Genetic variations in connection

Genetic variations in connection

Understanding the effects of Single Nucleotide Polymorphisms in their biological context

The research presented in this dissertation was conducted at NUTRIM School of Nutrition and Toxicology and Metabolism of Maastricht University and the department of Bioinformatics-BiGCaT.

Cover design: Remco Wetzels (Proefschriftmaken.nl)
Layout by: Elisa Cirillo
Printed by: Proefschriftmaken.nl
Published by: Proefschritfmaken.nl

ISBN: 978-90-829118-4-8

Genetic variations in connection

Understanding the effects of Single Nucleotide Polymorphisms in their biological context

DISSERTATION

to obtain the degree of Doctor at the Maastricht University, on the authority of the Rector Magnificus, Prof.dr. Rianne M. Letschert in accordance with the decision of the Board of Deans, to be defended in public on Friday 18th January 2019, 14.00 hours

by

Elisa Cirillo

Supervisor

Prof. Dr. Chris T. A. Evelo

Co-supervisors

Dr. Susan L.M. Coort

Dr. Laurence D. Parnell (Tufts University)

Assessment Comittee

Prof. Dr. Maurice P.A. Zeegers (Chair)

Prof. Dr. Ellen E. Blaak

Prof. Dr. Annemie M.W.J. Schols

Prof. Dr. Peter-Bram 't Hoen (Radboud University)

Dr. Marco Roos (Leiden University Medical Center)

TABLE OF CONTENTS

CHAPTER 1

General introduction

The beauty and the tragedy of a single nucleotide polymorphism

The beauty

Humans (*Homo sapiens*) have 99.5% to 99.8% of their genome essentially identical in everyone [19]. However, there is a large variety of different types of humans in the world. For example, it is generally not common to find two people that look exactly alike. These differences are even greater in the molecular and physiological mechanisms of the inner body. The actual human DNA variability is estimated from 0.2% to 0.5% of 3 billion nucleotides, but this small percentage reflects a very large number of variations in the DNA, from 6 to 15 million nucleotides. Part of these variations, both common and rare ones, are in the category of the Single Nucleotide Variants (SNVs) [19], in which are included the Single Nucleotide Polymorphisms (SNPs). A SNP is a DNA sequence variant, occurring when a single nucleotide (adenine, guanine, thymine, or cytosine) differs between members of a species or paired chromosomes in an individual. Thus, the beauty (Figure 1.1) of roughly 10 million variants is that they can potentially occur in numerous different combinations in various individuals. This is vastly more than enough to ensure individual uniqueness at the DNA level, while still representing a very small fraction of the total genome. Even in monozygotic twins, that essentially have the same chromosomal DNA sequence, it is possible to find differences in their physical characteristics. These differences reflect a DNA variability originating from somatic mutations, caused by small errors in DNA replication after the four- to eight-cell zygote stage; and from the action of the epigenetic mechanisms that play a role in the DNA plasticity and availability for the transcription process [6].

The tragedy

In general, the majority of the human SNPs are located between genes and often they do not have any deleterious effects on the functionality of the gene or its encoded proteins. The tragedy (Figure 1.1) occurs when the SNPs located in the coding area or those placed in the regulatory regions of a gene, such as promoter or splicing sites, show a deleterious impact on the transcription of the gene or the protein activity, for instance. When this happens in a gene that plays a key role in a biological pathway, serious diseases such as the Rett syndrome for the monogenic X-dominant *MECP2* gene [23] can occur. The scenario becomes convoluted in complex diseases such as: diabetes, obesity, and chronic obstructive pulmonary disease (COPD), to name a few, where the genetic background of the disease is most likely characterized by a combination of variations occurring in an unknown number of genes, usually interacting with various environ-

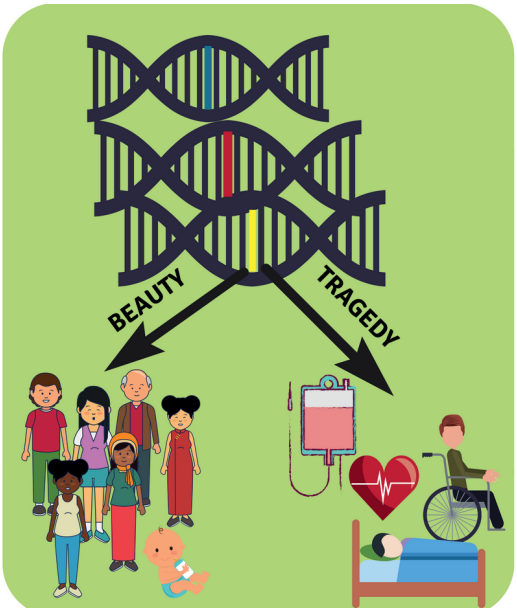

Fig. 1.1 Representation of the beauty and the tragedy of SNPs consequence.

mental factors [9]. Although complex disorders often cluster in families, they do not have a clear pattern of inheritance. For this reason, it is difficult to determine a person's risk of inheriting or passing on these disorders. Since 2005, population genetics has accelerated with the use of Genome Wide Association Study (GWAS) and the ability to examine the genotype-phenotype relationship with at first 100,000 genetic variants and now routinely 1 million (or more). Population genetic studies have been ongoing for decades but with much slower and less precise methods. These studies have been used to investigate the genetic background of individuals with complex diseases, and they are performed comparing the DNA sequences of genetically similar individuals (from the same population origin) with (cases) and without (controls) specific phenotypic traits [10]. In particular, the comparison is performed between the alleles of cases and controls that have been detected through the whole genome, resulting in a list of SNPs significantly associated with the specific trait or disease. The core data analyzed in the different chapters of this thesis are SNPs detected with GWAS studies and associated with different types of complex diseases and traits such as: Type 2 diabetes mellitus (T2DM) and obesity.

Genome-wide association studies as a tool to unravel complex diseases

GWAS studies are the first step, followed by functional genomics studies, that can be used either to identify causal or predictive factors for a given trait or to investigate the genetic architecture of that trait. For example, the genetic risk factors identified from the list of SNPs associated with a complex disease, allows development of susceptibility testing for disease prediction [16]. For clinicians, knowing the genetic susceptibility of a patient especially for complex diseases can improve the diagnosis and inform the choice of treatment [22]. Because these diseases have a strong environmental component, the correction of the lifestyle is considered both a form of prevention and a treatment. Indeed, the environmental changes could help to balance the genetic predisposition. In this regard, obtaining genetic information can launch a warning message to take action. On the other hand, from the researchers perspective GWAS studies are extensively used to identify the genetic factors contributing to disease phenotypes and to elucidate the extent to which those genetic factors affect the pathophysiology of a disease. This type of fundamental knowledge is at the base of drug design, especially in relation of pharmacogenetics. Inherited genetic variations present in drug targets or in enzymes that metabolize a drug [4] can affect the individual responses both in terms of adverse effects and therapeutic effects. Knowing the existence and the role of these variations (Figure 1.2) for specific diseases enables the design or selection of drugs for a personalized treatment [25]. After more than a decade of GWAS experience, researchers have recognized that the power of GWAS to identify within a population a true association between a SNP and a trait is dependent on the phenotypic variance explained by the SNP [24]. The phenotypic variance is determined by how strongly the two allelic variants differ in their phenotypic effect (the effect size or beta coefficient), and the allele frequency in the sample. However, in complex diseases some causal SNPs [10] are common variants with a small effect size. Specific considerations need to be included in the GWAS design in order to incorporate the causal allele. Some of the issues to tackle in a GWAS study design, in order to increase statistical power and enable the detection of more meaningful associations are: i) sample size, ii) incomplete genotyping, iii) genetic heritability and iv) confounding factors [13]. Finally, an unsolved problem in GWAS is to identify the causative gene that is influenced by the GWAS variants detected, which often is not the closest gene mapped to the variant. Considering properly those issues of GWAS study design and output, facilitates obtaining high quality results and true associations.

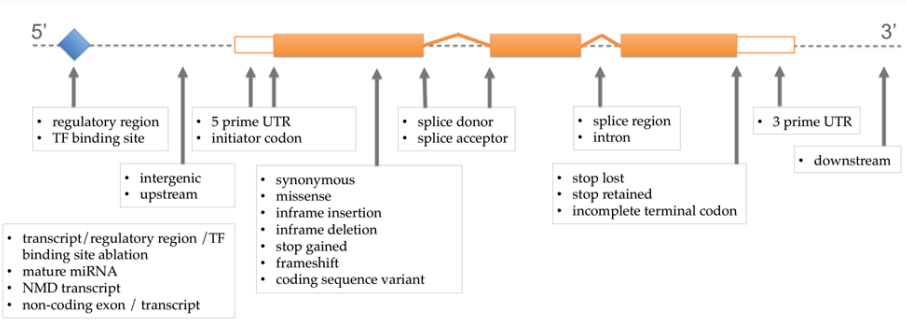

Fig. 1.2 Consequences of SNPs located inside and outside a gene region that is indicated in orange. Some of the consequences reflect the name of the genetic region, such as 3 prime UTR.

Data integration: Key to characterizing the effects of genetic variants

When the GWAS study is designed and performed according to the standards, a challenge remains to delineate the role of the significantly associated SNPs in the molecular scenario of a complex disease [24]. In order to achieve this goal Bioinformatics and System Biology approaches are needed, because these fields promote methodologies that facilitate the analysis and integration of multiple and disparate data types. The reason for data integration, related to significant SNPs identified in GWAS studies, is to enable the exploration of the biological meaning of the SNPs. In this regard, the combination of other biological data can enrich the description of the context in which the SNPs act. Integrating such multiple data types is a crucial aspect in the Bioinformatics and System Biology fields, but because biological data are diverse, complex and distributed in many different resources, this can be a challenge. This thesis will show how to link SNPs with various data from a diversity of resources, aiming to a meaningful description of the SNPs biological role especially in complex diseases. The data integrated with the GWAS data and presented in the following chapters are: i) biological pathways (containing gene products, metabolites and their interactions), ii) expression quantitative trait loci (eQTLs), iii) gene-environment interactions, iv) epigenetics and v) literature sources, see figure 1.3 .

Biological pathways

A biological pathway is a series of interactions between genes, proteins and metabolites in a cell. For many years, pathway diagrams helped researchers to illustrate and understand in which way molecular interactions of small molecules and proteins occur

in cells. This in turn supported an understanding of how cells influence each other. Pathway diagrams provide a description of the gene's interactions, with the potential to explore the consequences of genetic variations within the pathway's entities, such as an enzyme with altered kinetics. Currently, pathway representations are collected in databases such as: WikiPathways, KEGG and Reactome [8, 20, 21], and are linked to the biological knowledge related to genes and metabolites stored in other digital sources. Moreover, tools that perform integration and visualization of multiple data types on the pathway diagram are available, aiming to evaluate biological scenario in the context of processes described in pathway [11]. These data analysis and visualization automation advancements are applicable for genetic variation data as well. In addition, a specific methodology called pathway analysis is used to analyze the enrichment of the biological data in the pathway collections [14]. Meaning that a given dataset has (statistically) significantly more or fewer members of that pathway than expected. An example of a tool that performs such an analysis is PathVisio [11]. Pathway analysis enables the detection of relevant groups of related genes in case samples compared to controls, and the interaction between genes/proteins and metabolites. Recently, the GWAS data have been integrated in pathway analysis approaches [14, 15, 30, 31]. Currently the methodology has been perfected to obtain relevant pathway results based on the associations of SNPs to a disease [17]. However, pathway analysis has also limitations such as: knowledge biases and methodological challenges [32], but the field is developing to reduce such constraints. In this thesis the pathway context and the methodology are extensively used to better characterize the effect of the SNPs from the GWAS studies, with the aim of extending the initial GWAS results from individual genes to biological processes involved in the disease phenotype(s).

Expression quantitative trait loci

The fundamental challenge of SNP investigation is to understand how the variant exerts an effect on the phenotype. The mapping of eQTLs is an approach used to clarify if a variant has an influence on gene expression in a specific tissue [26]. The eQTL variants are assessed by looking at gene expression panels of genotyped individuals, and several statistical analyses have been developed since 2001 to refine this detection, combining genomics and transcriptomics data [27]. Currently, different online eQTL catalogues covering multiple tissues are available to facilitate researchers in this type of genetic analysis. In this thesis such resources are used and integrated with SNPs from GWAS, in order to elucidate the effect of variants that are located in non-coding regions of genes, in order to ascertain a potential regulatory role on transcription.

Epigenetics

Epigenetics is the study of heritable changes in the genome that do not involve modifications in the underlying DNA sequence. There are several types of epigenetic mechanisms that affect the modeling processes of the DNA structure such as DNA methylation [28], and histone modification [29]. Additionally, epigenetics involves microRNAs and control of translation of mRNA into protein [2]. Structural epigenetic DNA rearrangements like histone modifications, coordinate the accessibility of the DNA, opening functional regions to proteins and molecules that regulate the gene transcription. Variants located in these regulatory regions could disturb the regular functionality of the epigenetic mechanism [5]. For example, if a variant has an impact on the DNA sequence, that prevents the correct opening of the histones, this could prevent the transcription of the closes genes. Overall, the effect of these modifications: DNA methylation, histone acetylation and methylation among others, are strictly connected with the transcriptional control of the genes. In this thesis several databases with epigenetic data are consulted to investigate if a non-coding variants is located in regions with epigenetic activity. This information is used in combination with the eQTL data, because the variants that can exert an eQTL function are often located in epigenetically active areas.

