maps, numerical srd values were computed for 6 metabolite gene-metabolite pairs with Inf srd values. This was explained by only one enzyme missing from KEGG human metabolic pathway maps. Among them is the enzyme 1.1.1.237, which was found to be annotated in the HumanCyc database [98] (Table 2).

Altogether, the aforementioned examples highlight some limitations of KEGG metabolic pathway maps and the utility of PathQuant in identifying them and providing a basis for their correction for future applications.

2.5.5 Comparaison with other methods

Our method not only compares favourably, but also adds to the currently available approaches for data interpretation of mGWAS [73], namely manual biological annotation, GGM [75] and mining the unknown [95] and for computation of shortest path length between genes and metabolites (used in MetaboSignal [138]) for the following reasons.

Biological annotation: We have compared the srd metric of gene-metabolite associations calculated by PathQuant to the manual biological description of these associations made by Shin et al. (reported in the Supplementary Table 6 of Reference [72]), referred as biological annotation. While this manual annotation provided a detailed description of the disease linked to the gene and metabolite of a given association, this is based on the depth of the literature review and, thus, important information could be missed. For example, no manual annotation was made for CPS1 (carbamoyl-phosphate synthase 1, mitochondrial) which was associated with L-serine, creatine and betaine. In contrast, PathQuant annotated srd values of 1, 3 and 4 respectively. Hence, while both manual biological annotation and our method enable similar interpretation of mGWAS data sets, the srd metric offers the advantage of a systematic and quantitative annotation of these associations using the complete and curated knowledge of the KEGG metabolic pathway maps.

GGM and mining the unknown: Both methods infers connectivity between metabolites from metabolomics data using an undirected network in which each edge represents a pairwise partial correlation between measured metabolites [75] leading to the reconstruction of metabolic pathways. GGMs only map measured metabolites, while "*Mining the unknown"* method aims at identifying unknown metabolites from mGWAS, using GGM and metabolic pathway databases

such as KEGG, based on the assumption that gene-metabolite associations reveal biological functions [95]. Although both methods use a quantitative proximity-based approach similar to PathQuant, GGM evaluates only distance between two metabolites and not for gene-metabolite pairs. In addition, GGM does not annotate individual gene-metabolite associations based on biological knowledge from pathway databases, but rather recovers metabolic pathways in a data-driven approach. Altogether these differences between these two methods and PathQuant make their direct comparison difficult.

Nevertheless, in their study, Shin et al. have connected metabolites using GGMs and linked genes to the network based on gene-metabolite association [72]. The difference between PathQuant and GGM can be best illustrated using the genes CPS1, CBS (cystathionine-beta-synthase) and BHMT (betaine--homocysteine S-methyltransferase), which were found to be associated with the metabolite betaine and for which PathQuant calculated a srd of 4, 4 and 0, respectively (Fig. 3). A srd of 0 enables one to identify that betaine is catalyzed by the protein encoded by BHMT. From the resulting GGM network, shown in the online supplement of reference [96], it is showed 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. GGM cannot differentiate between those 3 genes (i.e. showed as equally important). Interestingly similar to our results, GGM reported unconnected metabolite pairs for which the distance was infinite, which they interpreted as potentially representing novel pathways [75].

MetaboSignal: This is the first published method that reported a shortest path metric between genes and metabolites [138]. Similar to PathQuant, MetaboSignal is a R package based on the same mathematical criterion of shortest path metric and uses metabolic pathways from the

KEGG database. There are, however, important differences between these two methods. MetaboSignal offers the advantage of combining both metabolic and signaling pathway maps. Specifically, its network is built as a multipartite graph, i.e. the nodes of the graph represents multiple types of information, specifically: metabolites, metabolic genes (gene encoding enzymes), signaling genes (e.g. kinase) and non-enzymatic reactions (spontaneous reactions). While this can provide novel information about the interaction between genes and metabolites that goes beyond the known enzymatic reaction and pathways, it nevertheless modifies the original and manually curated topology of the pathways (signaling and metabolic) reported by KEGG. As it was done and also emphasized by Dumas *et al.* [21], a manual curation of the created pathways, supported by biological data, is crucial to ensure their validity.

In contrast, PathQuant focuses only on metabolic pathways, but its advantage is that it is knowledge-driven. PathQuant converts a metabolic pathway map into a graph of biochemical reactions (also called compound graph [122]) with metabolites as nodes and genes, mapped to their corresponding encoded enzymes. It computes shortest path values from any given list of gene-metabolite pairs using any given individual metabolic pathway maps or a specified list of individual metabolic pathways of KEGG without combining them, thus keeping the original and curated topology of the metabolic pathways reported by KEGG, which has been curated.

It is noteworthy that a direct quantitative comparison of the shortest path metrics calculated by PathQuant and MetaboSignal using as input data the list of gene-metabolite pairs of the mGWAS from Shin *et al.* could not be performed for two main reasons : 1) MetaboSignal computes its shortest path metric using a graph that is built from a selected list of pathways signaling and metabolic pathways, 2) the human overview pathway map is not accepted as input

by MetaboSignal which is the pathway map where most of the associations from *Shin et al.* are mapped.

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 provide a systematic, objective and quantitative annotation of mGWAS reported gene-metabolite associations using a shortest path metric (srd) computation, which is based on mapping of these associations on the original topology or the well curated metabolic pathway maps of KEGG database. In fact, all discussed methods, namely GGM [75], "Mining the unknown"[95] and MetaboSignal[138], provide information that is complementary to PathQuant. In principle, all these methods could be combined for a more comprehensive and complementary approach to mGWAS data analysis.

2.5.6 Future perspectives

Through our evaluation of the KEGG database to explain reported gene-metabolite associations by the mGWAS study of Shin et al. [72], we encountered some limitations, which are for the most part inherent to KEGG. In this method, we used individual KEGG pathway maps and kept their topology in order to find paths that have been previously validated biochemically. One additional advantage of KEGG is that it is the only database for which pathway maps were built with side compounds definition [153]. This information, previously found in KEGG RPAIRS, eliminates shortcuts created by side compounds such as ATP or H_2O when computing srd [154]. Despite these advantages, we recognized that the addition of other databases for pathway mapping would likely enhance coverage and extend the application and value of the srd metric using PathQuant. For this, we have considered the Recon2 database [87, 117], which is the most complete reconstruction of human metabolism so far. In contrast to KEGG, Recon2 has the

advantage of including more gene encoding transporters and include cellular compartments. However, Recon2 does not define side compounds in reactions. To the best of our knowledge, there is no algorithm available for the strict annotation of these compounds, although srd can be computed by avoiding most of them. As an example, a weighted algorithm using carbon transfer between substrates and products could be used to determine main reactions [126]. As each database catalogues reactions, pathways, metabolites and genes differently [119, 155], they all require different algorithms for srd computation and comparison of resulting paths. Thus, for all these reasons, the validation of the srd metric in other pathway databases, including those with signaling pathways, was beyond the scope of this study, but this is part of the planned future development of PathQuant.

Besides the aforementioned limitations of KEGG database, there are additional considerations that are linked to the use of PathQuant with any mGWAS dataset and that should be put forward as both constraints and future perspectives for applications of PathQuant. First, PathQuant uses gene-metabolite associations, which was provided by Shin et al. [72], while mGWAS report SNP-metabolite associations, where the SNP is annotated to its putative causal gene. Finding the affected gene by an intronic or intergenic SNP is still a complete problem of the genetics field. In the mGWAS by Shin et al. [72], gene annotation for SNPs was achieved using 1) the knowledge of 4 databases, namely KEGG, EHMN (Edinburgh human metabolic network reconstruction) [102], PubMed and BRENDA (The Comprehensive Enzyme Information System) [156] and 2) if no information was found in step 1, the closest gene within 500kb around the SNP was chosen. In this regard, PathQuant could be used to compute srd for all genes present within a defined SNP genomic region. We have tested this approach for the 49 associations annotated with

PathQuant and found that 28/49 of the annotated genes by Shin et al. were the only one enabling computation of a finite value of srd within the SNP region. In cases for which PathQuant calculated a srd value for multiple genes in the genomic region, the gene annotated by Shin et al. was always the one with the smallest srd value (Additional file 1: Table S5).

2.6 Conclusion

This study provides a critical assessment of the utility of the KEGG database for calculation of the shortest reactional distance (srd) through pathway mapping for explaining gene-metabolite association pairs revealed by mGWAS. Using mGWAS data from Shin et al. [72], we show that the developed R package, PathQuant, enables a robust, systematic and automated computation of the shortest reactional distance (srd) values between any list of pairs of genes (encoding for an enzyme) and metabolites (statistically associated or not) mapped onto KEGG metabolic pathway maps, while keeping their original, well curated, topology. The computed srd metric provides an objective and quantitative annotation of these gene-metabolite pairs based on current biochemical knowledge. As expected, short srd values (<5), which reflect close proximity between a gene-metabolite pair, were found for most statistically significant associated pairs. However, there were also significantly associated gene-metabolite pairs that had large or infinite srd values, some of which could be explained by manual curation of these data. This highlighted some limitations and annotation errors of KEGG metabolic pathway map, and the utility of PathQuant in identifying them and providing a basis for their correction for future applications. Our method adds to the currently available approaches for data interpretation of mGWAS [73], namely manual biological annotation, GGM [75] and mining the unknown [95] and MetaboSignal [138]. Additional potential applications of PathQuant include srd computation for all genes present

within a defined SNP genomic region that has been associated to a metabolite to identify those that with close proximity within selected KEGG pathway maps. Future development of PathQuant includes the addition of other databases such as Recon2 for pathway mapping, as well as signaling pathways given that some signaling proteins are known to directly impact on enzymes' activity. The latter inclusion, in future work, would enhance coverage and extend the application and value of the srd metric.