Gene-environment interactions

Cells evolved mechanisms in response to environmental and external stimuli. They adjust their biochemistry in different ways such as: changes in the activities of preexisting enzyme molecules, changes in the rates of synthesis of new enzyme molecules, and changes in membrane-transport processes. The core of this response is related to the genetic control. For this reason, interactions between genes and environmental factors are measured to improve the assessment of both genotype and environmental influence on the phenotype [18]. Gene-environment (GxE) interactions describe a modifiable relationship between genetic variation and changes in phenotype due to external factors such as: diet, physical activity, smoking, sleep, alcohol intake, etc. This information can be applied in the clinic to take action in the health of the individual, especially if the aim is to modulate the adverse effects of a risk allele that participates in a genotype-phenotype relationship whereby risk is increased [12]. For this reason online resource of GxE interactions was used, in one of the chapter of this thesis, to delineate the influence of certain SNPs in a biological process after a specific dietary or physical activity.

Literature sources

Literature always has been the main resource of accumulated knowledge from where to start to formulate a scientific hypothesis or look for confirmatory or refuting evi-

dence. Nowadays, several databases such as PubMed Central https://www.ncbi.nlm.nih.gov/pmc/ and journal websites provide in most cases free access to articles. Moreover, with the advent of the semantic web field, the literature information is directly linked with the description of the biological entities placed in other databases like NCBI https://www.ncbi.nlm.nih.gov/ and Ensembl http://www.ensembl.org. These connections enable a quick consultation of the prior knowledge related to the biological entity that generally concerns wet or in silico laboratory experiments to confirm its existence and/or biological function. In the methodology presented in several thesis chapters, the literature knowledge is extensively used. In particular, a better description is presented of the role of genetic variants along with reports of any experimental validation of the variant effect.

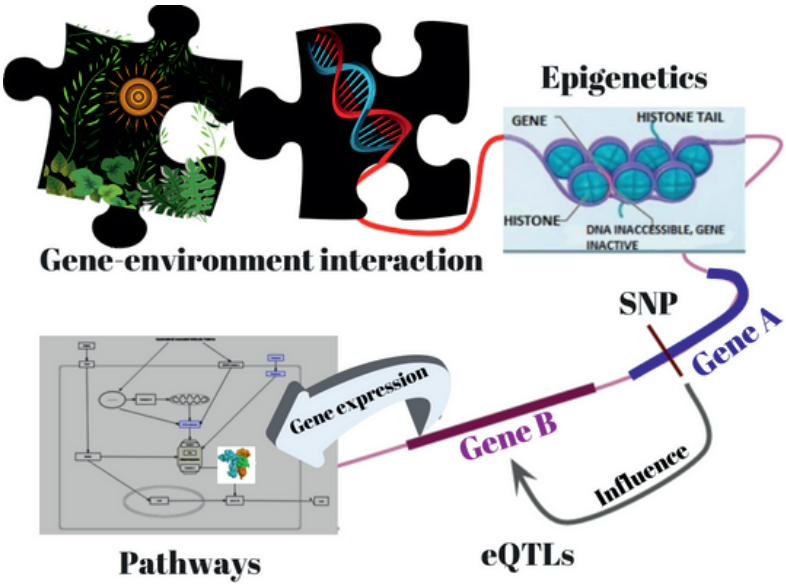

Fig. 1.3 Representation of the different types of data integrated in the analysis presented in this thesis. The data types include: gene-environment interactions, epigenetics mechanisms in particular histones modification, eQTLs in which SNPs influence the gene expression and biological pathways.

Visualizing the complexity with network analysis

The results from biological data analysis are extensively visualized with different types of representations such as graphs, plots, images, and animation [3]. Computer-based visualization tools are developing rapidly to facilitate data interpretation by using advanced data representation. Pathway analysis tools are an example in this regard. In

addition, those applications also support data analysis relying on proper statistical and computational methods. In order to implement a data analysis algorithm in a tool, it is essential to understand how the data are obtained. This is an aspect strictly related to the technologies used to generate the data. In this thesis the focus is on GWAS SNPs, obtained from DNA analysis. Although there is a variety of chemical, enzymatic or technological procedures used for determining the sequence of nucleotide bases in DNA [1], currently a big distinction is made between the platforms used such as: microarrays and next generation sequencing (NGS). The microarrays are also called DNA chip or biochip and it is a collection of microscopic DNA spots, called probes, attached to a solid surface that can hybridize a cDNA sample called target, under high-stringency conditions. The probe-target hybridization is detected and quantified by a fluorophore-, silver-, or chemo-luminescence-labeled target. The NGS uses different types of chemical or enzymatic approaches to obtain the sequences, the major differences with the microarrays are NGS technologies can sequence the entire human genome faster because of the massive parallel sequencing and NGS has the capacity for greater accuracy of the sequence, for which this technology is sometimes known as ”deep sequencing”. Despite these advantages, microarrays are still widely adopted for genotyping studies as they are substantially less expensive and require less complicated and less labor-intensive sample preparation than NGS. Moreover, for the detection of specifc well known variants SNP arrays are very useful because they give a precise yes/no signal even on a single array. These are practical advantages considering the processing of thousands of samples required for typical GWAS studies. The data from GWAS studies considered in the thesis were all obtained using microarrays. This choice influences not only the type of statistical analysis required in the computational tools, but also the way that the data are visualized as reported in the first part of the thesis.
Finally, network analysis is used as the primary methodology to support the data integration considered in this thesis and the resulting visualization. The biological data previously mentioned can be connected with the genes that carry the SNPs and their relationships can be displayed as a network. The interactions between the different biological entities can be represented either as nodes or edges that connect the nodes [7]. The network resulting from the data connection requires interpretation and this is the major focus of the second half of the thesis.

Outline of the thesis

The aim of the thesis is to investigate how SNPs from GWAS study can be analyzed towards the understanding of their role in pathways, and how extending the data pathway analysis and visualization can support this understanding. The biological scenarios, in which the role of SNPs is explored, are described for two related complex phenotypes:

T2DM and obesity.

In **Chapter 2** a review about how genetic variants are visualized and analyzed in pathway context is presented. In particular, several software packages that perform pathway analysis are evaluated and a variant use case of these tools is shown. We identified strengths and limitations of the technology for the researchers that want to use it or improve it. Then, in **Chapter 3** a gene to variant and variant to gene mapping database that can be used with the mapping tool BridgeDb is introduced. Such mapping tool is essential to enable the analysis and visualization of a variant in the pathways, since these are composed of genes and not variants, and more in general it facilitates the association of thousands of SNPs from the GWAS study to their genes. The rest of the thesis focuses on the development and application of workflows, in which SNPs from GWAS studies are re-analyzed in combination of several other biological data, in order to capture a better interpretation of the variants role. In this regard **Chapter 4** presents the design of a workflow based on pathway and network analysis, in which SNPs associated with T2DM are integrated with eQTLs and GxE interactions data. **Chapter 5** shows the same workflow applied to another GWAS study related to BMI, but in this case the data integrated are: eQTLs and epigenetic data. This data combination enabled capture of the biological interpretation of non-coding variants that have a regulatory role in the transcription process of a gene. **Chapter 6**, shows how SNPs associated with T2DM can be mapped together in a biological pathway, built based on literature information regarding a specific organ condition of T2DM individuals. The additional value is that these genes acting in the same pathway and presenting relevant T2DM genetic variations, are also detected as differentially co-expressed genes from a different transcriptomics study.

Finally, the **General Discussion** presents the significance of the results obtained applying pathway and network methods to SNPs from GWAS studies. In particular, what challenges in data integration methodologies still exist, and need to be solved in order to better describe the biological effect of the SNPs using existing data, and how this approach can help to support the development of precision medicine.

References

[1] Jay A. Shendure, Gregory J. Porreca, George M. Church, Andrew F. Gardner, Cynthia L. Hendrickson, and et al. Overview of DNA sequencing strategies. *Current Protocols in Molecular Biology*, (SUPPL.96):1–23, 2011.

[2] Jody C Chuang and Peter A Jones. Epigenetics and MicroRNAs. *Pediatric Research*, 61(5 Part 2):24R–29R, may 2007.

[3] Friedman V. Data visualiation and infographics. *Graphics, Monday inspiration*.

[4] Shabbir Ahmed, Zhan Zhou, Jie Zhou, and Shu-Qing Chen. Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine. *Genomics, Proteomics and Bioinformatics*, 14(5):298–313, oct 2016.

[5] Steven Henikoff and M Mitchell Smith. Histone variants and epigenetics. *Cold Spring Harbor perspectives in biology*, 7(1):a019364, jan 2015.

[6] Albert H.C. Wong, Irving I. Gottesman, and Arturas Petronis. Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human Molecular Genetics*, 14(suppl_1):R11–R18, apr 2005.

[7] Albert-László Barabási. Network Medicine From Obesity to the Diseasome. *New England Journal of Medicine*, 357(4):404–407, jul 2007.

[8] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, and et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, jan 2018.

[9] Arno G Motulsky. Genetics of complex diseases. *Journal of Zhejiang University. Science. B*, 7(2):167–8, feb 2006.

[10] William S. Bush and Jason H. Moore. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12), 2012.

[11] Martina Kutmon, Martijn P. van Iersel, Anwesha Bohler, Thomas Kelder, Nuno Nunes, and et al. PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLOS Computational Biology*, 11(2):e1004085, feb 2015.

[12] Laurence D Parnell, Britt A Blokker, Hassan S Dashti, Paula-Dene Nesbeth, Brittany Elle Cooper, and et al. CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits. *BioData Mining*, 7(1):21, dec 2014.

[13] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9:29, 2013.

[14] Miguel A. Garcia-Campos, Jesus Espinal-Enriquez, and Hernandez-Lemus. Pathway analysis: State of the art. *Frontiers in Physiology*, 6(DEC):1–16, 2015.

[15] Kai Wang, Mingyao Li, and Maja Bucan. Pathway-based approaches for analysis of genomewide association studies. *American journal of human genetics*, 81(6):1278–83, dec 2007.

[16] C. Marzuillo, C. De Vito, E. D'Andrea, A. Rosso, and P. Villari. Predictive genetic testing for complex diseases: a public health perspective. *QJM*, 107(2):93–97, feb 2014.

[17] Patrick Y.P. Kao, Kim Hung Leung, Lawrence W.C. Chan, Shea Ping Yip, and Maurice K.H. Yap. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochimica et Biophysica Acta - General Subjects*, 1861(2):335–353, 2017.

[18] R Ottman. Gene-environment interaction: definitions and study designs. *Preventive medicine*, 25(6):764–70, 1996.

[19] K. K. Kidd, A. J. Pakstis, W. C. Speed, and J. R. Kidd. Understanding Human DNA Sequence Variation. *Journal of Heredity*, 95(5):406–420, sep 2004.

[20] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, and et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, jan 2018.

[21] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, jan 2000.

[22] Steven J Schrodi, Shubhabrata Mukherjee, Ying Shan, Gerard Tromp, John J Sninsky, and et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Frontiers in genetics*, 5:162, 2014.

[23] Friederike Ehrhart, Susan L.M. Coort, Elisa Cirillo, Eric Smeets, Chris T. Evelo, and et al. Rett syndrome - Biological pathways leading from MECP2 to disorder phenotypes. *Orphanet Journal of Rare Diseases*, 11(1):1–13, 2016.

[24] Gerome Breen, Qingqin Li, Bryan L Roth, Patricio O'Donnell, Michael Didriksen, and et al. Translating genome-wide association findings into new therapeutics for psychiatry. *Nature Neuroscience 2016 19:11*, oct 2016.

[25] Ann K Daly. Pharmacogenetics: a general review on progress to date. *British Medical Bulletin*, 124(1):1–15, oct 2017.

[26] Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[27] Jacob J. Michaelson, Salvatore Loguercio, and Andreas Beyer. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, 48(3):265–276, 2009.

[28] Jordana T. Bell and Tim D. Spector. DNA methylation studies using twins: what are they telling us? *Genome biology*, 13(10):172, 2012.

[29] Eli J. Draizen, Alexey K. Shaytan, Leonardo Mariño-Ramírez, Paul B. Talbert, David Landsman, and et al. HistoneDB 2.0: A histone database with variants - An integrated resource to explore histones and their variants. *Database*, 2016:1–10, 2016.

[30] K Wang, M Li, H Hakonarson Nature Reviews Genetics, and undefined 2010. Analysing biological pathways in genome-wide association studies. *nature.com*.

[31] Michael A. Mooney and Beth Wilmot. Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(7):517–527, oct 2015.

[32] Lv Jin, Xiao-Yu Zuo, Wei-Yang Su, Xiao-Lei Zhao, Man-Qiong Yuan, and et al. Pathway-based Analysis Tools for Complex Diseases: A Review. *Genomics, Proteomics and Bioinformatics*, 12(5):210–220, 2014.

CHAPTER 2

A review of pathway-based analysis tools that visualize genetic variants

Elisa Cirillo[1*], Laurence D. Parnell[2], Chris T. Evelo[1]

1 Department of Bioinformatics BiGCaT, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands
2 Agricultural Research Service, USDA, Jean Mayer-USDA Human Nutrition Research Center on Aging at Tufts University, Boston, MA, USA

Published in: Frontiers in Genetics, doi: 10.3389/fgene.2017.00174

Abstract

Pathway analysis is a powerful method for data analysis in genomics, most often applied to gene expression analysis. It is also promising for genetic data analysis, including Genome Wide Association Study (GWAS) data, because it allows the interpretation of variants with respect to the biological processes in which the affected genes and proteins are involved. Such analyses support an interactive evaluation of the possible effects of variations on function, regulation or interaction of gene products. Current pathway analysis software often does not support variants data visualization in pathways as an alternate method to interpret GWAS results, nor specific statistical methods to facilitate GWAS analysis. In this review we first describe the visualization options of the tools that were identified by a literature review, in order to provide insight for improvements in this developing field. Tool evaluation was performed using a computational epistatic dataset of gene-gene interactions for obesity risk. Next, we report the necessity to include in these tools statistical methods for the pathway-based analysis in GWAS, expressly aiming to define features for more comprehensive pathway-based analysis tools. We conclude by recognizing that pathway analysis of GWAS data requires a sophisticated combination of the most useful and informative visual aspects of the various tools evaluated.