2.7 Supplementary figures

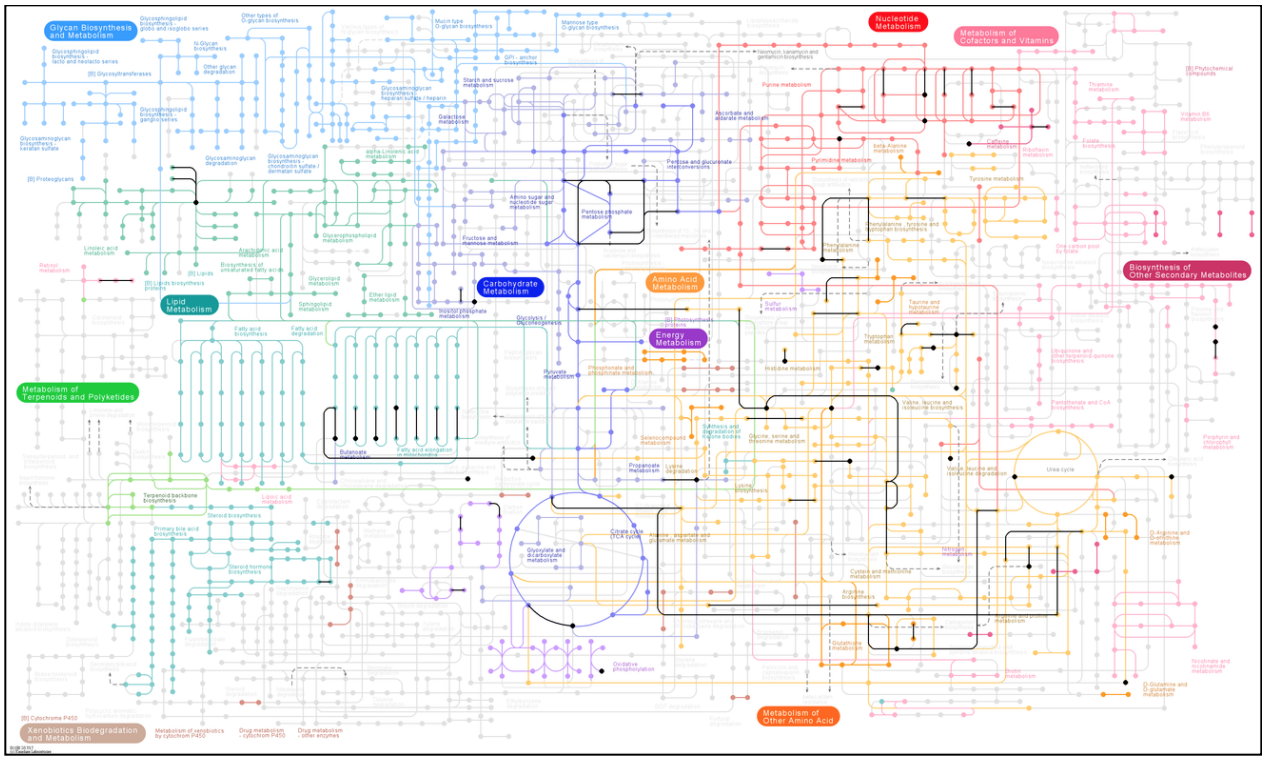

Figure 2.S1. The human overview pathway map of KEGG database, where nodes represent metabolites and edges represent genes and their encoded enzymes.
Grey colored nodes and edges refer to metabolites and genes/reactions not identified in human metabolism. Black highlighted nodes and edges represent metabolites and genes mapped from the Shin et al. dataset. This graph was generated using KEGG mapper [110]. (PNG 758 Ko)

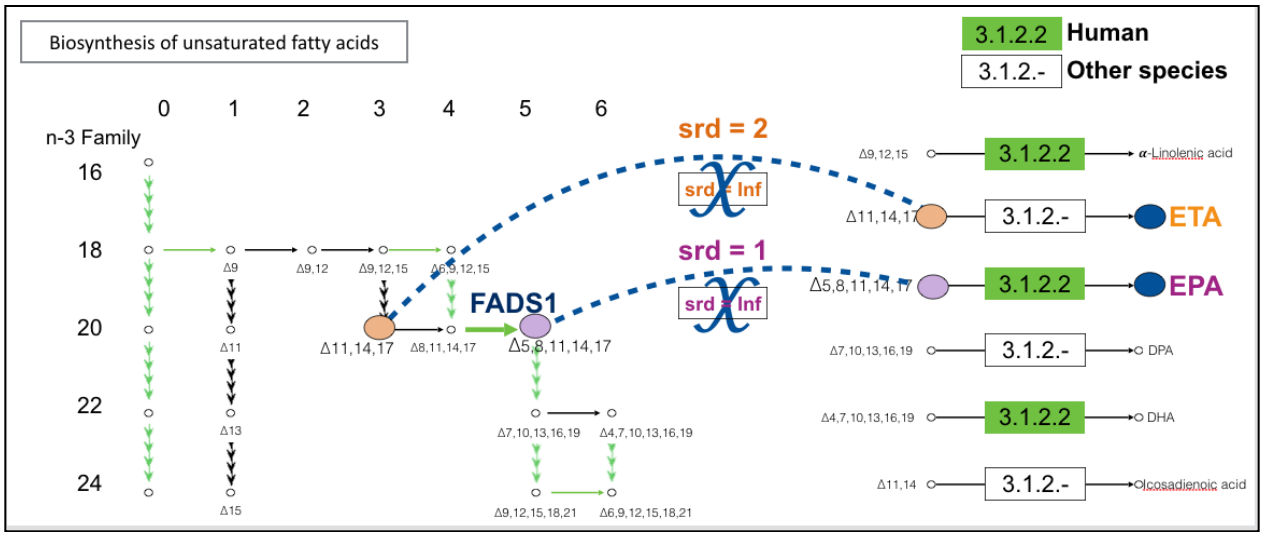

Figure 2.S2. Example of the topological limitation of the KEGG pathway map 'Biosynthesis of unsaturated fatty acids' and manual annotation of FADS1-Eicosatrienoic acid (ETA) and FADS1-Eicosapentaenoic acid (EPA) association pairs.

The metabolites are represented as nodes and genes mapped to their corresponding enzyme as edges. Metabolites that are common to the two different albeit unconnected maps are indicated as orange and purple nodes. Both the associations FADS1-Eicosatrienoic acid (ETA) and FADS1-Eicosapentaenoic acid (EPA) had an Inf srd annotated by PathQuant. Manual connection of enzymes catalyzing reactions involving these common metabolites allowed computation of srd(FADS1, ETA) of 2 and srd(FADS1, EPA) of 1. This graph was generated using KEGG mapper [110] and manually modified. (PNG 123 Ko).

2.8 Description of additional data files

Tables 2.S1-2.S6. Provides six supplemental tables as separate Excel spreadsheet, as well as an accompanying legend. Table S1 represent all genes, Table S2 all metabolites, Table S3 all associations reported by Shin et al. and identified in KEGG, Table S4 PathQuant output of Shin et al. dataset, Table S5 PathQuant output for every genes in 500 kb region of associations with an srd annotation and Table S6 srd annotation for associations annotated to an IEMS disease (XLSX 123 ko)

3. Discussion

L'objectif global de l'étude dans laquelle mon travail de maîtrise s'est inscrit était de développer un outil bio-informatique permettant de cibler un ensemble de métabolites donné en se basant sur la connaissance de l'architecture génétique du phénotype étudié. Mon premier objectif a été de construire une librairie R, *PathQuant,* dans le but de partager la méthode développée avec la communauté scientifique. Le second objectif de mon travail de maîtrise a été d'élargir la méthode développée en ajoutant à l'utilisation de la reconstruction du métabolisme (l'*overview*), l'ensemble des cartes de *KEGG* répertoriant des voies métaboliques. Finalement, mon dernier objectif a été d'accomplir une analyse approfondie des résultats obtenus par la méthode afin de mieux délimiter les forces et les limitations de la métrique *srd* pour l'annotation des paires entre gènes et métabolites.

Dans le contexte actuel de l'avancée des technologies "omiques", l'immense bassin de connaissances sur l'architecture génétique de centaines de phénotypes, généré principalement par l'explosion des études *GWAS* de la dernière décennie, font de ces connaissances un point de départ idéal pour le développement d'une méthode basée sur les connaissances. De plus, l'importance du développement d'approches combinant la génomique à d'autres "omiques" est mis en évidence par l'incapacité d'obtenir une compréhension précise des phénotypes au niveau cellulaire, moléculaire et physiologique avec la connaissance seule de leur architecture génétique [2, 39]. Ceci a mené la communauté scientifique à développer des approches systémiques favorisant la caractérisation et la combinaison des données provenant de plusieurs niveaux cellulaires. Dans ce travail, le métabolome a été le deuxième niveau cellulaire sélectionné, entre

autres, parce que le profil métabolique offre une représentation précise du phénotype observé par sa proximité avec ce dernier en plus d'offrir une compréhension de l'interaction qui existe entre le contrôle génétique et l'influence de l'environnement sur le phénotype. Ceci dit, contrairement à la génomique où le génome entier peut être séquencé et analysé relativement aisément, ce n'est pas le cas pour le métabolome. Ceci est expliqué par le fait que l'entièreté du métabolome ne peut être étudiée par un seul outil analytique. C'est pourquoi, il est primordial de développer des méthodes permettant la sélection de groupes ou de classes métaboliques à analyser pour le phénotype étudié.

Au meilleur de nos connaissances, il n'existe pas un outil idéal pouvant prédire un ensemble de métabolites pouvant être modulé par un phénotype par la seule connaissance de son architecture génétique ou par l'intermédiaire de tout autre type de données. C'est pourquoi, l'identification d'un profil métabolique à analyser en lien avec un phénotype est généralement faite par des spécialistes. Néanmoins, que ce soit un spécialiste ou une personne inexpériment, si cette tâche doit être accomplie manuellement, elle nécessite une profonde et fastidieuse recherche de la littérature scientifique. Considérant les raisons mentionnées ci-haut combinées à la récente explosion de la production des mégadonnées provenant de plusieurs niveaux cellulaires, le développement de solutions bio-informatiques permettant d'automatiser ce processus en se basant sur des méthodes systémiques qui prennent avantage des connaissances répertoriées dans les bases de données est crucial.

Le but initial du travail était de développer et d'utiliser une méthode systémique "multi-omique" basée sur les connaissances, combinant la génomique et la métabolomique, permettant de

comprendre les liens entre les gènes et les métabolites formant les associations rapportées par les récentes études *mGWAS;* une approche analytique rapportant des associations statistiques entre les *SNPs* (annotés à leur gène putatif) et les métabolites mesurés. L'hypothèse pour le développement de cette méthode était que les *SNPs,* annotés à leur gène putatif encodant des enzymes, auraient un impact sur les métabolites étant soit les substrats/produits des réactions catalysées par ces enzymes ou les métabolites étant à proximité dans les voies métaboliques. C'est pourquoi, notre méthode est basée sur le calcul du court chemin, *srd (shortest reactional path),* entre les gènes encodant des enzymes et leurs métabolites associés dans les voies métaboliques de la base de données *KEGG.* La validation de cette approche a été faite en utilisant le jeu de données de l'étude *mGWAS* réalisée par Shin *et al.* [72] qui était le plus complète et riche lors du début de ce projet [73]. Finalement, le manque d'outil bio-informatique a motivé le développement de la librairie R afin de partager notre méthode avec la communauté scientifique.

3.1 Considérations méthodologiques

3.1.1 Choix pour l'implémentation et la distribution de la méthode

Le langage R a été sélectionné pour le développement de cette méthode partagée avec la communauté scientifique sous le nom de PathQuant (*https://github.com/sandraTL/PathQuant*). Initialement, un code préliminaire avait été développé pour tester notre approche avec les données de Shin *et al.* [72] en se limitant à l'utilisation du réseau métabolique de l'*overview* de la base de données *KEGG*. Les raisons principales de l'utilisation du langage R pour le développement de la librairie *PathQuant* étaient: 1) la disponibilité d'un protocole simple d'utilisation afin de faciliter le bon développement d'une librairie R, mis en place par le projet *bioconductor* (*open source software for bioinformatics*) [157], 2) l'*IDE* (*Integrated Development*

Environment) *RStudio* dont le but est d'automatiser plusieurs fonctionnalités (ex: auto-complétion du code) afin d'accélérer et faciliter la programmation avec R, 3) la librairie *devtools* qui automatise la mise en place de fonction permettant le génération de la structure nécessaire à une librairie R, 4) les librairies *knitr* et *roxygen2* qui permettent de générer la documentation de la librairie de manière automatique, 5) la disponibilité des librairies *testthat* et *RUnit* pour la mise en place de tests unitaires pour assurer la qualité et le bon fonctionnement du code et 6) la popularité du langage R dans la communauté scientifique spécifiquement. D'autre part, Python est un langage également très utilisé par la communauté scientifique. Il est particulièrement indiqué lorsque le code doit s'intégrer dans une plateforme web ou si l'on veut développer des approches d'intelligences artificielles. Dans le cadre de notre projet, les deux langages auraient pu être utilisés, toutefois, au commencement du projet, Python n'offrait pas les outils équivalent à ceux offert par R pour le développement de librairie, c'est pourquoi, nous avons favorisé R.

Le gestionnaire de version Github [158] a été utilisé pour faciliter la gestion du développement de la librairie. Github est une plate-forme qui rend son contenu disponible gratuitement, ce qui permet de partager facilement le code développé avec la communauté scientifique, ce qui est essentiel pour une avancée rapide de nos connaissances et de la recherche.

3.2 Analyse critique de la méthode

La méthode présentée à la section 2.4 a été appliquée à un jeu de données provenant d'une étude *mGWAS,* puisque la construction des réseaux métaboliques est directement basée sur les données du génome et du métabolome. Le but de cette section est de mettre en lumière les avantages et les limitations des caractéristiques de notre méthode qui ont un impact important sur l'interprétation des résultats.

Toutefois, il faut noter que la méthode peut être appliquée à d'autres types d'études rapportant des associations entre gènes et métabolites qui pourraient être annotées par cet outil, telles que, les études combinant la transcriptomique à la métabolomique [65] ou l'épigénétique à la métabolomique [66]. Spécifiquement, la méthode peut annoter les associations statistiques qui sont réalisées entre les niveaux d'ARNm (annotés à leur gène putatif) et les métabolites mesurés et elle peut s'appliquer aux associations statistiques formées entre les sites de méthylation de l'ADN (annotés au gène dont il contrôle possiblement l'expression) et les métabolites mesurés.

3.2.1 Application de la méthode aux données *mGWAS*

Le principal objectif d'une étude utilisant une approche systémique combinant l'analyse de plusieurs niveaux cellulaires est d'approfondir et de préciser notre compréhension d'un phénotype donné. Le potentiel de nouvelles découvertes pouvant émerger de ces approches est immense et dans le cas spécifique des *mGWAS*, les résultats obtenus jusqu'à présent ont permis une meilleure compréhension du contrôle du génome sur le métabolisme. L'annotation des associations rapportées par les *mGWAS* par notre méthode pourrait permettre d'accélérer le processus d'analyse des liens biologiques entre les gènes et les métabolites.

L'étude de Shin *et al.* [72] a été réalisée avec les échantillons de plasma d'un total de 7 824 individus provenant de deux cohortes européennes: la *German KORA* (*Kooperative Gesundheitsforschung in der Region Augsburg*) et la *TWIN UK*. Un *GWAS* utilisant des puces de génotypage ayant 2,1 millions de *SNPs* a été réalisé et ces *SNPs* ont été imputés à l'aide des données provenant du projet *HapMap* [26]. D'autre part, les analyses de métabolomique ciblées et non-ciblées ont été réalisées par l'intermédiaire de plusieurs outils analytiques de spectrométrie de masse, notamment GC-MS et LC-MS. Les analyses de métabolomique ont permis de mesurer

un total de 529 métabolites dont 333 ont été identifiés (structure chimique connue) et 196 non identifiés (structure chimique inconnue). Au total, 299 associations statistiquement significatives impliquant 145 *SNPs* annotés à 132 gènes uniques et 187 métabolites uniques (136 identifiés et 51 non identifiés) ont été rapportées (Table 2.1). Parmi ces 299 associations, 52 étaient entre un *SNP* et un ratio métabolique. Ces derniers ont été séparés en deux associations après élimination des doublons ce qui malgré tout permet de totaliser un même nombre d'associations (299). Souvent, un métabolite faisant partie d'un ratio était également impliqué dans une association simple impliquant le même *SNP*. De plus, dans le cadre de notre méthode, les associations entre un *SNP* et un métabolite non identifié ont été éliminées du jeu de données puisqu'il n'était pas possible de les cartographier.

Construction du jeu de données

À ce jour, les normes pour la publication des résultats d'études de métabolomique ne sont pas précisément établies, ce qui fait en sorte que les métabolites sont majoritairement rapportés par leur nom commun. Puisque chaque nom commun a souvent plusieurs synonymes, la conversion de ce dernier en un identifiant adapté à la base de données visée est complexe. Toutefois, l'augmentation du nombre d'études de métabolomique a favorisé l'implantation de normes par le consortium *COSMOS* (*Coordinatino of Standards in MetabolomicS*) dans le but d'uniformiser la manière dont les résultats sont rapportés. Les identifiants suggérés sont ceux étant le plus précis pour identifier les métabolites tel que répertorié dans : *SMILES (Simplified Molecular-Input Line-Entry System*) et *InChI (IUPAC International Chemical Identifier*) car ils se basent sur la structure chimique des métabolites.

Toutefois, les normes proposées par *COSMOS* ne sont pas nécessairement appliquées et c'est ce qui complique l'utilisation de jeux de données publiés. Les données publiées par Shin *et al.* ne suivaient pas ces recommandations, dans ce jeu de données les métabolites ont été identifiés par leur nom commun et par leur identifiant provenant d'une base de données privée créée par la compagnie *Metabolon Inc (Durham, NC, USA)*. Puisque la base de données *Metabolon* n'est pas librement accessible, les noms communs des métabolites ont été manuellement convertis en leur identifiant *KEGG* et *HMDB* [49]. Pour un jeu de données ayant plusieurs centaines de métabolites, ce processus est long et fastidieux ce qui renforce la nécessité d'uniformiser les normes de publication des métabolites. La base de données *HMDB* a majoritairement été utilisée pour trouver l'identifiant *KEGG* puisqu'elle répertorie une liste des synonymes utilisés pour chaque métabolite ainsi que leurs identifiants pour plusieurs bases de données dont *SMILES, InChi et KEGG.*

Contrairement à la conversion des identifiants pour les métabolites, la conversion des identifiants pour les gènes est généralement plus simple car ceux-ci sont rapportés en suivant les normes établis par *HUGO (HUGO GENE Nomenclature Committee)* [159]. Les gènes rapportés par Shin *et al.* étaient identifiés par cet identifiant ainsi la conversion a été réalisé rapidement avec l'outil de *KEGG, LinkDB* [149].

3.2.2 Développement de la méthode

Dans le cadre de notre projet, l'utilisation d'une reconstruction du métabolisme basée sur les voies réactionnelles métaboliques de la cellule regroupant l'information génétique (annotation des gènes encodant les enzymes qui catalysent les réactions métaboliques) ainsi que l'information

métabolique (substrats/produits des réactions métaboliques catalysées par les enzymes encodées par les gènes) représentait la source de données la mieux adaptée pour entreprendre ce défi.

3.2.2.1 Choix de la base de données

Comme mentionné à la section 1.4, plusieurs bases de données répertorient les voies métaboliques ainsi que des reconstructions du métabolisme. Au moment du développement de cette méthode (2012), la base de données *KEGG* [110] a été sélectionnée. Ceci pour plusieurs raisons, notamment, pour la fiabilité de son contenu, pour son *API* [146] qui permet un accès simple et efficace pour l'extraction de son contenu, pour son outil de visualisation *KEGGMapper* qui facilite l'analyse des résultats ainsi que pour sa renommée dans la communauté scientifique.

L'*overview* de *KEGG* représentait un excellent point de départ, notamment parce qu'elle contient un grand nombre de voies métaboliques intermédiaires telles que le cycle de Krebs et la glycolyse. Toutefois, en plus de l'*overview*, *KEGG* contient plusieurs autres cartes qui répertorient des voies métaboliques de manière individuelle dont quelques-unes qui ne sont pas incluses dans l'*overview* par exemple "*Bionsynthesis of unsaturated fat acids - Reference pathway*" et dont certaines qui sont seulement partiellement représentées sur l'*overview* par exemple: *"Sulfur metabolism"*. C'est pourquoi, dans la continuité de ce travail, l'ensemble des cartes ont été utilisées afin d'approfondir l'analyse de la pertinence de la métrique *srd* pour l'annotation des associations gène-métabolite.

Les prochaines sous-sections ont pour but de mettre en évidence les caractéristiques de la base de données *KEGG* pour cartographier les gènes, les métabolites ainsi que les associations rapportées par les études *mGWAS* particulièrement pour les données de Shin *et al*.