Introduction

Pathway analysis for Genome Wide Association Study data

Today, pathway analysis is routine with software or web services that accept and analyse different omics data, transcriptomics, proteomics with protein-protein interactions, and metabolomics. Methods and tools used to visualise and analyse these three main kinds of high-throughput data have been reviewed [27]. Moreover, a decade ago genetic variation data, originating from analyses of Genome Wide Association Studies (GWAS), began to be incorporated into pathway analysis [4]. Recently, several step by step guides [3, 5, 31, 33] were published as reviews, describing and providing recommendations on how to use different pathway analysis methodologies, which are applicable to GWAS data. The main features to consider are: (i) make certain that GWAS analysis is performed according to standard guidelines; (ii) choose curated and up-to-date pathway collections; (iii) filter the list of gene sets to avoid bias related to size, a common limit is between 10 and 200 genes, and map the SNPs to genes based on location or linkage disequilibrium; (iv) choose the method according to the statistical hypothesis to be tested; (iv) report the results and if applicable visualise them in order to improve comprehension. Although genetic association research is advancing rapidly,

biological interpretation remains a challenge, especially when interpretation concerns connecting genetic findings with known biological processes [8]. Application of pathway analysis to GWAS data is a valid approach to meet this challenge for different reasons: first, because of the polygenic nature of complex diseases, such an approach held the promise to contextualize better the GWAS data and to suggest novel interpretations of the results based on prior knowledge of genes and pathways [31]. Second, a typical display of GWAS results consists of the few SNPs showing strong evidence for disease or phenotype association (generally p-value minor 1e-8), but it is also well-known that these few associated SNPs often have only a modest effect on disease risk [17]. Thus, examining the cumulative effects of numerous variants can empower detection of genetic risk factors for complex diseases [8]. Finally, genetic heterogeneity within an affected population is another well-known issue in GWAS. If sufficient loci exerting small effects are present in the same gene set, it may be possible to detect their cumulative effect by testing for associations at the pathway level [33].In general improving and standardizing the practice of this methodology not only will improve the comparability of the results of gene set analysis, but also will allow a better evaluation of related polymorphisms both in the same and in different but functionally related genes. This step potentially would increase the power to detect causal pathways and disease mechanisms, using SNPs with significant associations and those in linkage disequilibrium (LD) with functional variants. Moreover, it can point towards integration of omics data, where the additional molecular information could verify or predict the functional effects of the associating SNP [31].

We identified two shortcomings concerning pathway analysis for GWAS data: statistical methods for genetic variation analysis have not been combined commonly in pathway analysis tools, and visualization of genetic association data such as GWAS is missing in pathways analysis. Regarding the first shortcoming, Wang and colleagues [4] were among the first to publish a pathway-based GWAS analysis using a statistical method adapted for genetic variation data. The authors modified a Gene Set Enrichment Analysis (GSEA) algorithm, initially designed for pathway analysis of gene expression data [15]. Since the adaptation of GSEA by Wang, researchers have developed other statistical methods for pathway-focused analysis of associating SNPs. Currently, existing methodologies for the analysis of GWAS gene sets are based on over-representation analysis, enrichment analysis, functional class score, and pathway-topology [3, 5, 33]. The recommendation is to apply multiple methods to capture different genetic effects and identify robust gene set associations [33]. However, only a few of these new algorithms were implemented in user-friendly tools, possibly because pathway-based approaches still have many technical challenges to overcome [31].

With regard to the second shortcoming, we believe that visualization enhances interpretation of scientific data, understanding the conclusions drawn, and discussing follow-

up research questions [30]. Thus, interpretation of GWAS data would benefit from pathway-based approaches accepting of genetic variation so that allele-specific relationships are displayed. For example, one allele of a pathway entity might allow the bioprocess to continue while a second allele curtails pathway flux. Thus, interpretation of GWAS data would benefit from pathway-based approaches accepting of genetic variation so that allele-specific relationships are displayed. Then, visualising on a pathway map the effect of variants associated with elevated risk of disease, can indicate biological and biochemical insufficiencies (and/or vulnerabilities), which then can be made more informative if placed within depictions of the affected cell or organ. Lastly, there is epistasis, where two alleles mapping to different loci associate in concert with a phenotype, but where those two alleles individually show no phenotype association [13, 40]. Epistasis or a gene-gene interaction is yet another manner in which connections within a pathway are different in different individuals. Consider, for example, that pathway endpoints are a phenotype, clinical indicator of health or disease status, or disease itself. Then, the epistatic relationships can be indicated by epistatic- or e-edges that serve to connect distinct pathways or different nodes within a single pathway in this conditional relationship. The pathways linked by such e-edges would give support to co-function and/or co-regulation with regard to the given phenotype of interest. In addition, the nodes within the GWAS-identified pathways, i.e. the main effect associations, can be used to focus the genetic landscape in the search for epistatic relationships as opposed to searching for epistasis across the entire genome.

However, genetic variants currently cannot be combined easily in pathway representations because it is not clear how to visualise and interpret variation data once connected programmatically to pathway content. In this review we sought to investigate the attempts to find solutions for the second shortcoming: visualising genetic variations in a pathway context. First, we performed a systematic review of articles that analysed genetic variants using pathway based methods in order to identify and describe the visualization options of the tools resulting from this literature review. Secondly, we performed a use case in the tools identified, testing a computational derived epistatic dataset of gene-gene interactions for twelve candidate genes in obesity risk, in order to evaluate how genetic variant analysis of epistasis is tackled by the tools. Taking a visualisation point of view, we report the features and the potential of the different software. Reviewing the articles, we also collected current statistical methodologies that have been applied in pathway-based analysis of GWAS data, and we report those without discussing in detail.

Methods and Materials

This review follows criteria developed by the PRISMA statement (Moher et al., 2009).

Search strategy

In order to assemble an overview of visualization approaches used in studies that applied pathway-based analysis to genetic association studies fully reflecting current practices, a keyword search for Pathway Analysis in PubMed and Medline (July 2014) was conducted. The literature research was performed using EndNote X7. The search yielded 2,231 articles from January 2005 through August 2014, 2,184 remained after removing duplicates, 15 others were added based on suggestions by experts in the field. Subsequently, these articles were screened manually by reading title and abstract. We retained only those 264 articles describing pathway-based analysis with genetic variation, and these articles were studied in detail. Retaining the 65 most relevant papers, all from 2007 through 2014, we aggregated the results with key features of the analysis, summarized in Table 1 in the supplemental material. In order to update the manuscript with additional visualization tool for GWAS pathway analysis, we performed a second PubMed search in January 2017 using the keyword Pathway Analysis for title and abstract, and date of publication from August 2014 to present. We obtained 2,774 articles that were scanned by title. Several articles describing GWAS pathway analysis tools were found, but only one [1] presented visualization features. This one was included and described in the tool paragraph, and reported in Table 2.1 together with the other four tools previously identified. Details of the 65 relevant articles selected with the literature search are given in Table 1 of the supplemental material. Columns describe specific features extracted from each study: type of data and variants, algorithm used, and bioinformatics tools used with visualizations. Because we did not select the articles based on the type of variants utilised, but on the type of analysis performed (keyword used: Pathway Analysis), we also identified articles where the variants participating in the genotype-phenotype association originated from sources other than SNParrays. In the 65 articles: 57 were based only on GWAS data, 4 on GWAS plus expression data, 1 on GWAS plus epigenetic data, 2 used known somatic mutations, and 1 using Next Generation Sequencing data. In all studies the resulting SNPs were investigated using pathway-based analysis, and only 3 studies also analysed copy number variants and/or indels [6, 14, 35].

Table 2.1 Summary of the main features of the pathway-based analysis tools evaluated.

Features	Caleydo	IPA	MetaCore	PathVisio	Path
Availability	Free download	Private	Private	Free download	Free download
Type of genetic variants data	CNVs	SNPs	SNPs	SNPs	SNPs
Variants data format	.cvs, .txt, .gct	.xsl, xslx, .txt	VCF	.csv .txt	LINKAGE pre-mapped, QTDT
Pathway collections	KEGG, Wikipathways	Private collection	Private collection	Wikipathways	KEGG
Applications for pathway-analysis	Enroute, Entourage	Enrichment Analysis	Enrichtment Analysis Workflow	Enrichment Analysis	UNPHASED
Gene description	Present	Present	Present	Link to the gene database	Present
Variants data visualized on pathway	YES	YES	YES	YES	not known because of the bug
Variants description	Not present	Not present	Not present	Links to the variants database	not known because of the bug
Linkage Disequilibrium map	Not present	Not present	Not present	Not present	Present
Presence of other omics data	YES	YES	YES	YES	NO

Overview of pathway analysis tools for genetic variation data

Although some algorithms are available as web services or installable software, no generally accepted implementation for visualising results exists. From the literature search we found the following bioinformatics tools able to visualise the significant variants in a pathway: IPATM of QIAGENs Ingenuity Pathway Analysis (QIAGEN Redwood City, `http://www.qiagen.com/ingenuity`) [28, 42, 43], MetaCoreTM from Thomson Reuters (`http://www.thomsonreuters.com/metacore`) [16], Path (`http://www.genapha.icapture.ubc.ca/`) [18, 41], and Pathvisio 3 [1]. In general, very few tools support pathway visualization of genetic variants. In addition, the Gehlenborg et al. (2010) [27] review mentions a visualization tool not found in the articles reviewed. This tool is called Caleydo (`http://www.caleydo.org/`) and it depicts only CNVs. We describe in the Results section the five tools mentioned here with a specific focus on the visualization options for the genetic variants. However, some relevant command line tools were also detected in the literature search, but we do not describe these because of the absence of user-friendly visualization features.
We also evaluate three of the five tools selected from the literature search, using an available epistatic dataset [40]. Because the tools do not only require different formats, but also have different features, we could not use this dataset for Caleydo and Path. For these tools the evaluation of the visualization was assessed using the default dataset provided by the software and the tutorials.

Dataset of epistatic interaction

An epistatic dataset from De et al. (2015) [40] is chosen to evaluate the SNP visualisation in the biological pathways of three tools retrieved from a literature search: IPA, MetaCore and PathVisio. The dataset consists of a list of SNPs with significant epistasis interactions (SNP-SNP connections) calculated from a gene-gene interaction epistasis network of twelve candidate genes for obesity risk (*BDNF, ETV5, FAIM2, FTO, GNPDA2, KTCD15, MC4R, MTCH2, NEGR1, SEC16B, SH2B1, TMEM18*). SNPs were extracted from the twelve genes following specific criteria: 500kb window around the gene UTR, exclusion of SNP with minor frequency allele below 0.05, exclusion of SNP that shows linkage disequilibrium of r2 above 0.8, and imputation of missing genotypes. The resulting SNP dataset in the study was 1,191 SNPs with genotype data available for 1,141 obese individuals (Body Mass Index above 30 kg/m2). A Statistical Epistasis Network [36] was utilized to characterize the interactions between genetic variants from the twelve obesity genes, resulting in a list of 58 SNPs with significant mutual information. This value is a measure of the independent or main effect of a SNP on a phenotype. It can be used to study the interaction effect between pairs of SNPs and the degree to which phenotypic variance can be understood when both genotypes are

combined. We used the 58 SNPs as input to the three tools selected for the visualisation evaluation. Describing the advantages and disadvantages of the tool features, we try to understand which tool can facilitate the interpretation of the SNPs in the pathway context.

Results

Pathway-based analysis tools with visualization options

The evaluation of five pathway-based analysis tools Caleydo, IPA, MetaCore, Path and PathVisio that support incorporation of genetic association data demonstrates: first, how polymorphism data can be visualised and analysed in a pathway-based environment, and second, how different information and experimental data can be combined for analysis and visualization. The purposes of evaluation relate directly to the need to combine GWAS results with biological context in order to better understand results in a disease context. Pathway content provides the biological processes in which GWAS-identified genes are known to be involved and shows other genes related by common function that may not pass GWAS significance thresholds. Integration of other types of genomics data as accepted by these tools, often in combination with bioinformatic pipelines for data processing, permit evaluation of different transcriptomics outcomes in subjects with a specific genotype or phenotype, and some tools allow also integration of metabolomics results.
The five tools are designed to visualise the data on different pathway collections originating from different databases. Path refers to KEGG (`http://www.genome.jp/kegg`) [34], PathVisio to WikiPathways (`http://www.wikipathways.org/`) [1] and Reactome [7], Caleydo to both KEGG and WikiPathways; while MetaCore and IPA use their respective curated pathway collections.