Classification des associations

Les *SNPs* impliqués dans les associations rapportées par les *mGWAS* sont annotés de manière putative à des gènes destinés à diverses fonctions cellulaires (enzymes, transports, facteurs de transcription, etc). Toutefois, les voies métaboliques de *KEGG* sont majoritairement construites à partir des réactions catalysées par les enzymes, à l'exception de quelques transporteurs tels que: *SLC33A1* (*solute carrier family 33 member 1*) et *SLC27A5* (*solute carrier family 27 member 5*). C'est pourquoi, le développement de cette méthode a nécessité une classification des associations utilisées dans le but de l'appliquer seulement à celles étant pertinentes, c'est-à-dire, les associations dont le *SNP* est annoté de manière putative à un gène qui encode une enzyme. Ainsi, les 132 gènes impliqués dans les associations rapportées par Shin *et al.* ont été classés selon trois catégories, nommément : 1) les enzymes (86), 2) les transporteurs (20) et 3) les autres (33) (Table 2.1). Au total, 86 associations ont été utilisées parce qu'elles impliquaient un des 86 gènes encodant une enzyme et un métabolite identifié. Ces 86 associations étaient formées par 50 gènes uniques et 66 métabolites uniques (Table 2.1). Toutefois, l'utilisation d'une reconstruction du métabolisme telle que *Recon2*, qui différencie les compartiments cellulaires, aurait permis une augmentation du nombre d'associations utilisées pour le calcul du *srd* puisque cette reconstruction contient les réactions de transport entre les différents compartiments cellulaires (discutés à la section 3.3). Les prochaines sections prendront seulement les 86 associations mentionnées ci-haut lors de l'analyse.

Cartographie des gènes de Shin et al.

Comme mentionné dans la section précédente, 50 gènes uniques étaient impliqués dans les 86 associations d'intérêts et ils étaient tous catalogués dans *KEGG*. Toutefois, 46 (92%) des 50

gènes présents dans les cartes répertoriaient les voies métaboliques. Plus exactement, quatre gènes : *ASPG* (asparaginase), *ETFDH* (*electron-transferring-flavoprotein dehydrogenase*), *MARCH8* (*E3 ubiquitin-protein ligase MARCH1/8*) and *PPM1K* (*protein phosphatase 1K*) n'étaient présents sur aucune carte de *KEGG*. Ainsi, aucune valeur *srd* n'a pu être calculée pour les associations impliquant ces gènes, ils ont été annotés par *NA* (*not applicable*).

L'attribution par Shin *et al.* d'un gène putatif aux *SNPs* faisant partie des associations rapportées a été basée sur les connaissances des bases de données *KEGG* et *BRENDA*. Si aucun gène putatif n'était identifié, alors le gène le plus près physiquement du *SNP* dans la région de 500 kb autour de ce dernier a été sélectionné.

Cartographie des métabolites de Shin et al.

En ce qui concerne les métabolites, 54 (82%) des 66 ont été cartographié dans une carte répertoriée par *KEGG*. Considérant les limitations méthodologiques et technologiques, mentionnées à la section 1.1.2, pour l'étude du métabolome dans son entièreté, il est attendu que le pourcentage de métabolites cartographiés soit moindre que celui des gènes car les voies métaboliques sont fragmentaires. Il faut également noter que 7 de ces 12 métabolites non-cartographiés sont des lipides et que cette classe de métabolites a été moins bien investigué par le passé dû aux limitations technologiques. Les associations comportant ces 12 métabolites ont été annotées par *NA*.

Cartographie des associations de Shin et al.

Au total, une valeur *srd* a été calculée pour 49 des 86 associations considérées, dont 35 ont été calculée en utilisant l'*overview* de *KEGG* uniquement. Le calcul d'une valeur *srd* n'a pas pu être

réalisé pour 37 de ces 86 associations, notamment pour les raisons suivantes : 1) les associations dont le gène ou le métabolite n'était pas cartographié sur une carte de *KEGG* (discuté aux sections 3.2.2.1.2 et 3.2.2.1.3) et 2) les associations dont le gène et le métabolite sont cartographiés sur des cartes différentes. Lorsque ceci est le cas, il n'est pas possible de calculer une valeur de *srd*. Une solution à cette limitation serait de relier l'ensemble des cartes pour former une carte globale. D'un point de vue algorithmique, ceci se fait aisément, toutefois, la vérification manuelle de l'ensemble des nouvelles voies métaboliques formées par cette approche est nécessaire et essentielle pour assurer la pertinence du contenu biologique d'une telle carte. Ce processus de vérification demande un effort considérable, c'est pourquoi, cette solution n'a pas été appliquée à cette étape-ci du développement de la méthode. Toutefois, cela représenterait une avancée potentiellement intéressante pour le projet.

3.2.3 Analyse des valeurs *srd* obtenues pour le jeu de donnée *mGWAS*

L'application du calcul du court chemin (*srd*) dans les voies métaboliques de *KEGG* peut mener à plusieurs résultats qui ont été séparés en quatre catégories: 1) $0 < srd \leq 5$, 2) $srd > 5$, 3) infini et 4) *NA*. La définition des deux premières catégories a été faite en fonction des propriétés topologiques du réseau métabolique de l'*overview* puisque la majorité des associations ont été cartographiées sur cette carte et parce qu'elle représente la majorité de ce qui est connu du réseau métabolique de l'humain. Entre autres, le court chemin moyen de l'*overview* est 15 tandis que le plus long *srd* de l'*overview* (diamètre) est de 46, ainsi on estime qu'une valeur de *srd* inférieur ou égale à cinq est significativement plus courte que ce qui est obtenu moyennement dans le réseau. De plus, une étude récente étudiant le réseau métabolique d'*E. coli* suggère que la suppression de l'expression d'un gène aurait un impact sur les métabolites se trouvant jusqu'à une distance de

cinq réactions de l'enzyme encodée par ce gène [137]. Ils précisent que l'impact est majoritairement plus important lorsque le métabolites se trouve à une distance inférieure ou égale à deux réactions.

Les associations appartenant aux trois premières catégories ont été cartographiées sur les voies métaboliques de *KEGG* avec l'outil *KEGGMapper* et les chemins ont été analysés manuellement. L'annotation d'un *NA* pour les associations appartenant à la quatrième catégorie a été expliquée en détail dans les sections 3.2.2.1.1 à 3.2.2.1.3.

3.2.3.1 $0 < srd \leq 5$

L'annotation d'une valeur *srd* inférieure à 5 (catégorie 1) suggère que l'expression du gène a une influence pertinente sur la concentration du métabolite auquel il est associé. Cette affirmation a été soutenue par les vérifications manuelles qui ont été réalisées mais aussi par le fait que plusieurs des associations ayant un petit *srd* impliquent des gènes causaux de plusieurs maladies innées du métabolisme (*IEMS*) et leurs métabolites associés sont déjà identifiés comme bio-marqueurs métaboliques de ces maladies (Table 2.S6). Par exemple, le jeu de donnée de Shin *et al.* contenait l'association entre le gène *IVD - Isovaleric acid dehydrogenase* et le métabolite 3-méthylbutanoyl-CoA, comme illustré à la figure 3.1, ce métabolite est le substrat utilisé par l'enzyme encodée par le gène IVD. La perte de l'expression du gène IVD cause la maladie innée du métabolisme nommée acidémie isovalérique qui est également caractérisée par une accumulation du composé 3-méthylbutyrylcarnitine dans le plasma. Toutefois, puisque les composés carnitines ne sont généralement pas retrouvés sur la voie du métabolisme de *KEGG,* le

calcul entre le gène et l'équivalent en acyl-CoA du composé est fait, comme expliqué à la section 2.5.1.

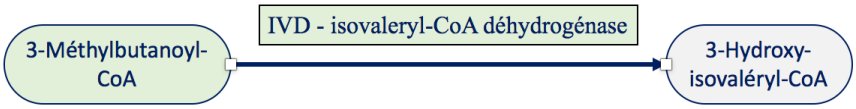

Figure 3.1 Réaction catalysée par l'enzyme isovaléryl-CoA déhydrogénase encodée par le gène IVD.
Les composés en vert représente l'association rapportés par Shin *et al.*

3.2.3.2 *srd* > 5

Lorsque la valeur de *srd* est plus élevée que cinq (catégorie 2), une influence relativement directe de l'expression du gène sur le métabolite associé est moins évidente. Ainsi, une analyse manuelle pour déterminer si le chemin utilisé pour le calcul du *srd* est biologiquement valide ou non est nécessaire. Les cas pour lesquels le chemin n'est pas valide, on peut supposer que le chemin liant le gène et le métabolite n'est pas répertorié dans la base de données utilisées ou on peut supposer soit que l'associations est fausse. Tandis que les cas pour lesquels le chemin est valide, les valeurs *srd* étant plus longue que la moyenne permettent d'identifier les voies métaboliques où il possible de retrouver de telles valeurs. Il serait éventuellement possible de fragmenter la recherche des métabolites selon le profil génétique en fonction des voies métaboliques ou se retrouve les gènes données.

3.2.3.3 Infini

Un total de 11 associations ont été annotées avec une valeur *srd* infinie. L'annotation d'une valeur *srd* infinie indique qu'il n'existe pas de chemin entre le gène et le métabolite d'une association donnée dans la carte où ils ont été cartographiés. Ceci peut vouloir dire qu'il n'existe pas de liens

biologique entre le gène et le métabolite ou que la base de données est incomplète et que la voie réactionnelle reliant ce gène et ce métabolite n'a pas encore été annotée. C'est pourquoi, une analyse manuelle des 11 associations a été réalisée et a permis d'obtenir des informations pertinentes pour la génération d'hypothèses ainsi que pour l'évaluation du contenu de la base de données utilisée.

3.2.3.3.1 Génération d'hypothèses

Pour analyser les 11 associations visées, nous avons mis à profit une particularité de *KEGG* par l'utilisation des cartes globales qui contiennent les réactions annotées à l'ensemble des espèces répertoriées. Prenons par exemple l'*overview*, illustré à la Figure 2.S1, les sommets (métabolites) et les arêtes (gènes/réactions) colorés représentent ce qui est annoté à l'humain. Tandis que les sommets et les arêtes en gris représentent ce qui est annoté aux autres espèces excluant l'humain. Ceci dit, il est possible d'utiliser les cartes complètes et c'est ce qui a été fait pour analyser les associations pour lesquelles des valeurs infinies ont été obtenues. Pour 7 des 11 associations, nous avons remarqué qu'une seule enzyme était absente chez l'humain pour permettre le calcul du *srd* (Table 2.2). Une de ces enzymes, l'hydroxyphénylpyruvate réductase (e.c 1.1.1.237), a d'ailleurs été annotée chez l'humain dans la base de données *HumanCyc*. Ces enzymes représentent des hypothèses qui pourraient permettre de raffiner le contenu de la carte et pour comprendre l'association statistique donnée.

3.2.3.3.2 Évaluation du contenu de *KEGG*

L'analyse des associations avec des valeurs infinies a aussi permis d'identifier des problèmes dans la topologie de la voie métabolique suivante: "*Bionsynthesis of unsaturated fat acids - Reference pathway*". L'exploration du lipidome est assez récente, donc le contenu des voies

réactionnelles lipidiques est moins bien détaillé comparativement aux voies métaboliques très connues (ex: cycle de Krebs, glycolyse), ce qui engendre des problèmes topologiques. Entre autres, la voie métabolique mentionnée ci-haut contient plusieurs sous-graphes non connectés qui, en réalité, devrait être connectés puisqu'ils partagent des métabolites en commun. Ceci dit, deux associations avec une valeur *srd* infinie ont été cartographiées sur cette voie métabolique et il ne manquait aucune enzyme pour convertir leur valeur infinie en une valeur finie appartenant à la première catégorie, par exemple l'association FADS1-EPA(acide icosapentaénoïque) (Figure 2.S2, Table 2.2).

3.2.4 Les avantages de notre méthode en comparaison avec les autres méthodes

À ce jour, il existe des outils permettant d'analyser les données "omiques", par exemple, les méthodes d'enrichissement des gènes (*GSEA*) [82] ou des métabolites (*MSEA*) [81] pour identifier les voies métaboliques pouvant être fortement impliquées dans un phénotype. Ces méthodes permettent une évaluation individuelle de chaque type de donnée "omiques", elles ne sont donc pas comparables à *PathQuant* dont le but est d'analyser les données combinant la génomique et la métabolomique. Lorsque le développement de la méthode a été débuté, au meilleur de nos connaissances, il n'existait pas de méthode similaire à *PathQuant*, toutefois une nouvelle méthode nommée *MetaboSignal* [138] a récemment été publiée et elle pourrait être comparée à *PathQuant* parce qu'elle fait un calcul du court chemin entre un gène et un métabolite en utilisant les cartes de la base de données *KEGG.* Dans les sous-sections suivantes, *MetaboSignal* ainsi que les méthodes actuellement utilisées pour analyser les données *mGWAS* seront comparées à *PathQuant.*

PathQuant vs. MetaboSignal

La différence principale entre *PathQuant* et *MetaboSingal* est que cette dernière s'applique exclusivement sur un réseau combinant au minimum deux cartes de *KEGG* dont une doit être une voie métabolique et l'autre doit être une voie de signalisation, d'autres cartes pouvant être ajoutées. La combinaison des cartes sélectionnées se fait automatiquement par *MetaboSignal* en joignant les arêtes (réactions) ayant des métabolites en communs. Puis le calcul du court chemin est réalisé. Toutefois, comme discuté à la section "*Cartographie des associations de Shin et al.",* la connexion de plusieurs cartes change leur topologie et peut créer des voies réactionnelles erronées ce qui nécessite une vérification minutieuse du réseau créé par ce processus, ce qui est ardu et demande un travail considérable. Contrairement à *MetaboSignal, PathQuant* calcule le court chemin entre un gène et un métabolite sur chaque carte répertoriant des voies métaboliques séparément, ce qui conserve la topologie des voies métaboliques répertoriées par *KEGG*. Actuellement, *PathQuant,* n'utilise pas les voies de signalisation de *KEGG,* mais l'utilisation de ces cartes ajouterait une profondeur à la méthode et permettrait d'augmenter le nombre d'association annotées par la méthode.

Comparaison avec les méthodes utilisées pour analyser les données mGWAS

Deux stratégies ont été utilisées jusqu'à maintenant pour analyser les résultats des *mGWAS* [94]: 1) *GGM* et 2) *mining the unknown.* Ces deux méthodes ont été comparées à PathQuant à la section 2.5.5. En effet, *GGM* et *mining the unknown* sont des approches basées sur les données qui utilisent des corrélations statistiques pour reconstruire les réseaux métaboliques. Elles sont très avantageuses pour visualiser les différentes interactions statistiques entre les métabolites et les gènes rapportés par les associations. Toutefois, ces deux méthodes se comparent difficilement

à *PathQuant* puisqu'elles ne sont pas basées sur les connaissances mais plutôt sur les données. *PathQuant* est basée sur la cartographie des associations sur les réseaux métaboliques connus qui contiennent l'ensemble des gènes et des métabolites impliqués dans les réseaux et non le sous-ensemble présent dans les associations rapportées. La limitation la plus importante de notre méthode pour l'analyse d'un jeu de données *mGWAS* est qu'elle est restreinte par le contenu de la base de données *KEGG* et donc l'ensemble des associations n'est pas nécessairement entièrement analysé. Les approches *GGM* et *mining the unknown* peuvent donc être complémentaires à *PathQuant* puisque leurs objectifs sont différents.

3.3 Application préliminaire du calcul du *srd* avec *Recon2*

Lors d'un stage de deux semaines à Toulouse en France dans le laboratoire de Dr Fabien Jourdan à l'INRA (Institut National de Recherche Agronomique) qui se spécialise dans les réseaux métaboliques, j'ai eu la chance de tester la méthode du calcul du *srd* sur la base de données *Recon2* avec l'aide précieuse de Clément Frainay, candidat au PhD. Notre objectif était d'obtenir des résultats préliminaires en appliquant notre méthode à *Recon2* pour comparer principalement la cartographie des associations entre les deux réseaux. Pour faciliter ce travail quelques modifications topologiques au réseau métabolique de *Recon2* ont été réalisées. Dans les prochaines sous-sections, la construction du réseau modifié de *Recon2*, l'algorithme utilisé pour calculer le *srd* ainsi qu'une comparaison des résultats obtenus dans les deux bases de données seront abordés.

3.3.1 Construction du réseau de la base de données *Recon2*

Il y a quatre différences majeures entre *Recon2* et *KEGG* qui déterminent les algorithmes utilisés ainsi que l'interprétation des résultats : 1) *Recon2* annote les réactions et les

métabolites à leurs compartiments cellulaires respectifs, 2) elle annote les réactions de transport des métabolites entre les différents compartiments, 3) la définition des réactions inclut les cofacteurs et 4) l'annotation du sens des réactions enzymatiques. Pour obtenir des résultats préliminaires sur la base de données *Recon2* et pour permettre une comparaison des résultats obtenus entre cette dernière et *KEGG*, le réseau de *Recon2* a été modifié de la façon suivante pour obtenir deux réseaux similaires sans diminuer la cartographie des associations. Premièrement, l'ensemble des réactions et des métabolites a été annoté comme se produisant dans le cytosol et les doublons ont été éliminés. Puis, l'ensemble des réactions de transport des métabolites entre les différents compartiments a été enlevé pour faciliter dans un premier temps la comparaison des deux réseaux sur la base d'un référentiel similaire (voies métaboliques). Ces deux étapes ont permis de supprimer la différenciation entre les compartiments cellulaires, toutefois un développement futur de PathQuant serait de prendre avantage de la présence des transporteurs et des compartiments et adapter l'algorithme du calcul du *srd* en conséquence. Pour les cofacteurs, il n'est pas possible de simplement les retirer du réseau puisqu'ils ne sont pas annotés comme étant des cofacteurs, toutefois deux cofacteurs ubiquitaires connus ont été enlevés, l'H_2O et le CO_2. Finalement, pour cette première utilisation de *Recon2*, l'annotation du sens des réactions a simplement été supprimé pour deux raisons: 1) pour permettre la comparaison avec l'*overview* de *KEGG* dont les réactions n'ont pas de sens et 2) parce que l'exactitude de l'annotation du sens des réactions n'est pas complètement fiable, par exemple, les réactions pour lesquelles le sens n'est pas connu sont annotées comme étant bidirectionnelles [160].

3.3.2 Calcul du *srd*

Le calcul des valeurs *srd* a été fait en utilisant l'algorithme du court chemin modifié en ajoutant des poids sur les arêtes et en ignorant le sens des réactions annoté par *Recon2*. Les poids sur les arêtes ont été ajoutés pour permettre d'éviter d'utiliser les chemins passant par des cofacteurs. Les poids sont définis dans l'intervalle entre 0 et 1 et sont attribués selon la conservation du nombre de carbones entre le substrat et le produit de la réaction. Ainsi, un poids de 1 est attribué à une arête pour laquelle le substrat et le produit ont le même nombre de carbones, un poids de 0 est attribué si au moins un des deux métabolites n'a aucun atome de carbone et un poids supérieur à 0 et inférieur à 1 est attribué proportionnellement à la différence de carbones entre le substrat et le produit [126].

3.3.3 Comparaison préliminaire des résultats obtenus

Au total, 49 associations ont été cartographiées dans le réseau métabolique de la base de données *Recon2* et donc 49 valeurs *srd* ont été calculées et leur distribution est illustré à la figure 3.2 A. La distribution des valeurs *srd* obtenues avec *KEGG* pour le même ensemble est représenté à la figure 3.2 B. Entre autres, 31 associations étaient également retrouvées avec *KEGG* dont 24 avait une valeur *srd* identique, tandis que les 7 autres associations avaient des valeurs *srd* différentes qui variaient d'au plus quatre réactions entre *KEGG* et *Recon2* (*srd Recon2* = *srd* de *KEGG* ± 4). De plus, la valeur *srd* médiane des résultats trouvés avec *Recon2* était de 1, ce qui est identique à la médiane des résultats trouvés avec *KEGG*. Au total, 17 associations ne pouvant pas être annotées par *KEGG* mais annotés en utilisant *Recon2* ont été obtenues, 11 des associations implique un lipide. Ces résultats préliminaires démontrent la capacité de la base de données Recon2 pour augmenter la couverture d'annotation des associations, particulièrement les

associations impliquant des lipides. Toutefois, une analyse approfondie des résultats permettrait de mieux comprendre les différences trouvées entre les deux bases de données et permettrait de mieux délimiter les forces et les limitations de l'utilisation de Recon2. Entre autres, une comparaison des chemins empruntés pour le calcul du *srd* dans chaque base de données serait également intéressante.