Tool-specific visualization details

MetaCore is a software suite suitable for functional analysis of different omics data, including expression data and genetic variation data. One of MetaCores relevant applications for pathway analysis is the Enrichment Analysis Workflow, which calculates enrichment p-values in different types of gene sets within the uploaded dataset. These gene sets originate from curated pathways, networks of related genes derived primarily from literature evaluation and from the Gene Ontology lexicon. We performed an example analysis using the 58 SNPs with significant epistasis interactions as input. As the tool accepts variants in a VCF file, we formatted the input data accordingly. The results of this analysis recognized 13 objects, limited to just one SNP per gene. Differ-

ent outputs such as pathway maps, gene ontology (GO) processes, process networks, and diseases (symbolized by biomarkers) are listed as part of the result. All list items are clickable, allowing more detailed visualization of the different items. The resulting pathway maps are ordered by enrichment p-value, with false discovery rate (FDR) corrections. The FDR calculation considers the p-value of each network map and its rank given the total number of maps in the entire set of pathway maps. The list also contains the ratio of significant genes in the dataset over the number of genes in the pathway. If one pathway in the list is selected, a pathway map is displayed. In our example the first pathway of the list is Retinal ganglion cell damage in glaucoma in which two genes appear to be colored bright and illustrated that they present the input SNPs with a red colored bar. Clicking a gene symbol displays detailed information about the description of the gene and encoded protein for human, mouse and rat. Clicking the red bar yields details for the uploaded data of that gene, in this case the SNP rs ID. In the example pathway two genes show data: BDNF with rs10835210 and ASIC1 with rs1108923. It is remarkable to notice that ASIC1 is not in the list of the twelve obesity genes of the study selected. Indeed, the SNPs from the obesity-epistasis dataset [40] were extracted taking into account a window of 500kb from the obesity genes, but MetaCore assigned SNPs only positioned within a gene region. This is also the reason why the total SNPs identified by the analysis is thirteen and not twelve. In this case rs1108923 is selected in the dataset because it maps to the upstream region of the obesity gene FAIM2, but the tool considers this variant to be within the region of ASIC1.

QIAGENs Ingenuity Pathway Analysis, IPA is a web-based application for data analysis in pathway context. Although the IPA environment is amenable to different types of analysis (i.e. Metabolomics, microRNA, Toxicology, etc.), our objective is to highlight aspects of pathway analysis. After uploading the list of 58 SNPs with the significant epistasis interactions value, the program automatically displays an overview page with information such as the number of SNPs recognized by the tool, in this case 22 SNPs of 58 were mapped. In addition, a table is shown with Entrez gene IDs and affiliated information such as cellular location, type of gene, and interacting drug. Clicking on one of the gene names listed, it displays a link to a description gene page for human, mouse and rat, in which additional information about the gene functionality are provided. In this overview page there is a possibility to perform different analysis as was mentioned above. We opted to the Core Analysis that includes the enrichment pathway analysis. However, such analysis takes into account the genes in which the 22 SNPs were mapped and not the SNPs themselves. The result page, as in MetaCore, lists several output such as: canonical pathways, diseases and function, regulators, and networks. The canonical pathway visualization is a list of enriched pathways ranked by p-value and percentage of the overlapping genes mapped against the total number of those in that pathway. Selecting a pathway prompts IPA to offer several views that depict different items within

the top significant pathways such as bar charts, and stacked bar charts. The pathway visualization is displayed under the network tab, where genes with different colors and shapes are shown as clickable nodes that link with additional information related to that gene, including biochemical elements, metabolites, and references curated by IPA team. At this level, further information about SNPs related to the genes is not visualised and reported.

PathVisio 3 is a pathway editor, visualization and analysis software. PathVisio core features related to visualisation are listed in a main panel where pathway diagrams can be drawn, and the entities of the pathway can be displayed in different ways according to advance data visualisation options. There is a side panel called backpage where data and other visualization features are shown. Some of these features are related to the advanced options provided by plugins. Developed by any user, these plugins are extensions of the PathVisio system that do not change its core functionalities. Two of these plugins, BiomartConnect (`https://www.pathvisio.org/plugin/biomartconnect/`) and RegInt plugin (`https://www.pathvisio.org/plugin/regint-plugin/`), add functionalities related to genetic variants. BiomartConnect enables visualization of biological information in the backpage, retrieved with the Ensembl BioMart tool (`http://www.ensembl.org/biomart/martview`), with which variants also are accessible. With this plugin the variants, stored in the Ensembl database and located in any gene selected from a pathway diagram, are visualised in the backpage. Moreover, additional SNP information like SIFT and PolyPHEN predictive scores is available and possible to display in the backpage. The RegInt plugin enables one to upload and visualise user data on the pathway, in the form of an interaction file. This file contained a data column listing the 58 SNPs and another listing the genes in which those SNPs are located. For the detailed input format check plugin instructions in Github (`https://github.com/PathVisio/RegInt-Plugin/wiki/User-Guide`). We used the RegInt plugin to display the 58 epistatic SNPs. First, in the main panel, we opened a pathway diagram presenting at least one of the genes related to the 58 SNPs from the WikiPathways collection (`www.wikipathways.org`), a pathway database linked to the software. Then, in the backpage the SNPs related to the gene selected in the pathway are displayed. The number of SNPs visualised depends on the data uploaded. In our case we selected from Wikipathways the brain-derived neurotrophic factor signaling pathway (WP2380), that presents two (BDNF and SH2B1) of the twelve genes of the epistatis dataset. From a biological prospective this type of visualization allows two types of investigation: one at the gene level where the relation between genes with significant epistatic SNPs can be explored in the pathway. The other one at the SNPs level, where the list of the epistatic SNPs is shown in the backpage and their effects can be explored further. Moreover, a SNP hyperlink that connects to a variant database in which the SNP description is

provided, is a useful feature to speed the research into SNP function.
Caleydo is an open source software with three applications for data visualization: StratomeX [39], enRoute [6], and Entourage [32]. StratomeX organizes different data from cancer patients, and retrieves disease information from TCGA datasets (`http://www.cancergenome.nih.gov/`). Packages that are of interest for pathway analysis are the Entourage view, which investigates interdependencies between pathways, and the enRoute view, which analyses experimental data in pathway context. The Entourage view compares pathway maps selected from the same or different pathway collections. A notable aspect is the visualization of pathway interconnectivity between selected pathways for specific genes. This useful feature enables deeper insight because it depicts how a gene or even a subpath observed in one pathway might have different roles in an interconnected process. These interconnections are intuitively displayed with colored lines that connect the selected subset of genes or single genes from the main pathway to their occurrence in other pathways. Lastly, enRoute allows selection of a subset of genes in a pathway, and these selected genes can be associated with experimental data from TCGA in which CNVs also are shown. Caleydo provides this type of visualization and analysis only for a specific set of experimental data (i.e. TCGA dataset), and for this reason it was not possible to upload the list of 58 epistatic obesity SNPs for the use case.
Path is specifically designed for GWAS analysis, connects GWAS results with information retrieved from nine common bioinformatics resources (NCBI, OMIM, KEGG, UCSC Genome Browser, Seattle SNPs, PharmGKB, Genetic Association Database, dbSNP, The Innate Immune Database), and supports visualisation of the integrated data. Path uses UNPHASED [24] for statistical analysis and retrieving information on SNP-SNP associations from the different bioinformatics resources. The only pathway resource included is KEGG. Visualizations mainly consist of charts, plots and summary tables that list genes, SNPs, SNP associations and gene-gene interactions. Importantly, Path is specifically directed towards GWAS studies, showing specific association results, and lists of genes, SNPs and LD plots. The pathway visualization using KEGG data shows genes with significant SNPs highlighted in red. Currently, not all the features of Path work properly due to unfixed bugs, that the authors decided to do not address at the moment. For this reason it was not possible to perform the use case with the epistatic obesity SNPs.

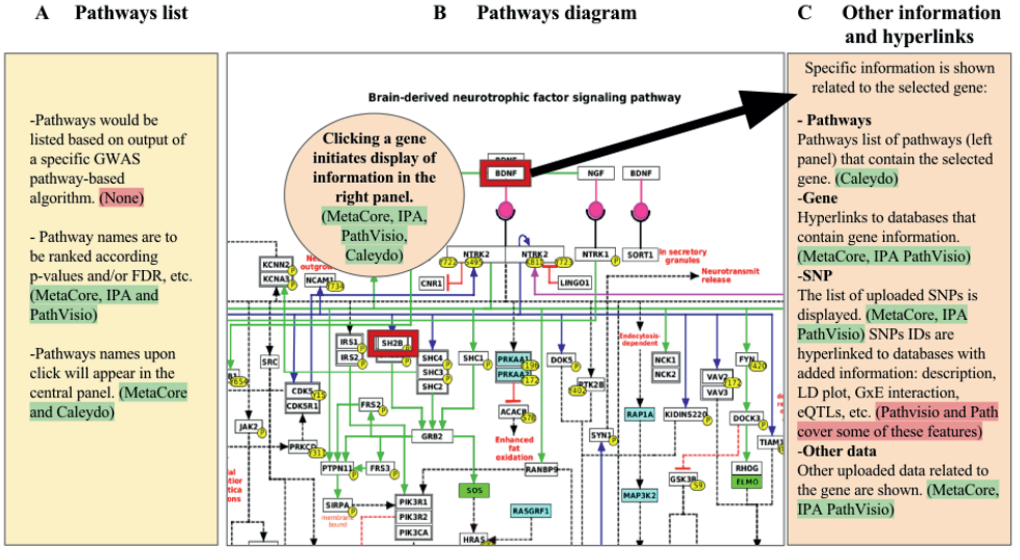

Fig. 2.1 Mock-up visualization of the combination of useful features to apply for GWAS visualization and analysis in pathway-based tools. The panels show: (A) list of pathways obtained from a specific GWAS pathway analysis algorithm; (B) pathway diagram selected from one of the pathways listed in the panel (A), where genes with GWAS hits are highlighted (red border); (C) other information with hyperlinks related to several types of data with regard to the gene selected from the panel (B), that could be displayed in expandable/collapsible lists. Highlighted green are the tools in which the specific feature described is present, red highlights indicate features that are either not present or partially present in the tools reviewed.

Statistical methods in pathway analysis tools

The GWAS variants from the 65 articles retrieved by literature search, were evaluated for pathway assignment using different algorithms that were not always well described. When they were, the authors always provided the p-value of the variant from the genotype-phenotype association [20]. The different algorithms used in the pathway-based methods aggregated SNP or gene scores to assign a p-value to a pathway. The association of a SNP to a particular gene is normally evaluated using a cutoff for SNP significance in a specific gene neighborhood region. Then, p-values assigned to each pathway can be calibrated and adjusted for some biological event such as LD patterns and co-location of functionally related genes. Such biological events can be evaluated differently by different algorithms, which can affect the results and suggest other conclusions. Researchers have developed different statistical methods for analysis of associating SNPs (Table 1, supplemental material). Approaches include LD calibration and identification of associated pathways [26], and comparison of different algorithms, which revealed advantages and disadvantages of the statistics used for a specific GWAS dataset [21, 23, 37, 38]. These articles compare different statistical methods tested in GWAS datasets, evaluating the lists of enriched pathways. Although not all algorithms listed in Table 1 of the supplemental material have been compared, we reported the conclusive judgment of the comparison performed in certain studies. Some of the most sensible statistical methods include the adaptive rank truncated product (ARTP) [38], the modified summary statistic (mSUMSTAT) [26], and the raw data-based algorithms implemented in PLINK (PLINK set-based test) [21, 37]. These algorithms were shown to be the most powerful for detecting genes that could be used further by pathway analysis tools [21, 26, 38]. It is difficult to make a single and objective preference of one specific method because results of pathway-based analysis for GWAS data vary by method. Even the overlap of shared pathways can be quite limited because each algorithm has its own evaluation focus on disease associations [37], and some examples concern different calculations of values, including pathway p-values in ARTP, or the mean value of a gene with the significant SNP in mSUMSTAT.

From the tools analysed, MetaCore, IPA and PathVisio present a statistical analysis of the data provided. Instead Caleydo and Path provide only data visualization on pathway graph and not statistical methods for pathway analysis. MetaCore, PathVisio and IPA perform pathway analysis in an automated fashion. The first tool uses an over-representation method on the gene list annotated from the variants present in the Variant Calling Format (VCF) provided as input. MetaCore employs a hyper-geometric model to determine the significance of the enrichments. PathVisio also uses an over-representation analysis and it is based on methods adopted in the MAPPFinder tool [19] with settings designed for gene expression data. Finally, IPA utilizes a method

for combining p-values. In the over-representation test, an association for each gene in the dataset is first calculated, then a threshold is used to determine which genes are significantly associated. The proportion of significantly associated genes within a target pathway is compared to the proportion of significantly associated genes among all genes outside the target pathway.
Alternatively, in the method applied in IPA, a p-value associated with a pathway is calculated using the right-tailed Fisher Exact Test. This p-value measures the likelihood that the association between a set of genes with a significant SNP identified by GWAS and a pathway arose by chance. In this method, the p-value for a given process annotation is calculated by considering (i) the number of genes with a significant SNP that participate in that process and (ii) the total number of genes that are known to be assigned to that process in the selected reference set. Further details on how IPA identifies pathways reaching significance were not provided (IPA webpage, 23rd June 2016, date last access).