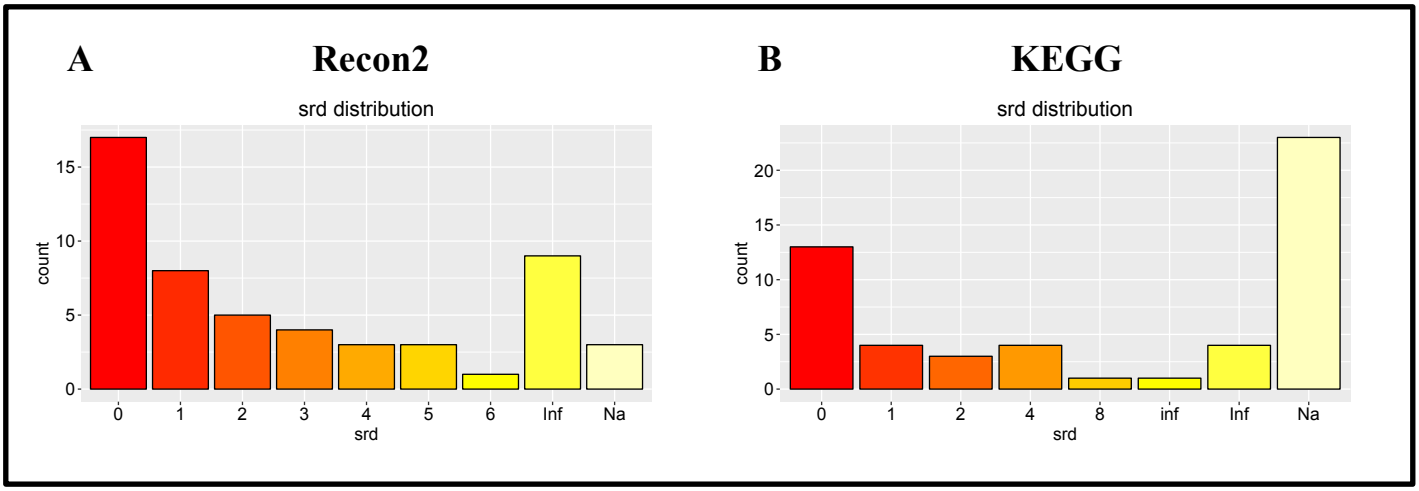

Figure 3.2 Distribution des valeurs srd obtenues avec *Recon2* et *KEGG*.

A. Distribution des valeurs srd obtenues en les calculant avec la base de données Recon2. **B.** Distribution des valeurs srd en obtenues avec *KEGG* pour le même ensemble d'associations cartographié dans la base de données Recon2.

3.4 Développements futurs

3.4.1 Augmenter la cartographie des associations

Un développement à court terme de la méthode serait d'ajouter l'utilisation de la base de données *Recon2*. Cette reconstruction du métabolisme a grandement évoluée [87, 112] depuis le développement de notre méthode et plusieurs de ses caractéristiques pourraient aider à améliorer notre compréhension au niveau biologique de l'annotation de la métrique *srd* aux paires ou aux associations entre gène et métabolite. Les caractéristiques sont les suivantes: 1) l'ensemble des

réactions métaboliques de la base de données est compris dans un même réseau métabolique qui comprend l'ensemble des voies métaboliques, 2) *Recon2* contient un nombre de métabolites plus élevé que *KEGG*, notamment, elle contient une plus grande diversité lipidique, incluant les acylcarnitines qui ne sont généralement pas rapportés par *KEGG*, 3) *Recon2* différencie les compartiments cellulaires et 4) répertorie les gènes encodant les transporteurs. Pour l'ensemble de ces raisons, *Recon2* offre un réseau métabolique qui se rapproche à l'heure actuelle le plus de la réalité *in vivo* du métabolisme humain.

La récente émergence de la lipidomique, une sous-catégorie de la métabolomique, engendre le développement de nouvelles méthodes pour mesurer et quantifier ces espèces métaboliques. À ce jour, on estime que les lipides représentent environ 75% - 80% du métabolome chez l'humain [161] et qu'ils ont jouent un rôle dans un large éventail de phénotypes. Toutefois, puisque peu de lipides sont dans les bases de données, les voies métaboliques lipidiques sont en conséquence peu présentes dans les reconstructions du métabolisme. Ainsi, la mise en place d'initiatives pour améliorer la présence et la précision du contenu des voies métaboliques est primordiale et serait un grand avantage pour améliorer les performance de PathQuant puisque 30% des associations rapportés par Shin *et al.* impliquaient un lipide.

3.5 Autres applications

L'annotation de paires ou associations gène-métabolite par la métrique *srd* s'avère intéressante dans différents contextes, entre autres deux options seront discutées ici: la prédiction d'une liste de métabolites à mesurer à partir de l'architecture génétique d'un phénotype et l'annotation d'un gène putatif aux *SNPs*.

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

Définir un profil métabolique ciblé à analyser pour un phénotype donné n'est pas simple et souvent une approche de métabolique non-ciblée est utilisée pour explorer le profil métabolique global du phénotype. Cependant, l'analyse des données provenant de l'approche non-ciblée est complexe et requiert un certain temps. C'est pourquoi, le développement d'approches permettant de mieux cibler le spectre des métabolites à analyser pour un phénotype donné serait bénéfique pour réduire la production de mégadonnées.

Les résultats obtenus en analysant les données provenant des *mGWAS* suggèrent que l'expression des gènes encodant des enzymes a un impact sur les métabolites à proximité de la réaction catalysée par l'enzyme. Ceci dit, il serait possible pour un phénotype dont on connait l'architecture génétique, de cartographier les gènes encodant des enzymes et de sélectionner l'ensemble des métabolites ayant une valeur *srd* inférieure à cinq pour définir des classes de métabolites étant pertinentes à analyser avec une étude de métabolomique. Il faut noter que cinq est une valeur arbitraire, il serait également possible d'être plus stringent en utilisant 1 ou 2 comme valeur maximale. Cette sélection de métabolite serait un point de départ intéressant pour aider les spécialistes dans le traitement des mégadonnées, mais également pour les scientifiques non-spécialistes souhaitant utiliser la métabolomique pour mieux comprendre un phénotype étudié.

3.5.2 Annotation d'un gène supposé associé aux *SNPs*

L'annotation d'un gène putatif pour un *SNP* donné est une étape cruciale pour une bonne compréhension de l'interaction entre gène et métabolite. Dans l'étude de Shin *et al*., l'annotation

a été basée sur les connaissances présentes dans les bases de données *KEGG* et *BRENDA*. Toutefois, aucun outil n'a été mentionné pour accomplir cette tâche, nous avons donc testé si la métrique *srd* serait pertinente pour faire une telle annotation. Comme discuté à la section 2.5.7, nous avons bâti une liste des gènes candidats présents dans la zone de 500 kb autour de chaque *SNP* faisant partie d'une association annotée par une valeur *srd* (Table 2.S5). Nous avons trouvé que 28 des 49 *SNPs* étaient annotés au seul gène candidat encodant une enzyme. De plus, les *SNPs* ayant plusieurs gènes candidats encodant une enzyme dans la zone de 500 kb étaient majoritairement annotés au gène ayant la plus petite valeur *srd*. Ces résultats préliminaires suggèrent que le calcul du *srd* pourrait être utile pour automatiser l'annotation des *SNPs* à leur gène putatif. Une vérification manuelle de l'annotation serait néanmoins nécessaire puisque seulement les gènes encodant des enzymes peuvent être évalués avec la métrique *srd.* Un gène ayant un autre type de fonction cellulaire pourrait s'avérer être le gène associé au *SNP* donné.

Conclusion

Dans le cadre de cette étude, nous avons développé *PathQuant*, une librairie R permettant d'annoter des paires entre gène et métabolite. Spécifiquement, notre approche permet d'annoter une valeur *srd* à des paires gène-métabolite en calculant le court chemin réactionnel les séparant sur les voies métaboliques de la base de données *KEGG*. Cette méthode a été développée selon l'hypothèse que les gènes qui encodent des enzymes catalysant des réactions métaboliques peuvent avoir un impact important sur les concentrations des métabolites à leur proximité dans les voies métaboliques. Notre approche a été testée en utilisant un jeu de données *mGWAS* publié par Shin *et al.* rapportant des associations statistiques entre *SNP* (annoté à leur gène supposé associé) et métabolite. Dans ce jeu de données, 79 associations impliquaient un gène encodant une enzyme et un métabolite identifié étant présent dans la base de données *KEGG*, de ces dernières, 49 ont pu être annotées d'une valeur *srd* dont la médiane est de 1. Nous avons également analyser méticuleusement les associations annotées par une valeur *srd* étant plus grandes que cinq ou infini. Ceci a permis de découvrir les différentes utilités et le potentiel de la méthode pour évaluer le contenu des bases de données ainsi que pour générer des hypothèses quant au contenu potentiellement manquant de leurs voies métaboliques. De plus, nous avons produit des résultats préliminaires en appliquant la méthode avec la base de données *Recon2* et le même jeu de données avec lequel des résultats similaires ont été obtenus avec une médiane de 1.

Nous avons développé *PathQuant* une libraire R robuste et facilement modifiable qui au meilleur de nos connaissances est le seul outil permettant d'annoter des paires gène-métabolite en utilisant les voies métaboliques de la base de données *KEGG* tout en respectant leur topologie. Cet outil permet également de reproduire facilement les résultats obtenus. À la lumière de notre analyse,

cette annotation permet de rapidement évaluer le lien biologique entre un gène et métabolite présent dans les voies métaboliques de la base de données utilisées. Toutefois, il est nécessaire d'améliorer le nombre d'associations pouvant être annotées par la métrique en incluant l'utilisation d'autres bases de données, particulièrement nous proposons l'utilisation de *Recon2*. Toutefois, les différences topologiques entre *KEGG* et *Recon2* obligent une utilisation et une comparaison minutieuse de leurs résultats. En somme, l'annotation des paires *mGWAS* par la métrique *srd* a permis de mieux comprendre le lien biologique entre les gènes et les métabolites. De plus, cette métrique pourrait être utilisée pour prédire un ensemble de métabolites appropriés à analyser en fonction de l'architecture génétique d'un phénotype donné en plus de permettre d'émettre un hypothèse sur le gène qui serait associé à un *SNP* donné.

Bibliographie

1. Watson, J.D. and F.H. Crick, Genetical implications of the structure of deoxyribonucleic acid. Nature, 1953. **171**(4361): p. 964-7.

2. Visscher, P.M., et al., Five years of GWAS discovery. Am J Hum Genet, 2012. **90**(1): p. 7-24.

3. Avery, O.T., C.M. Macleod, and M. McCarty, Studies on the chemical nature oof the substance inducing transformation of pneumoccocal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med, 1944. **79**(2): p. 137-58.

4. Griffith, F., The Significance of Pneumococcal Types. J Hyg (Lond), 1928. **27**(2): p. 113-59.

5. Watson, J.D. and F.H. Crick, Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 1953. **171**(4356): p. 737-8.

6. Buescher, J.M., et al., Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism. Science, 2012. **335**(6072): p. 1099-103.

7. Buescher, J.M. and E.M. Driggers, Integration of omics: more than the sum of its parts. Cancer Metab, 2016. **4**: p. 4.

8. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome Biol, 2017. **18**(1): p. 83.

9. Morange, M., What history tells us XIII. Fifty years of the Central Dogma. J Biosci, 2008. **33**(2): p. 171-5.

10. MW, C., *Human Retroviruses.* 1996. Medical Microbiology. 4th edition.: p. Chapter 62.

11. Hu, W.S. and S.H. Hughes, *HIV-1 reverse transcription.* Cold Spring Harb Perspect Med, 2012. **2**(10).

12. Bernstein, B.E., et al., The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol, 2010. **28**(10): p. 1045-8.

13. Portela, A. and M. Esteller, Epigenetic modifications and human disease. Nat Biotechnol, 2010. **28**(10): p. 1057-68.

14. Pearson, H., Meet the human metabolome. Nature, 2007. **446**(7131): p. 8.

15. Zamboni, N., A. Saghatelian, and G.J. Patti, Defining the metabolome: size, flux, and regulation. Mol Cell, 2015. **58**(4): p. 699-706.

16. Viant, M.R., et al., How close are we to complete annotation of metabolomes? Curr Opin Chem Biol, 2017. **36**: p. 64-69.

17. Satterlee, J.S., D. Schubeler, and H.H. Ng, Tackling the epigenome: challenges and opportunities for collaboration. Nat Biotechnol, 2010. **28**(10): p. 1039-44.

18. Veenstra, T.D., Metabolomics: the final frontier? Genome Med, 2012. **4**(4): p. 40.

19. Kan, M., M. Shumyatcher, and B.E. Himes, Using omics approaches to understand pulmonary diseases. Respir Res, 2017. **18**(1): p. 149.

20. Bush, W.S. and J.H. Moore, Chapter 11: Genome-wide association studies. PLoS Comput Biol, 2012. **8**(12): p. e1002822.

21. Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001. **409**(6822): p. 860-921.

22. Venter, J.C., et al., The sequence of the human genome. Science, 2001. **291**(5507): p. 1304-51.

23. Finishing the euchromatic sequence of the human genome. Nature, 2004. **431**(7011): p. 931-45.

24. Heather, J.M. and B. Chain, The sequence of sequencers: The history of sequencing DNA. Genomics, 2016. **107**(1): p. 1-8.

25. Kodama, Y., M. Shumway, and R. Leinonen, The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res, 2012. **40**(Database issue): p. D54-6.

26. Frazer, K.A., et al., A second generation human haplotype map of over 3.1 million SNPs. Nature, 2007. **449**(7164): p. 851-61.

27. Auton, A., et al., A global reference for human genetic variation. Nature, 2015. **526**(7571): p. 68-74.

28. Loh, P.R., et al., Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet, 2016. **48**(11): p. 1443-1448.

29. Wang, Q., Q. Lu, and H. Zhao, A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Front Genet, 2015. **6**: p. 149.

30. Klein, R.J., et al., Complement factor H polymorphism in age-related macular degeneration. Science, 2005. **308**(5720): p. 385-9.

31. MacArthur, J., et al., The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res, 2017. **45**(D1): p. D896-d901.

32. Hirschhorn, J.N. and M.J. Daly, Genome-wide association studies for common diseases and complex traits. Nat Rev Genet, 2005. **6**(2): p. 95-108.

33. Auer, P.L. and G. Lettre, Rare variant association studies: considerations, challenges and opportunities. Genome Med, 2015. **7**(1): p. 16.

34. Ward, L.D. and M. Kellis, HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res, 2012. **40**(Database issue): p. D930-4.

35. Boyle, A.P., et al., Annotation of functional variation in personal genomes using RegulomeDB. Genome Res, 2012. **22**(9): p. 1790-7.

36. Mootha, V.K., et al., Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci U S A, 2003. **100**(2): p. 605-10.

37. Thompson Legault, J., et al., A Metabolic Signature of Mitochondrial Dysfunction Revealed through a Monogenic Form of Leigh Syndrome. Cell Rep, 2015. **13**(5): p. 981-9.

38. Cuillerier, A., et al., Loss of Hepatic Lrpprc Alters Mitochondrial Bioenergetics, Regulation of Permeability Transition and Trans-Membrane Ros Diffusion. Hum Mol Genet, 2017.

39. Visscher, P.M., et al., 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet, 2017. **101**(1): p. 5-22.

40. Fiehn, O., Metabolomics--the link between genotypes and phenotypes. Plant Mol Biol, 2002. **48**(1-2): p. 155-71.

41. Semmar, N., Metabotype Concept: Flexibility, Usefulness and Meaning in Different Biological Populations, Metabolomics. InTech, 2012.

42. Patti, G.J., O. Yanes, and G. Siuzdak, Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol, 2012. **13**(4): p. 263-9.

43. Misra, B.B. and J.J. van der Hooft, Updates in metabolomics tools and resources: 2014-2015. Electrophoresis, 2016. **37**(1): p. 86-110.

44. Lim, E., et al., T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic Acids Res, 2010. **38**(Database issue): p. D781-6.

45. Degtyarenko, K., et al., ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res, 2008. **36**(Database issue): p. D344-50.

46. Degtyarenko, K., et al., ChEBI: an open bioinformatics and cheminformatics resource. Curr Protoc Bioinformatics, 2009. **Chapter 14**: p. Unit 14.9.

47. Ogata, H., et al., KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 1999. **27**(1): p. 29-34.

48. Kim, S., et al., PubChem Substance and Compound databases. Nucleic Acids Res, 2016. **44**(D1): p. D1202-13.

49. Wishart, D.S., et al., HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res, 2013. **41**(Database issue): p. D801-7.

50. Frolkis, A., et al., SMPDB: The Small Molecule Pathway Database. Nucleic Acids Res, 2010. **38**(Database issue): p. D480-7.

51. Fuhrer, T. and N. Zamboni, High-throughput discovery metabolomics. Curr Opin Biotechnol, 2015. **31**: p. 73-8.

52. Alonso, A., S. Marsal, and A. Julia, Analytical methods in untargeted metabolomics: state of the art in 2015. Front Bioeng Biotechnol, 2015. **3**: p. 23.

53. Cambiaghi, A., M. Ferrario, and M. Masseroli, Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Brief Bioinform, 2017. **18**(3): p. 498-510.

54. Wishart, D.S., *Quantitative metabolomics using NMR.* Trends in Analytical Chemistry, 2008.

55. Scalbert, A., et al., Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. Metabolomics, 2009. **5**(4): p. 435-458.

56. Katajamaa, M. and M. Oresic, Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics, 2005. **6**: p. 179.

57. Bowen, B.P. and T.R. Northen, Dealing with the unknown: metabolomics and metabolite atlases. J Am Soc Mass Spectrom, 2010. **21**(9): p. 1471-6.

58. Guma, M., S. Tiziani, and G.S. Firestein, Metabolomics in rheumatic diseases: desperately seeking biomarkers. Nat Rev Rheumatol, 2016. **12**(5): p. 269-81.

59. Codrea, M.C., et al., Tools for computational processing of LC-MS datasets: a user's perspective. Comput Methods Programs Biomed, 2007. **86**(3): p. 281-90.

60. Xia, J. and D.S. Wishart, Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. Curr Protoc Bioinformatics, 2011. **Chapter 14**: p. Unit 14.10.

61. Xia, J. and D.S. Wishart, Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. Curr Protoc Bioinformatics, 2016. **55**: p. 14.10.1-14.10.91.

62. Wishart, D.S., et al., HMDB: the Human Metabolome Database. Nucleic Acids Res, 2007. **35**(Database issue): p. D521-6.

63. Noelia Clemente Plaza1, M.R.G.-G., 3 and Rosa María Martínez-Espinosa1*, *Impact of the "Omics Sciences" in Medicine: New Era for Integrative Medicine.* Journal of Clinical Microbiology and Biochemical Technology, 2017. **3(1):**: p. 009-013.

64. Adamski, J., Genome-wide association studies with metabolomics. Genome Med, 2012. **4**(4): p. 34.

65. Ji, B., et al., Transcriptomic and metabolomic profiling of chicken adipose tissue in response to insulin neutralization and fasting. BMC Genomics, 2012. **13**: p. 441.

66. Petersen, A.K., et al., Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. Hum Mol Genet, 2014. **23**(2): p. 534-45.

67. Braun, K.V.E., et al., Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. Clin Epigenetics, 2017. **9**: p. 15.

68. Inouye, M., et al., Metabonomic, transcriptomic, and genomic variation of a population cohort. Mol Syst Biol, 2010. **6**: p. 441.

69. Derks, K.W., J.H. Hoeijmakers, and J. Pothof, The DNA damage response: the omics era and its impact. DNA Repair (Amst), 2014. **19**: p. 214-20.

70. Wanichthanarak, K., J.F. Fahrmann, and D. Grapov, Genomic, Proteomic, and Metabolomic Data Integration Strategies. Biomark Insights, 2015. **10**(Suppl 4): p. 1-6.

71. Ebrahim, A., et al., Multi-omic data integration enables discovery of hidden biological regularities. Nat Commun, 2016. **7**: p. 13091.

72. Shin, S.Y., et al., An atlas of genetic influences on human blood metabolites. Nat Genet, 2014. **46**(6): p. 543-50.

73. Kastenmuller, G., et al., Genetics of human metabolism: an update. Hum Mol Genet, 2015. **24**(R1): p. R93-r101.

74. Long, T., et al., Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nat Genet, 2017. **49**(4): p. 568-578.

75. Krumsiek, J., et al., Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol, 2011. **5**: p. 21.

76. Kumari, S., et al., Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes. BMC Bioinformatics, 2016. **17**: p. 132.

77. O'Brien, E.J., et al., Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol, 2013. **9**: p. 693.

78. O'Brien, E.J., J.M. Monk, and B.O. Palsson, Using Genome-scale Models to Predict Biological Capabilities. Cell, 2015. **161**(5): p. 971-87.

79. Langfelder, P. and S. Horvath, WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 2008. **9**: p. 559.

80. Rotival, M. and E. Petretto, Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits. Brief Funct Genomics, 2014. **13**(1): p. 66-78.

81. Xia, J. and D.S. Wishart, MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W71-7.

82. Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

83. Croft, D., et al., Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res, 2011. **39**(Database issue): p. D691-7.

84. Nookaew, I., et al., Genome-scale metabolic models of Saccharomyces cerevisiae. Methods Mol Biol, 2011. **759**: p. 445-63.

85. Gebauer, J., et al., A Genome-Scale Database and Reconstruction of Caenorhabditis elegans Metabolism. Cell Syst, 2016. **2**(5): p. 312-22.

86. Sigurdsson, M.I., et al., A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. BMC Syst Biol, 2010. **4**: p. 140.

87. Thiele, I., et al., A community-driven global reconstruction of human metabolism. Nat Biotechnol, 2013. **31**(5): p. 419-25.

88. Bordbar, A., et al., A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. BMC Syst Biol, 2011. **5**: p. 180.

89. Ryu, J.Y., H.U. Kim, and S.Y. Lee, Reconstruction of genome-scale human metabolic models using omics data. Integr Biol (Camb), 2015. **7**(8): p. 859-68.