Discussion

Overview of the comparison: benefits and limitations of the tools

Comparing the five tools described above makes evident that each uses different interactive ways to combine experimental data with information about genes, metabolites and pathway relationships (Table2.11). A mock visualization of the beneficial and applicable features observed in the different tools (green highlight), and the new characteristics that enhance the visualization and analysis of GWAS data in pathway-based analysis tools is shown in Figure 2.1. The five investigated investigated tools share some similar and effective visualization approaches, such as depicting significant pathways that contain genes in the analysed data by list view. These lists are generally ranked by enrichment ratios, p-values or FDR scores. Another common and useful strategy is to highlight genes for which pathway data are uploaded by the user, with an option to uncover gene details via hyperlinks.effective visualization approaches, such as depicting significant pathways that contain genes in the analysed data by list view. These lists are generally ranked by enrichment ratios, p-values or FDR scores. Another common and useful strategy is to highlight genes for which pathway data are uploaded by the user, with an option to uncover gene details via hyperlinks.
A general problem in pathway-based visualizations is the efficient display of information about genes that appear in multiple pathways and thereby interconnect those pathways. Caleydo offers an attractive solution in allowing interactive and automatic visualization of subpathways of genes present in other pathways. Caleydo uses this subpathway approach to indicate when the dataset has information about genes in a given

pathway. This demonstrates how experimental data can be combined with different types of knowledge about gene relationships and permits an increased understanding of experimental results that might act in concert. Caleydo provides this type of visualization and analysis only for a specific set of experimental data (i.e. TGCA dataset). It would be a large improvement if this same approach were used to automatically select the relevant genes in the pathways based on the GWAS statistical parameters such as SNP p-value or effect size beta, which in turn could offer an assessment of allele effects on pathway output, or other omics datasets.
A strength of PathVisio, on the other hand, is its enabling of this feature to permit visualisation of variants in pathways sourced either from a public repository like Ensembl or from user data. However, PathVisio lacks the interactive visualization that links entities of different pathways, as it described in Caleydo. In this context MetaCore depicts related experimental effects of genes known to be connected via membership in a pathway, protein-protein interactions, co-citation, or co-expression in other experimental datasets with network visualization. MetaCores network settings can be used to view or hide specific interaction mechanisms, such as binding, influence on expression, phosphorylation, or cleavage. IPAs approach is similar to that of MetaCore. After running the enrichment analysis, IPA lists the most represented processes, such as canonical pathways, networks, upstream regulators, diseases, and biological functions. In this way the user subjectively decides which information to use and how to integrate it. Finally, Path has some methods to integrate GWAS data in pathway analysis. Paths basic data visualization of pathways uses the common strategies described above, and data integration focuses specifically on genetic information and on gene-gene interactions. Paths representation also includes an LD plot, useful and important support for GWAS interpretation.

Suggested improvements for data integration in pathway-based analysis tools

As early as 2005, the importance of effective approaches to visualization was noted through interviews and observations of current work practices [29]. That report highlighted different aspects of pathway visualization, and suggested future developments to improve the researchers job. Our comparisons indicate that most of those recommendations have been implemented. Two examples are the options to automatically search for relevant pathways containing genes from an uploaded dataset, and access to periodically updated pathway libraries. We have presented different types of visual strategies used in currently available tools that, for a specific gene set, support the connection with various kinds of pathway information including significant pathways, metabolites involved therein, and related diseases. With many different types of high-throughput data

now readily available, including gene expression, metabolomics and protein-protein interactions, methods for integrated analysis and visualization are greatly needed [25]. Visual strategies are particularly important for data from high-throughput experiments that provide information about many genes, facilitating evaluation of potential interactions between affected genes. This potentiality can speed the investigation of the SNP effect in the pathway. Indeed, highlighting the relevant items related to the research question can reduce the process of investigating pathways singly. Moreover, alternative visualizations such as pathway hierarchies and network analysis can also reduce the long list of relevant pathways resulting from a pathway analysis. However, researchers still must investigate those pathways one by one, in order to understand how a SNP influences gene function in the entire process. MetaCore and IPA are examples that use networks to visualise the data integration. However, genetic variants cannot be used readily with these methods, because the data uploaded are not completely recognized. Adding the variants option to these tools would allow the user to contextualize the function of the genetic polymorphisms on different molecular levels. In addition, when data such as SNP-SNP interactions become available, pathway tools that present a network visualization option (i.e. MetaCore and IPA) could support display of epistatic interactions from a set of SNPs located in genes that function in the same pathway. In general, some specific omics data integration methods that support inclusion of genetic variants in a pathway already exist. In this context it is suitable to mention BioXM from Biomax Informatics [22] because it semantically integrates existing knowledge such as genotype-phenotype relations or signal transduction pathways, and organizes data into structured networks that are connected with clinical and experimental data (e.g. metabolites or proteomics datasets). With regard to the pathway collection, BioXM is flexible in that, it can display any pathway data, but requires input of pathway enrichment statistics from other sources. BioXM, on the other hand, is designed for flexibility and can integrate and display a wide range of relationships between entities, including pathways and genetic variants, but linking those two has not been demonstrated with GWAS data.

New types of genetic variant interactions for pathway-based analysis tools

Additional characteristics regarding genetic variant interactions currently are rarely depicted in pathway visualizations: edgetics, gene-environment and epistatic interactions. Edgetics is a new term referring to network perturbation models focusing on specific alterations of the molecular interactions resulting from genetic variants [12]. This perturbation model might improve understanding of how mutations associating with complex diseases affect biological networks or interactome properties [9]. With network

visualization already developed in some of the presented tools, it would be exciting to see this model implemented as a new feature.

Another area in which pathway visualization of genetic associations can be improved involves gene-environment interactions (GxE), where the genotype-phenotype association exists only under certain environmental conditions. A recently published catalog of GxEs for numerous cardiometabolic phenotypes showed the wide extent under which the genotype-phenotype association can be modified by factors such as diet, exercise, sleep and many other exposures and lifestyle factors [2]. For identical traits, that study noted sparse overlap of SNPs contributing to main-effect associations from GWAS compared to those supporting GxE interactions. In such instances, the pathway edges linking the GxE gene to the phenotype obviously would be conditional, and in many examples would contain entities such as glucose, palmitic acid or linoleic acid, which are constituents of standard metabolic pathways. Finally, epistatic interactions were used here as a use case to test the visualization tool. As a result PathVisio, MetaCore and IPA are the tools that support upload of variant data, and highlight those variants in the pathways of the genes related to the uploaded SNPs. This feature aids investigation of the effect of the epistatic SNPs within the genes and their pathways. However, only PathVisio is able to provide the complete list of variants present in the uploaded data. Indeed, IPA identifies the genes related to the SNPs without showing the SNPs, and MetaCore performed a SNP-gene mapping that resulted in a selection of genes not included in the original dataset. Concerning IPA, it is notable to mention that Ingenuity developed another software specifically dedicated to variant investigation called Variant Analysis that was not detected by the review literature search, but discovered only through the Ingenuity website. In addition, the PathVisio RegInt plugin, even if it can upload the complete dataset, fails to automatically provide to the users the overview of the total pathways that present at least one of the genes with the SNPs. This feature is supported by IPA and MetaCore. The epistatic obesity use case shows that IPA, MetaCore and PathVisio have several features that permit the visualization of genetic variants in pathways. However, these features are not harmonized in one tool. On the one hand, this is a reasonable outcome because the tools were not built with the aim to analyse genetic variants. On the other hand, it is remarkable to notice that these tools already have some characteristics that, with improvements, could permit such complexities of variant analysis. In summary, such conditional relationships as epistasis, GxE interactions and edgetics will need to be considered for pathway-based visualization of association data because genome-wide approaches to identify such genetic elements are rapidly maturing [9–11, 13].

Conclusions

In conclusion, what is especially needed regarding the GWAS data visualization in pathway-based analysis tools are two important items (red highlight in Figure 2.1). One, there must be development and integration in the tools of specific statistical methods for GWAS pathway analysis (red highlight in Figure 2.1). One, there must be development and integration in the tools of specific statistical methods for GWAS pathway analysis. Two, improving strategies for combined visualization of genetic data with other omics data in a pathways context will vastly facilitate interpretation of results. For the first point, as indicated in Results and listed in Table 1 of supplemental material, some accepted statistical methods used for pathway analysis of GWAS data have been described. Our recommendation is to include at least one of these algorithms in pathway-based analysis tools that focus on GWAS data. This will enhance pathway-based analyses by increasing accuracy to detect significant pathways because of the specificity of the statistics for GWAS data. Additionally, it is necessary that results are visualised properly, and subpaths of genes with consideration of significant SNPs in the affected pathways. Next, the necessity to identify a strategy of combining genetic variants with other omics data could be addressed by permitting immediate evaluation of significant SNPs in the pathway context. While a detailed report of functional information is already provided for genes in a pathway, this needs to be extended to SNPs. Examples of SNP information that could be useful to add include: (i) incorporation of data or links to databases that contain association data from other sources, including data mined from GWAS databases, epistasis and gene-environment interactions, eQTL data, and allele-specific drug and micronutrient responses; (ii) SNP function and description; (iii) LD plot images anchored to the chromosomal region where the SNP maps. Lastly, other improvements in visualising genotype-phenotype associations will involve extending the phenotype information to co-morbidities, and data from electronic health records and public health agencies.

The main aim of this review is to give an overview of the current state of the tools that visualise GWAS data in a pathway context. We attempted to identify and describe the visualization options of the tools that resulted from a literature review in order to provide suggestions for improvements in this developing field (Figure 2.1). We also have reported the necessity to include in these tools statistical methods for the pathway-based analysis in GWAS, aiming to define features for more comprehensive pathway-based analysis tools.

Conflict of Interest statement

E Cirillo, L D Parnell, and C T Evelo state that there are no conflicts of interest and there is no goal to endorse a commercial entity. E Cirillo and C T Evelo are part of the team that developed PathVisio, one of the visualization tools evaluated in this review.

Acknowledgements

The authors thank Prof. Dr. Maurice Zeegers of the Department of Complex Genetics, Maastricht University for comments and corrections. Any opinions, findings, conclusion, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the U.S. Department of Agriculture. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The USDA is an equal opportunity provider and employer.

Author Contributions section

EC performed the analysis and wrote the paper, LP and CE revised critically the work, provided final approval with agreements on the content

References

[1] Martina Kutmon, Martijn P. van Iersel, Anwesha Bohler, Thomas Kelder, Nuno Nunes, and et al. PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLOS Computational Biology*, 11(2):e1004085, feb 2015.

[2] Laurence D Parnell, Britt A Blokker, Hassan S Dashti, Paula-Dene Nesbeth, Brittany Elle Cooper, and et al. CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits. *BioData Mining*, 7(1):21, dec 2014.

[3] Miguel A. Garcia-Campos, Jesus Espinal-Enriquez, and Hernandez-Lemus. Pathway analysis: State of the art. *Frontiers in Physiology*, 6(DEC):1–16, 2015.

[4] Kai Wang, Mingyao Li, and Maja Bucan. Pathway-based approaches for analysis of genomewide association studies. *American journal of human genetics*, 81(6):1278–83, dec 2007.

[5] Patrick Y.P. Kao, Kim Hung Leung, Lawrence W.C. Chan, Shea Ping Yip, and Maurice K.H. Yap. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochimica et Biophysica Acta - General Subjects*, 1861(2):335–353, 2017.

[6] S Hong Lee, Stephan Ripke, and Benjamin M et al. Neale. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, 45(9):984–994, sep 2013.

[7] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, and et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, jan 2018.

[8] Teri A. Manolio, Rex L. Chisholm, Brad Ozenberger, Dan M. Roden, Marc S. Williams, and et al. Implementing genomic medicine in the clinic: the future is here. *Genetics in Medicine*, 15(4):258–267, apr 2013.

[9] Florian Markowetz. How to Understand the Cell by Breaking It: Network Analysis of Gene Perturbation Screens. *PLoS Computational Biology*, 6(2):e1000655, feb 2010.

[10] Rishika De, Ting Hu, Jason H. Moore, and Diane Gilbert-Diamond. Characterizing gene-gene interactions in a statistical epistasis network of twelve candidate genes for obesity. *BioData Mining*, 8(1):45, jun 2015.

[11] Laurence D Parnell, Britt A Blokker, Hassan S Dashti, Paula-Dene Nesbeth, Brittany Elle Cooper, and et al. CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits. *BioData Mining*, 7(1):21, dec 2014.

[12] Quan Zhong, Nicolas Simonis, Qian-Ru Li, Benoit Charloteaux, Fabien Heuze, and et al. Edgetic perturbation models of human inherited disorders. *Molecular systems biology*, 5(1):321, jan 2009.

[13] Wen-Hua Wei, Gibran Hemani, and Chris S. Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733, nov 2014.

[14] Mark D M Leiserson, Jonathan V. Eldridge, Sohini Ramachandran, and Benjamin J. Raphael. Network analysis of GWAS data. *Current Opinion in Genetics and Development*, 23(6):602–610, 2013.

[15] A Subramanian, P Tamayo, VK Mootha, and S Mukherjee. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA. 102, 15 545-15. 2005.

[16] Gwan Gyu Song and Young Ho Lee. Pathway analysis of genome-wide association studies for Parkinson's disease. *Molecular Biology Reports*, 40(3):2599–2607, mar 2013.

[17] Q Zhong, N Simonis, QR Li Molecular systems ..., and undefined 2009. Edgetic perturbation models of human inherited disorders. *msb.embopress.org*.

[18] D Zamar, B Tripp, G Ellis, D Daley Bioinformatics, and undefined 2009. Path: a tool to facilitate pathway-based genetic association analysis. *academic.oup.com*.

[19] SW Doniger, N Salomonis Genome ..., and undefined 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *genomebiology.biomedcentral.com*.

[20] Kai Yu, Qizhai Li, Andrew W. Bergen, Ruth M. Pfeiffer, Philip S. Rosenberg, Neil Caporaso, and et al. Pathway analysis by adaptive combination of P -values. *Genetic Epidemiology*, 33(8):700–709, dec 2009.

[21] Hongsheng Gui, Miaoxin Li, Pak C Sham, and Stacey S Cherny. Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's Disease dataset. *BMC Research Notes*, 4(1):386, 2011.

[22] Dieter Maier, Wenzel Kalus, Martin Wolff, Susana G Kalko, Josep Roca, and et al. Knowledge management for Systems Biology a general and visually driven framework applied to translational medicine. *BMC Systems Biology*, 5(1):38, 2011.

[23] Gordon Fehringer, Geoffrey Liu, Laurent Briollais, Paul Brennan, Christopher I. Amos, and et al. Comparison of Pathway Analysis Approaches Using Lung Cancer GWAS Data Sets. *PLoS ONE*, 7(2):e31816, feb 2012.

[24] Frank Dudbridge. User guide. 2006.

[25] D Gomez-Cabrero, I Abugessaisa, and D Maier. Data integration in the era of omics: current and future challenges. 2014.

[26] OA Panagiotou, JPA Ioannidis International journal ..., and undefined 2011. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *academic.oup.com*.

[27] N Gehlenborg, SI O'donoghue, NS Baliga Nature ..., and undefined 2010. Visualization of omics data for systems biology. *nature.com*.

[28] T Inada, M Koga, H Ishiguro, Y Horiuchi Pharmacogenetics ..., and undefined 2008. Pathway-based association analysis of genome-wide screening data suggest that genes associated with the γ-aminobutyric acid receptor signaling pathway are. *journals.lww.com*.

[29] Purvi Saraiya, Chris North, and Karen Duca. Visualizing Biological Pathways: Requirements Analysis, Systems Evaluation and Research Agenda. *Information Visualization*, 4(3):191–205, sep 2005.

[30] JM Villaveces, , P Koti Advances applications ..., and undefined 2015. Tools for visualization and analysis of molecular networks, pathways, and-omics data. *ncbi.nlm.nih.gov*.

[31] K Wang, M Li, H Hakonarson Nature Reviews Genetics, and undefined 2010. Analysing biological pathways in genome-wide association studies. *nature.com*.

[32] Christian Partl, Alexander Lex, Marc Streit, Denis Kalkofen, Karl Kashofer, and Dieter Schmalstieg. enRoute: dynamic path extraction from biological pathway maps for exploring heterogeneous experimental datasets. *BMC Bioinformatics*, 14(Suppl 19):S3, 2013.

[33] Michael A. Mooney and Beth Wilmot. Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(7):517–527, oct 2015.

[34] M Kanehisa, S Goto, Y Sato, M Furumichi Nucleic acids ..., and undefined 2011. KEGG for integration and interpretation of large-scale molecular data sets. *academic.oup.com*.

[35] Sujoy Ghosh, Juan C. Vivar, Mark A. Sarzynski, Yun Ju Sung, James A. Timmons, and et al. Integrative pathway analysis of a genome-wide association study of V o _{2max} response to exercise training. *Journal of Applied Physiology*, 115(9):1343–1359, nov 2013.

[36] Ting Hu, Nicholas A Sinnott-Armstrong, Jeff W Kiralis, Angeline S Andrew, Margaret R Karagas, and Jason H Moore. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, 12(1):364, 2011.

[37] Peilin Jia, Yang Liu, and Zhongming Zhao. Integrative pathway analysis of genome-wide association studies and gene expression data in prostate cancer. *BMC Systems Biology*, 6(Suppl 3):S13, 2012.

[38] Marina Evangelou, Augusto Rendon, Willem H. Ouwehand, Lorenz Wernisch, and Frank Dudbridge. Comparison of Methods for Competitive Tests of Pathway Analysis. *PLoS ONE*, 7(7):e41018, jul 2012.

[39] A. Lex, M. Streit, H.-J. Schulz, C. Partl, D. Schmalstieg, P.J. Park, and N. Gehlenborg. StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Computer Graphics Forum*, 31(3pt3):1175–1184, jun 2012.

[40] Rishika De, Ting Hu, Jason H. Moore, and Diane Gilbert-Diamond. Characterizing gene-gene interactions in a statistical epistasis network of twelve candidate genes for obesity. *BioData Mining*, 8(1):45, jun 2015.

[41] D Daley, M Lemire, L Akhabir, M Chan-Yeung, JQ He Human Genetics, and undefined 2009. Analyses of associations with asthma in four asthma population samples from Canada and Australia. *Springer*.

[42] J Helleman, M Smid Gynecologic . . . , and undefined 2010. Pathway analysis of gene lists associated with platinum-based chemotherapy resistance in ovarian cancer: the big picture. *gynecologiconcology-online.net*.

[43] Julius S Ngwa, Alisa K Manning, Jonna L Grimsby, Chen Lu, Wei V Zhuang, and et al. Pathway analysis following association study. *BMC Proceedings*, 5(Suppl 9):S18, 2011.

CHAPTER 3

Providing gene-to-variant and variant-to-gene database identifier mappings to use with BridgeDb mapping services

Friederike Ehrhart[1,2*], Jonathan Mlius[1*], Elisa Cirillo[1], Martina Kutmon[1,3], Egon Willighagen[1], Susan L. Coort[1], Leopold G.M. Curfs[2], Chris T. Evelo[1,2,3]

1 Department of Bioinformatics - BiGCaT, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands.
2 GKC - Rett Expertise Centre, Maastricht University Medical Center, Maastricht, The Netherlands.
3 Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands.
*** These authors contributed equally**

Published in: F1000 Research, doi: 10.12688/f1000research.15708.1

Abstract

Database identifier mapping services are important to make database information interoperable. BridgeDb offers such a service. Available mapping for BridgeDb link 1. genes and gene products identifiers, 2. metabolite identifiers and InChI structure description, and 3. identifiers for biochemical reactions and interactions between multiple resources that use such IDs while the mappings are obtained from multiple sources. In this study we created BridgeDb mapping databases for selections of genes-to-variants (and variants-to-genes) based on the variants described in Ensembl. Moreover, we demonstrated the use of these mappings in different software tools like R, PathVisio, Cytoscape and a local installation using Docker. The variant mapping databases are now available from the BridgeDb mapping database repository (`http://bridgedb.org/data/gene_database/`) and updated according to the regular BridgeDb mapping update schedule.

Introduction

Many bioinformatics software tools rely on database identifier mapping, for instance for 1) recognition and mapping of identifiers used in experimental data to the corresponding identifiers present in secondary sources like pathways or ontology classes or 2) simply to combine data from different sources that use different identifiers. BridgeDb is a database identifier mapping tool that is available as a Java framework and as an installable web service [2]. Tools that integrate BridgeDb are for instance: the community curated pathway resource WikiPathways [3], the modular pathway editor and pathway analysis tool PathVisio [1], and the network tool Cytoscape used to visualize, extend and evaluate biological networks. Depending on the available mappings BridgeDb can provide the mapping between identifiers from various data sources, also when these link to different molecular levels, e.g. gene to protein. BridgeDb can also be deployed as a web service. Moreover, it is available in a semantic web version, the Identifier Mapping Service (IMS), which can be used inside the Open PHACTS platform but can also be deployed from a software container [14]. Mappings for BridgeDb are already available for gene products for many species (produced from the respective Ensembl genome annotations [10]), for metabolite identifiers (produced from HMDB [8]) and ChEBI [5]), and for reaction identifiers (produced from Rhea [9]). The BridgeDb mapping databases are linking pins between tools that support genetic variants, genes, and pathways analysis helping to visualize a complex biological context such those typical of the multifaceted (genetic) diseases. Gene-to-variant mapping was not yet available for BridgeDb. Such mappings can be especially useful to work with genetic variations, for instance when evaluating traits with a complicated genetic background like blood pressure, suscepti-

bility to heart failure, or diabetes type 2 development. Single nucleotide polymorphism (SNP) can be responsible for phenotypic variations. In extreme cases this can be the cause of rare genetic disorders. For example, several SNPs in the human *DMD* gene can be responsible of Duchenne muscular dystrophy (DMD), a severe congenital disorder which leads to severe physical impairment [20]. Since BridgeDb can stack mappings, the combination of the new gene-to-variant mapping database with the collection that was already available offers versatile mappings for variants to a large set of different human gene and gene product identifiers. The main objective of this work was to provide mappings between gene identifiers and variant identifiers in both directions. The steps needed to achieve this were: 1) select the best source for the mappings, 2) collect data from the selected source, 3) annotate the result with provenance data about the process, the source, and the source version, and 4) finally to release the new BridgeDb mapping database and integrate that in the regular BridgeDb mapping database update schedule. Target users for the resulting mappings are 1) bioinformaticians and developers, working on new approaches for data integration, if these use human genetic (variant) information; 2) members and users of ELIXIR data interoperability services, including the implementations in the tools mentioned that perform analyses based on human genetic variant data, for instance for the analysis of common multifaceted genetic diseases or in the rare disease field; and 3) researchers who access and query molecular data resulting from the analysis above.

Methods

The gene-to-variant database uses mappings between Ensembl and dbSNP [21]. The Ensembl gene-to-dbSNP variant mappings present in Ensembl were used as the source. The released database is based on Ensembl r91, dbSNP b150, and the human genome assembly GRCh38. Although Ensembl provides more genetic variation from different sources, we focused on dbSNP as this variation database is regularly updated and adjusted to the actual Ensembl genome built. We compared both sources (Ensembl and dbSNP) and made sure that Ensembl provides all dbSNP available variants. So, we are able to rely only on the Ensembl API as a source for the extraction of the data necessary for creation of this mapping database. To prevent problems introduced by the user interfaces we used database dumps for this comparison. The data dump was obtained from the Ensembl ftp server. For the first version, we used Ensembl 91, gene annotation with Gencode 27. The vcf (variant call format) file is the one relevant for our mapping. It contains the dbSNP identifier with its additional attributes and the associated Ensembl transcript identifier. By querying the Ensembl platform web service, we can access the gene identifier of the transcript. Combined, this leads to mappings between variants and genes. The size of the complete mapping database exceeded 150

Gb (for Ensembl 91), so we decided to create several different subsets: exonic variants, missense variants, protein truncating variants (PTV), PTV and missense variants, and variants with a PolyPhen score ¿0.908 indicating Probably Damaging. Other selections can be created easily on individual demand. The created database contains the link between the Ensembl gene identifiers and the dbSNP variant identifiers including a selection of attributes (MAF (minor allele frequency), chromosome, variant alleles, and chromosome position start/end). For the rare cases where a variation is associated to more than one gene, the variant is also associated to these genes in the BridgeDb database. For example, rs199773918 overlaps in the exons of two genes (ENSG00000173366 and ENSG00000239732), and in the exonic variant BridgeDb mapping both genes show up. Nevertheless, in our selection of variants it may happen that not all of them show up due to different variant effect classifications in the different genes. As an example, rs199773918 is a variant that overlaps in the following genes: *TPR* (ENSG00000047410) and *PRG4* (ENSG00000116690). This variant is a 3 prime UTR variant of *TPR* and a missense variant of *PRG4*. It can be found in both genes variant tables but due to our selection it will show up only once in the missense variant dataset.

Implementation

Database creation:

An open-source Java program to create the gene-to-variant database is available on GitHub (`https://github.com/BiGCAT-UM/BridgeDbVariantDatabase`). After downloading the vcf file form Ensembl, users create a configuration file with several parameters. Then the database creation program will parse the vcf file, retrieve additional information through the Ensembl web service and create the BridgeDb mapping database. Due to the large amount of mappings, the tool commits the mappings to the database in batches to keep the required memory low.

Operation:

The database creation workflow is depicted in Figure Database creation workflow. The gene-variant mapping database is built on the variant call format (vcf) file provided by Ensembl. After running the database creation tool, the database can be used in all the different use cases. . The vcf file can be downloaded from the Ensembl FTP. The “Homo_sapiens_incl_consequences.vcf.gz” file is used.

System requirements:

The database creation tool runs with Java and requires more memory than usually given to a Java process. We advise users to allocate 3-4GB of memory at least when running the database creation tool (-Xmx4G).

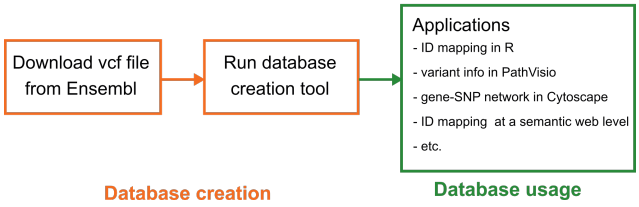

Fig. 3.1 Database creation workflow. The gene-variant mapping database is built on the variant call format (vcf) file provided by Ensembl. After running the database creation tool, the database can be used in all the different use cases.

Results

The resulting BridgeDb mapping databases are available as a Derby database from here: `http://bridgedb.org/data/gene_database/`. The new mappings are available for all the BridgeDb implementations mentioned above (PathVisio, Cytoscape, R package, web service, and the IMS). The mapping databases are freely available for download under CC-BY license. Application examples of the use of the variant BridgeDb database are given in the following section. We created gene-to-variant mapping databases for the variant classes given in Table 3.1. Any other subset of variant classes can be created on demand using the tool described in the Methods section. Any other subset of variant classes can be created on demand using the script given in the previous chapter.