90. Yugi, K., et al., Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. Trends Biotechnol, 2016. **34**(4): p. 276-90.

91. Adamski, J. and K. Suhre, Metabolomics platforms for genome wide association studies--linking the genome to the metabolome. Curr Opin Biotechnol, 2013. **24**(1): p. 39-47.

92. Petersen, A.K., et al., On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. BMC Bioinformatics, 2012. **13**: p. 120.

93. Hsia, D.Y., Phenylketonuria: the phenylalanine-tyrosine ratio in the detection of the heterozygous carrier. J Ment Defic Res, 1958. **2**(1): p. 8-16.

94. Suhre, K., J. Raffler, and G. Kastenmuller, Biochemical insights from population studies with genetics and metabolomics. Arch Biochem Biophys, 2016. **589**: p. 168-76.

95. Krumsiek, J., et al., Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. PLoS Genet, 2012. **8**(10): p. e1003005.

96. Shin S-Y, F.E., Petersen A-K, Krumsiek J, Santos R, Huang J, et al., GGM network - Online Supplement. http://metabolomics.helmholtz-muenchen.de/gwa/si/network.php. Accessed 15 Apr 2015.

97. Dagley, S., E.D. Nicholson, and V.J. Caldwell, *An Introduction to Metabolic Pathway.* Oxford and Edinburgh: Blackwell scientific Publications, 1970.

98. Romero, P., et al., Computational prediction of human metabolic pathways from the complete human genome. Genome Biol, 2005. **6**(1): p. R2.

99. Sud, M., et al., LMSD: LIPID MAPS structure database. Nucleic Acids Res, 2007. **35**(Database issue): p. D527-32.

100. Chang, A., et al., BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic Acids Res, 2015. **43**(Database issue): p. D439-46.

101. Kelder, T., et al., WikiPathways: building research communities on biological pathways. Nucleic Acids Res, 2012. **40**(Database issue): p. D1301-7.

102. Hao, T., et al., Compartmentalization of the Edinburgh Human Metabolic Network. BMC Bioinformatics, 2010. **11**: p. 393.

103. Pornputtapong, N., I. Nookaew, and J. Nielsen, Human metabolic atlas: an online resource for human metabolism. Database (Oxford), 2015. **2015**: p. bav068.

104. Tipton, K. and S. Boyce, History of the enzyme nomenclature system. Bioinformatics, 2000. **16**(1): p. 34-40.

105. Brown, H.A. and L.J. Marnett, Introduction to lipid biochemistry, metabolism, and signaling. Chem Rev, 2011. **111**(10): p. 5817-20.

106. Orth, J.D. and B.O. Palsson, Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng, 2010. **107**(3): p. 403-12.

107. Satish Kumar, V., M.S. Dasika, and C.D. Maranas, Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics, 2007. **8**: p. 212.

108. Caspi, R., et al., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res, 2014. **42**(Database issue): p. D459-71.

109. Gene Ontology Consortium: going forward. Nucleic Acids Res, 2015. **43**(Database issue): p. D1049-56.

110. Kanehisa, M., et al., KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.

111. Thiele, I. and B.O. Palsson, Reconstruction annotation jamborees: a community approach to systems biology. Mol Syst Biol, 2010. **6**: p. 361.

112. Duarte, N.C., et al., Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1777-82.

113. Bordbar, A. and B.O. Palsson, Using the reconstructed genome-scale human metabolic network to study physiology and pathology. J Intern Med, 2012. **271**(2): p. 131-41.

114. Wu, M. and C. Chan, Human metabolic network: reconstruction, simulation, and applications in systems biology. Metabolites, 2012. **2**(1): p. 242-53.

115. Quek, L.E., et al., Reducing Recon 2 for steady-state flux analysis of HEK cell culture. J Biotechnol, 2014. **184**: p. 172-8.

116. Noronha, A., et al., ReconMap: an interactive visualization of human metabolism. Bioinformatics, 2017. **33**(4): p. 605-607.

117. Swainston, N., et al., Recon 2.2: from reconstruction to model of human metabolism. Metabolomics, 2016. **12**: p. 109.

118. Hucka, M., et al., The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics, 2003. **19**(4): p. 524-31.

119. Stobbe, M.D., et al., Critical assessment of human metabolic pathway databases: a stepping stone for future integration. BMC Syst Biol, 2011. **5**: p. 165.

120. Stobbe, M.D., et al., Improving the description of metabolic networks: the TCA cycle as example. Faseb j, 2012. **26**(9): p. 3625-36.

121. Mason, O. and M. Verwoerd, Graph theory and networks in Biology. IET Syst Biol, 2007. **1**(2): p. 89-119.

122. Lacroix, V., et al., An introduction to metabolic networks and their structural analysis. IEEE/ ACM Trans Comput Biol Bioinform, 2008. **5**(4): p. 594-617.

123. Li, B.Q., et al., Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PLoS One, 2012. **7**(4): p. e33393.

124. Jiang, Y., et al., Identifying gastric cancer related genes using the shortest path algorithm and protein-protein interaction network. Biomed Res Int, 2014. **2014**: p. 371397.

125. Stam, C.J., et al., Small-world networks and functional connectivity in Alzheimer's disease. Cereb Cortex, 2007. **17**(1): p. 92-9.

126. Frainay, C. and F. Jourdan, Computational methods to identify metabolic sub-networks based on metabolomic profiles. Brief Bioinform, 2016.

127. Ravasz, E., et al., Hierarchical organization of modularity in metabolic networks. Science, 2002. **297**(5586): p. 1551-5.

128. Jeong, H., et al., The large-scale organization of metabolic networks. Nature, 2000. **407**(6804): p. 651-4.

129. SS., S., The Algorithm Design Manual. Springer London, 2008.

130. Jostins, L., et al., Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature, 2012. **491**(7422): p. 119-24.

131. Budhu, A., et al., Integrated metabolite and gene expression profiles identify lipid biomarkers associated with progression of hepatocellular carcinoma and patient outcomes. Gastroenterology, 2013. **144**(5): p. 1066-1075.e1.

132. Suhre, K. and C. Gieger, Genetic variation in metabolic phenotypes: study designs and applications. Nat Rev Genet, 2012. **13**(11): p. 759-69.

133. Heemskerk, M.M., et al., Reanalysis of mGWAS results and in vitro validation show that lactate dehydrogenase interacts with branched-chain amino acid metabolism. Eur J Hum Genet, 2016. **24**(1): p. 142-5.

134. Pavlopoulos, G.A., et al., Using graph theory to analyze biological networks. BioData Min, 2011. **4**: p. 10.

135. EW., D., A note on two problems in connexion with graphs. Numer. Math., 1959.

136. Dumas, M.E., et al., Topological analysis of metabolic networks integrating co-segregating transcriptomes and metabolomes in type 2 diabetic rat congenic series. Genome Med, 2016. **8**(1): p. 101.

137. Fuhrer, T., et al., Genomewide landscape of gene-metabolome associations in Escherichia coli. Mol Syst Biol, 2017. **13**(1): p. 907.

138. Rodriguez-Martinez, A., et al., MetaboSignal: a network-based approach for topological analysis of metabotype regulation via metabolic and signaling pathways. Bioinformatics, 2017. **33**(5): p. 773-775.

139. Li, B.Q., et al., Identification of retinoblastoma related genes with shortest path in a protein-protein interaction network. Biochimie, 2012. **94**(9): p. 1910-7.

140. Li, B.Q., et al., Identification of lung-cancer-related genes with the shortest path approach in a protein-protein interaction network. Biomed Res Int, 2013. **2013**: p. 267375.

141. Xu, Y., et al., Identification of thyroid carcinoma related genes with mRMR and shortest path approaches. PLoS One, 2014. **9**(4): p. e94022.

142. Yuan, F., et al., Mining for Candidate Genes Related to Pancreatic Cancer Using Protein-Protein Interactions and a Shortest Path Approach. Biomed Res Int, 2015. **2015**: p. 623121.

143. Yuan, F., et al., Identification of Candidate Genes Related to Inflammatory Bowel Disease Using Minimum Redundancy Maximum Relevance, Incremental Feature Selection, and the Shortest-Path Approach. Biomed Res Int, 2017. **2017**: p. 5741948.

144. R: The R Project for Statistical Computing. https://www.r-project.org/. Accessed 17 Feb 2017.

145. PathQuant. https://github.com/sandraTL/PathQuant. Accessed 1rst Feb 2017.

146. KEGG API. http://www.kegg.jp/kegg/docs/keggapi.html. Accessed 21 Oct 2014.

147. Kanehisa, M., Chemical and genomic evolution of enzyme-catalyzed reaction networks. FEBS Lett, 2013. **587**(17): p. 2731-7.

148. Cottret, L., et al., MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W132-7.

149. LinkDB. http://www.genome.jp/linkdb/. Accessed 21 Oct 2014.

150. Friolet, R., H. Hoppeler, and S. Krahenbuhl, 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. J Clin Invest, 1994. **94**(4): p. 1490-5.

151. Mootha, V.K. and J.N. Hirschhorn, *Inborn variation in metabolism.* Nat Genet, 2010. **42**(2): p. 97-8.

152. Prakash, S., Human metabolic individuality in biomedical and pharmaceutical research. Circ Cardiovasc Genet, 2011. **4**(6): p. 714-5.

153. Cottret, L. and F. Jourdan, Graph methods for the investigation of metabolic networks in parasitology. Parasitology, 2010. **137**(9): p. 1393-407.

154. Faust, K., D. Croes, and J. van Helden, *Metabolic pathfinding using RPAIR annotation.* J Mol Biol, 2009. **388**(2): p. 390-414.

155. Stobbe, M.D., et al., Knowledge representation in metabolic pathway databases. Brief Bioinform, 2014. **15**(3): p. 455-70.

156. Schomburg, I., et al., BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res, 2013. **41**(Database issue): p. D764-72.

157. Reimers, M. and V.J. Carey, Bioconductor: an open source framework for bioinformatics and computational biology. Methods Enzymol, 2006. **411**: p. 119-34.

158. D.R., B., *Github.* Journal of the Medical Library Association : JMLA.

159. Eyre, T.A., et al., The HUGO Gene Nomenclature Database, 2006 updates. Nucleic Acids Res, 2006. **34**(Database issue): p. D319-21.

160. Stobbe, M.D., Metabolic Pathway Databases: A Word of Caution. 2015: p. 27-63.

161. Quehenberger, O. and E.A. Dennis, *The human plasma lipidome.* N Engl J Med, 2011. **365**(19): p. 1812-23.