Use cases

To test and demonstrate the application of the variant BridgeDb database, we downloaded the database from BridgeDb. The gene-to-variant (and variant-to-gene) queries are shown in four different tools: R command line [11], PathVisio [1], Cytoscape [4] and the local IMS installation using Docker, in order to provide an overview of the flexibility of the mapping database in different environments. A genetic variant of the rare disease Duchenne muscular dystrophy (DMD) was selected from the gene-disease association database DisGeNET [15]. The rs104894790 [13] SNP was chosen because

Table 3.1 Gene-to-variant mapping databases (status Ensembl 91, to be updated regularly)

SNP selection	File	Date	Size
Exonic variants	SNP_r91_Exon.bridge.zip	2018-06-04	1.1G
Missense variants	SNP_r91_Missence.bridge.zip	2018-06-07	620M
Protein truncating variants	SNP_r91_PTV.bridge.zip	2018-06-07	75M
Protein truncating variants and missense	SNP_r91_PTV_Missense.bridge.zip	2018-06-07	620M
All variants with a PolyPhen score above 0.9	SNP_r91_PolyPhen.bridge.zip	2018-06-07	227M

it presented a high number of citations and a stop gain damaging effect on the genes protein product.

R

The SNP, rs104894790, as described above was used to query the Ensembl identifier for the gene(s) in which it is located (variant-to-gene query). The query was performed in R command line, after the installation of the BridgeDb R package (example R script in Supplementary File 1) (R version 3.5.1). The result shows that the variant is positioned only in one gene: dystrophin (*DMD*, ENSG00000198947). *DMD* is one of the largest genes in the human DNA (about 2.2 Mb), and is composed of 79 exons and has 32 known transcripts of which 20 are protein coding. Because the output is identifiers, it can be easily linked to other R packages such as mygene [12] which normally wraps around the mygene.info web service [6].

PathVisio

We used PathVisio (version 3.3.0) (Figure 3.2), a biological pathway analysis tool that allows drawing, editing and analyzing biological pathways, to demonstrate how the new gene-variant database can be used to evaluate variants in a pathway context. PathVisio, like Cytoscape, has the BridgeDb functionality integrated in the core. For the purpose of the demonstration, we first selected pathways that contain the *DMD* gene from the R example. Five pathways were found: two striated muscle contraction pathways (WikiPathways identifiers: WP3795 and WP383), Ectoderm differentiation (WP2858), Extracellular matrix organization (WP2703) and Arrhythmogenic right ventricular cardiomyopathy (WP2118). In principle, a new PathVisio plugin could now be developed that searches pathways that contain genes with selected variants automatically, or the plugin could show all variants from an analysis sets on a given pathway. For the exam-

ple, one of the striated muscle contraction pathways (WP383) was selected and visualized. Next, the BridgeDb variant database was loaded, using the BridgeDbConfig plugin (`http://www.pathvisio.org/plugin/bridgedbconfig-plugin/`). After selecting a gene in the pathway, the backpage tab of the right hand side panel now shows the list of hyperlinks obtained from the BridgeDb database that point to different information sources linked to the gene selected. Figure 3.2 shows the backpage with the list of the 720 SNPs (from the BridgeDb with a PolyPhen ¿ above 0.908, file "SNP_r91_PolyPhen.bridge") for the selected *DMD* gene. All the SNPs in the backpage have a hyperlink to the corresponding dbSNP page.

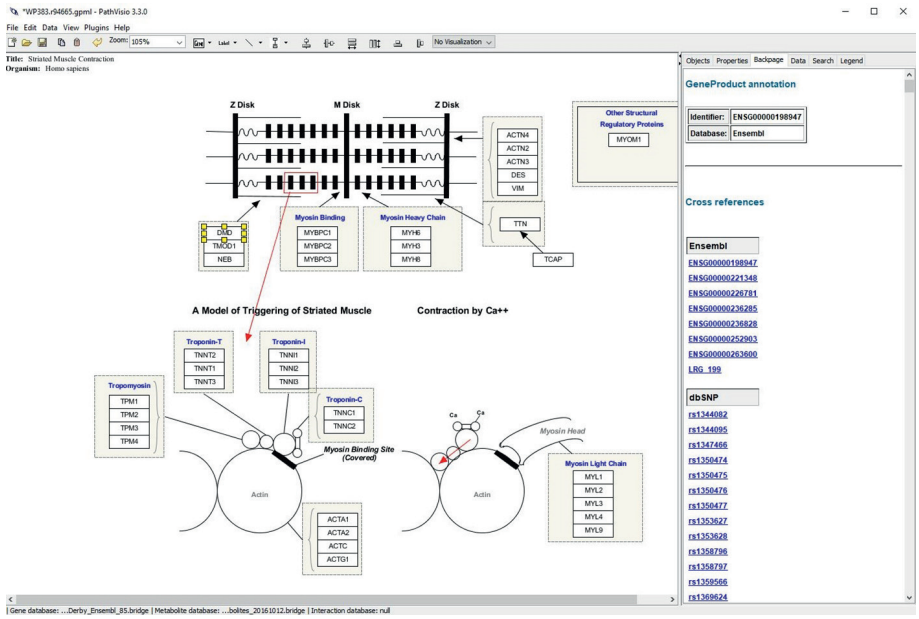

Fig. 3.2 PathVisio shows the diagram of the pathway WP383 from WikiPathways collection is shown in the left panel of the tool. When the DMD gene is selected a list of hyperlinks from different sources are displayed in the back page of the left panel. In this case a list of SNPs located in the gene are visualized.

Cytoscape

An alternative gene-to-variant visualization is provided using Cytoscape (version 3.6.1), a popular tool for (biological) network analysis and visualization (Figure 3.3). The BridgeDb app for Cytoscape is available at: `http://apps.cytoscape.org/apps/bridgedb`. A node with the Ensembl gene identifier of DMD was created and the 720 SNPs were mapped to the gene using the BridgeDb app interface. A gene-variant network was created using the list of variants mapped. Moreover, the app can be

used to configure the selection of several attribute columns related to the variant nodes such as: chromosome location, minor allele frequency, and variant allele. In this example figure, we visualize the PolyPhen score as the node fill color of the variants. For simplicity, the rs-numbers are not displayed.

BridgeDb Identifier Mapping Service (IMS)

Finally, we here show that identifier mapping linking variants to genes and vice versa can also be done at a semantic web level, we here demonstrate how an online BridgeDb Identifier Mapping Service (IMS) can be set up. The IMS technology was developed in the Open PHACTS project to link drug discovery related data sets, including a Docker image [2, 16, 17]. Here, identifier mappings are defined by link sets, which specify which identifiers are mapped. However, unlike traditional BridgeDb mapping files, these link sets also specify why the two identifiers are mapped, allowing them to be used as scientific lenses [17]. Because the IMS works at a semantic web level, identifiers are represented by uniform resource identifiers (URIs). Moreover, the IMS is aware of URI equivalence defined by the MIRIAM registry [18]. This means that even when a mapping file does not provide mappings for a certain URI, one would still get a number of equivalent URIs, following knowledge from MIRIAM database. And, when a single mapping is found in the link sets, equivalent URIs for the mapped URIs it returned. The IMS provide a targetUriPattern parameter allowing you to restrict the number of mapped URIs. We developed a tutorial explaining how to set up an IMS instance with the variant-gene mappings (available from GitHub). The instance is run locally using a Docker container developed by Open PHACTS, which is available from DockerHub. After the Docker image is started, it provides a web interface and an API. The web interface has a ”Check Mapping for an URI” page where the URI can be given to be mapped, the return format (XML, JSON, or HTML), and optionally a lensUri (see [17]), and the aforementioned targetUriPattern. However, it is more convenient to use this API from other tools, as demonstrated with a second R script (Supplementary File 1). This R script uses the curl [19] and jsonlite [7] packages to interact with the IMS. The first package is used to call the IMS webservice and the second to convert the returned JSON into a data model more easily handled in R. The example consists of two API calls: the first part finds 603 variants for the *DMD* gene (Ensembl ID ENSG00000198947); the second example takes a single variant (dbSNP ID rs769658853) and looks up the matching gene.

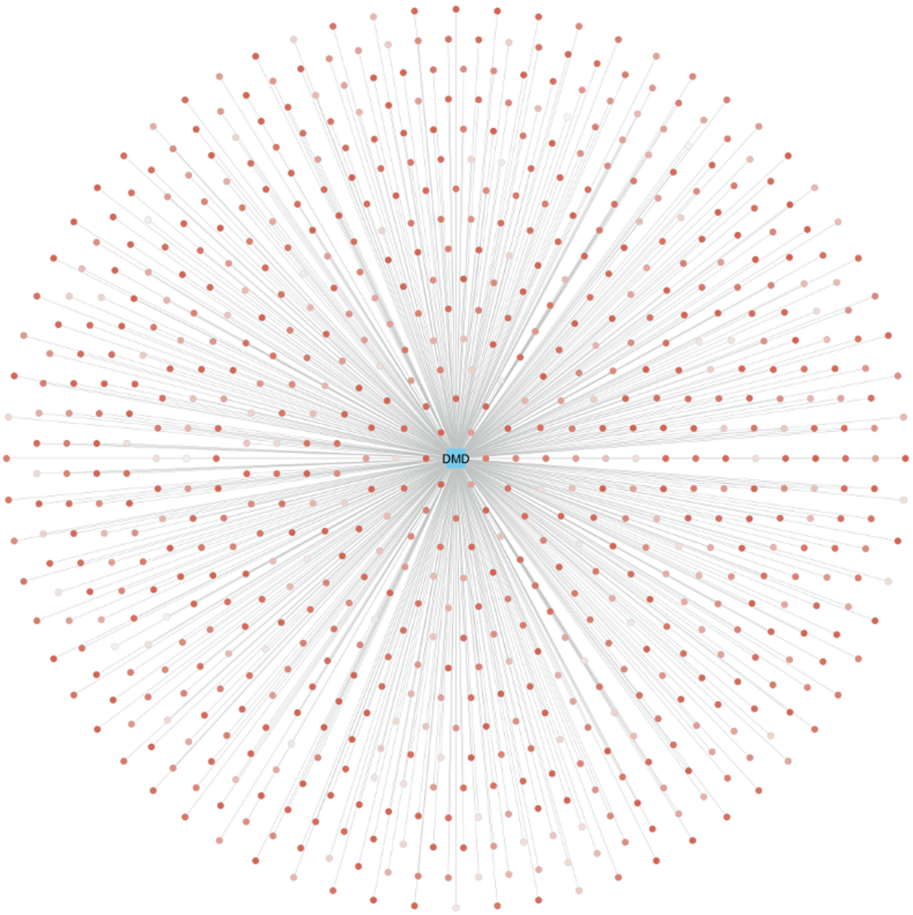

Fig. 3.3 Cytoscape displays the gene-variants network in which the red node is the DMD gene and the blue nodes are the 720 variants with a PhenScore ¿ 0.9 (pathogenic) known for this the gene. Below the main screen there is a table that shows the nodes ID and the attributes retrieved from the BridgeDb mapping.

Discussion

The BridgeDb toolset provides several apps and tools designed for different purposes, while mapping databases are available to link different database IDs for genes and gene products, metabolites, and reactions and interactions. A mapping database in the BridgeDb software environment, capable of linking genes to their variants and vice versa, was not yet available. The new database is expected to be useful to enhance the biological interpretation of genetic variant data (as shown with the example of the *DMD* gene) for instance when using apps that evaluate biological pathways, use the

classification of genes according to ontology terms, or in the R environment when performing gene and variant related statistical evaluation. With this newly created mapping database and the transitivity function of BridgeDb, the user can map between three different layers: e.g. variant-gene-protein. This approach can support multi-omics analysis for various biomedical applications, and tools like Cytoscape and PathVisio can be used immediately to benefit from this. We intend to keep the content up-to-date by regular updates. The human variant mapping database is already incorporated into the quarterly BridgeDb mapping database update. Also other variant sets including more than only the currently included protein truncating and missense variants can be created on user community (or individual) demand.

Data availability

The new gene-to-variant mapping databases are available here: `http://bridgedb.org/data/gene_database/` License:Apache 2.0 licence (`http://www.apache.org/licenses/LICENSE-2.0.html`)

Software availability

Source code for making of the mapping databases:`https://github.com/BiGCAT-UM/BridgeDbVariantDatabaseR-package` for BridgeDb is available here: `https://github.com/BiGCAT-UM/bridgedb-r`

Author contributions

JM created the BridgeDb mapping database, FE study coordination, rare disease examples and testing, EC tested the BridgeDb mapping database on the R command line, PathVisio and Cytoscape examples, EW developed the R use cases, MK contributed in choice of source database and selection of SNPs, SM critical review, LC critical review and rare disease expertise, CE implementation and testing plan and coordination, all authors contributed to writing.

Competing interests

This work was funded by ELIXIR, the European research infrastructure for life-science data.

Grant information

The authors declare no competing interests.

Acknowledgement

The authors would like to thank the BridgeDb development team. This work heavily leaned on previous work done by the dbNP and Ensembl teams who curated the actual mappings and on the original BridgeDb development team, especially Martijn van Iersel.

References

[1] Martina Kutmon, Martijn P. van Iersel, Anwesha Bohler, Thomas Kelder, Nuno Nunes, and et al. PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLOS Computational Biology*, 11(2):e1004085, feb 2015.

[2] Martijn P van Iersel, Alexander R Pico, Thomas Kelder, Jianjiong Gao, Isaac Ho, and et al. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1):5, jan 2010.

[3] Denise N. Slenter, Martina Kutmon, Kristina Hanspers, and et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, (November):1–7, 2017.

[4] Paul Shannon, Andrew Markiel, 2 Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, and et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, (13):2498–2504, 2003.

[5] Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, and et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41(D1):D456–D463, nov 2012.

[6] Jiwen Xin, Adam Mark, Cyrus Afrasiabi, Ginger Tsueng, Moritz Juchler, and et al. High-performance web services for querying gene and variant annotation. *Genome Biology*, 17(1):91, dec 2016.

[7] Jeroen Ooms. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. mar 2014.

[8] David S. Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, and et al. HMDB 3.0The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807, nov 2012.

[9] Anne Morgat, Thierry Lombardot, Kristian B. Axelsen, Lucila Aimo, Anne Niknejad, and et al. Updates in Rhea an expert curated resource of biochemical reactions. *Nucleic Acids Research*, 45(D1):D415–D418, jan 2017.

[10] Bronwen L. Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, and et al. The Ensembl gene annotation system. *Database*, 2016:baw093, jun 2016.

[11] R Core Team. R: A language and environment for statistical computing.

[12] A. et al. Mark. mygene: Access MyGene.Info_ services. R package version 1.12.0.

[13] U Lenk, R Hanke, U Kräft, K Grade, I Grunewald, and A Speer. Non-isotopic analysis of single strand conformation polymorphism (SSCP) in the exon 13 region of the human dystrophin gene. *Journal of medical genetics*, 30(11):951–4, nov 1993.

[14] Alasdair J.G. Gray, Paul Groth, Antonis Loizou, Sune Askjaer, Christian Brenninkmeijer, and et al. Applying linked data approaches to pharmacology: Architectural decisions and implementation. *Semantic Web*, 5(2):101–113, jan 2014.

[15] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, and et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839, jan 2017.

[16] Antony J. Williams, Lee Harland, Paul Groth, Stephen Pettifer, Christine Chichester, and et al. Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today*, 17(21-22):1188–1198, nov 2012.

[17] Colin Batchelor, Christian Y. A. Brenninkmeijer, Christine Chichester, Mark Davies, Daniela Digles, and et al. Scientific Lenses to Support Multiple Views over Linked Chemistry Data. pages 98–113. 2014.

[18] N. Juty, N. Le Novere, and C. Laibe. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Research*, 40(D1):D580–D586, jan 2012.

[19] Jeroen Ooms. curl: A Modern and Flexible Web Client for R. 2017.

[20] Francesca Magri, Alessandra Govoni, Maria Grazia D'Angelo, Roberto Del Bo, Serena Ghezzi, and et al. Genotype and phenotype characterization in a large dystrophinopathic cohort with extended follow-up. *Journal of Neurology*, 258(9):1610–1623, sep 2011.

[21] Adrienne Kitts, Lon Phan, Ward Minghong, and John Bradley Holmes. *The database of short genetic variation (dbSNP).* Number 2nd. Bethesda (MD), 2nd edition, 2013.

CHAPTER 4

From SNPs to Pathways: Biological interpretation of Type 2 Diabetes (T2DM) Genome Wide Association Study (GWAS) results

Elisa Cirillo[1*], Martina Kutmon[1,2], Manuel Gonzalez Hernandez[1], Tom Hooimeijer[1], Michiel E Adriaens[2], Lars MT Eijssen[1], Laurence D. Parnell[3], Susan L. Coort[1], Chris T. Evelo[1,2]

1 Department of Bioinformatics - BiGCaT, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands.
2 Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands
3 Agricultural Research Service, USDA, Jean Mayer-USDA Human Nutrition Research Center on Aging at Tufts University, Boston, MA, USA.

Published in: Plos One, doi: 10.1371/journal.pone.0193515.

Abstract

Genome-wide association studies (GWAS) have become a common method for discovery of gene-disease relationships, in particular for complex diseases like Type 2 Diabetes Mellitus (T2DM). The experience with GWAS analysis has revealed that the genetic risk for complex diseases involves cumulative, small effects of many genes and only some genes with a moderate effect. In order to explore the complexity of the relationships between T2DM genes and their potential function at the process level as effected by polymorphism effects, a secondary analysis of a GWAS meta-analysis is presented. Network analysis, pathway information and integration of different types of biological information such as eQTLs and gene-environment interactions are used to elucidate the biological context of the genetic variants and to perform an analysis based on data visualization. We selected a T2DM dataset from a GWAS meta-analysis, and extracted 1,971 SNPs associated with T2DM. We mapped 580 SNPs to 360 genes, and then selected 460 pathways containing these genes from the curated collection of WikiPathways. We then created and analyzed SNP-gene and SNP-gene-pathway network modules in Cytoscape. A focus on genes with robust connections to pathways permitted identification of many T2DM pertinent pathways. However, numerous genes lack literature evidence of association with T2DM. We also speculate on the genes in specific network structures obtained in the SNP-gene network, such as gene-SNP-gene modules. Finally, we selected genes relevant to T2DM from our SNP-gene-pathway network, using different sources that reveal gene-environment interactions and eQTLs. We confirmed functions relevant to T2DM for many genes and have identified some - *LPL* and *APOB* - that require further validation to clarify their involvement in T2DM.

Introduction

GWAS and pathway analysis

Since 2005 analysis of genetic variations in complex diseases has been conducted with genome-wide association studies (GWAS) [49]. Such an analysis consists of genotyping the genomic DNA of individuals divided into case and control groups according to a specific trait or phenotype. A genome scan is performed using a set of genetic variation probes of at least 100,000 single nucleotide polymorphisms (SNPs) to a million or more, and more recently genomic sequencing-based approaches to detect SNPs have been added [1]. Thereafter, computational methods are applied to SNPs related to the investigated phenotype, resulting in a list of SNPs significantly associated with the phenotype. Despite the limitations of such studies [11], GWAS is still a valuable method that provides insights to delineate the molecular scenario of complex diseases like type

2 diabetes mellitus (T2DM) and to support risk prediction [28].

Nevertheless, it remains challenging to perform secondary analysis on GWAS results with the aim of obtaining a detailed biological understanding of the SNPs function and role in a disease [42]. Pathway analysis is an example of secondary analysis that has been applied to GWAS since 2007 [4], where the SNPs are contextualized in biological processes through the genes to which they are assigned [13]. Garcia-Campos et al. (2015) and Kai et al.(2015) [3, 5] published reviews describing how to use different pathway analysis methodologies, which are applicable to GWAS data. Three basic steps are performed in these pathway analysis methods: (1) gene set are chosen by the user for instance from Gene Ontology annotations [17], KEGG or WikiPathways pathways [16], (2) genetic variants are mapped onto the genes, and (3) gene set statistics are performed. There are two different approaches for the gene set statistics: in a one-step approach, gene set p-values are calculated directly from genotype data, whereas in a two-step approach first single gene p-values are determined, from which the final gene set p-values are computed.

Another way to discriminate gene set statistical methods is by the difference in the hypothesis tested. The hypothesis tested is either whether the observed pathway is associated with the phenotype (often referred to as a self-contained approach or an association method), or whether the genes within a pathway are significantly enriched in comparison of other genes (referred to as competitive approach or enrichment method). In both cases the output is a list of pathways ranked by their significance based on the statistical test performed. Gene set enrichment was used to obtain interesting and germane pathway results linked to diabetes [18, 26]. However, looking only at the highest ranked pathways does not assure an accounting of all genes detected by the significant association signals, interpretation of which could be relevant to understand the phenotype. In general, it is often suggested that the output list needs to be checked manually, pathway by pathway and gene by gene. It then becomes time-consuming and error prone to account for all the possible relations that different pathways and genes present between each other.

We propose an approach based on network analysis and visualization where we display biological pathways identified by the presence of genes in which the SNPs from T2DM GWAS meta-analysis are mapped. The list of pathways is derived simply from the fact that one or more genes associated with a significant GWAS SNP signal are present in that pathway. This allows us to create a SNP-gene-pathway network that includes all the pathways where a significant SNP was found.

Furthermore, in recent years development of methods for testing hypotheses about the molecular mechanisms of a phenotype from the GWAS results, has promoted several secondary approaches besides those related to pathway-based analysis [41]. An example is expression quantitative trait loci (eQTL) analysis that can enhance the characteri-

zation of GWAS variants, in particular the non-coding ones. eQTLs are loci that contain sequence variants that are found to affect the expression of genes. They are identified by relating gene expression measurements to genotyping information in panels of individuals [6]. eQTL databases, such as GTEx portal [12], provide the opportunity to link GWAS results to the transcriptome level, in which a GWAS hit matching an eQTL for a given gene, brings up the hypothesis that the expression of this gene influences the particular phenotype. This transcript level can be analyzed in tissues relevant to the phenotype of interest. Pre-identified gene-phenotype and gene-environment interactions also form important resources because they allow us to both confirm the relation between the environmental causes that could lead to the gene network found (for instance determined by nutrition), and the relation between the gene network and the disease phenotype, in order to identify subnetworks related to more specific aspects of the phenotype (*e.g.* inflammation). CardioGxE is an important resource for such gene-environment and gene-phenotype interactions [9]. We use GTEx and CardioGxE to better understand the SNP and gene connections in the network, in relation to pathway context, environmental causes and the T2DM phenotype.

Key pathways in Type 2 Diabetes Mellitus

We chose to perform an analysis on T2DM data because this is a highly investigated complex disease. There are many and different types of T2DM data and biological information published in articles or stored in databases that can be re-used and integrated for secondary analysis [19]. T2DM is the inability to regulate glucose levels in the blood associated with the development of insulin resistance. This insulin resistance can be systemic or tissue specific. The high glucose levels progressively stress the pancreatic beta-cells, which respond by increasing secretion of insulin. Insulin induces glucose uptake in skeletal muscle, and regulates both glucose production in the liver and the release of free fatty acids from adipose tissue. The insulin imbalance leads to complications related to those organs. Pathway analysis results of T2DM GWAS studies [26, 50] have identified molecular pathways involved in the tissues previously mentioned such as: pancreas, liver, adipose and skeletal muscle. For example, the G-protein signaling pathways are known to activate genes like mitogen-activated protein kinases (MAPK) resulting in insulin resistance, regulation of lipid metabolism, and calcium signaling that converges in AKT signaling and promotes glucose uptake in response to insulin [26]. Another example is the neural development processes found to be enriched with well-known T2DM genes like *TCF7L2* that also have a role in the WNT-signaling pathway, involved in the regulation of pancreatic development [50].
Identifying the genetic influence on the pathways implicated in T2DM pathophysiology is an influential step to determine the genetic predisposition of this complex disease,

and offers targets for development of pharmacological agents. In this study we apply a novel network biology approach to identify genes and pathways relevant in T2DM based on GWAS results. To gain more insights into the possible role of identified genes in T2DM, we use several databases and literature search tools. Lastly, we identified a number of genes with known influence on T2DM phenotypes and others with potential molecular roles in T2DM, but which require further validation.

Materials and Methods

GWAS dataset

The GWAS results described in the current study are taken from a human GWAS meta-analysis conducted by Johnson and O'Donnell in 2009 [21]. The authors used a custom computer analysis to extract and collect 56,411 significant SNP-phenotype associations, in an publicly available GWAS database, from 118 previously published GWAS studies related to different phenotypes. As stated in the paper "the database represents results from an heterogenous set of studies with varied amounts and types of data available". The description of the included studies, and the information on how the meta-analysis was conducted (*i.e.* search strategy, study quality, heterogeneity between study variance, etc) are reported in the Method and Material section of the original paper [20]. We extracted 1,971 SNPs (in August 2016) associated with T2DM, which came from nine of the T2DM GWAS articles collected [24, 27, 30, 35, 36, 39, 43, 46, 51], and a total of 22,363 samples were considered. From this pool of SNPs 1,621 SNPs are from populations with European ancestry (and hence highly relevant to LD analysis performed with CEU data), 195 SNPs are from a MEX population and 155 SNPs are from American Indians. Study information for these articles related to: number of cases and controls, genotyping arrays used, phenotype descriptions, replication samples, analytic strategies, data availability, URLs, publication date and contact information are listed in additional file 2 and 3 of the dataset publication [21].
Consequently, in December 2017, an additional 757 SNPs associated with T2DM were retrieved from the GWAS Catalog (`https://www.ebi.ac.uk/gwas/`) and their genes and pathways were investigated. Those SNPs are not present in the Johnson and O'Donnell dataset, because they were detected in studies performed after their analysis. We report the full list of genes and pathways related to the 757 SNPs as Supplemental material in Table S1.

Workflow of the analysis

Our workflow of the data analysis is presented in Figure 4.1 showing the different steps and tools used. We started our analysis by mapping the 1,971 T2DM-linked SNPs to the genome using the Variant Effect Prediction tool (VEP[15]). This tool gives both the chromosomal location and the known consequences of the variants in the gene sequence (as defined by Sequence Ontology [25]), transcripts, proteins, and regulatory regions (`http://www.ensembl.org/info/docs/tools/vep/index.html`). We obtained 716 variants located in intergenic regions and 1,255 SNPs positioned within 1 kbp up and downstream of the 5' and 3' UTR of 1,046 genes.

We used Ensembl BioMart to retrieve the Ensembl gene identifier and gene name for the 1,046 genes linked to the T2DM SNPs [29]. Next, we checked their pathway involvement using the human complete WikiPathways curated collection [16] (Analysis performed in October 2016, 710 pathways). We identified 368 of 1,048 genes in 460 different pathways.

The 678 genes not present in any of the pathways were annotated with Gene Ontology (GO) terms using GOElite [7] in which 672 genes were detected in the three top level Gene Ontology (GO) trees: molecular function, biological process and cellular component. The GO annotation of these 672 genes was obtained running the GOElite analysis with default parameters (Z-score cutoff for initial filtering above 1.96, 3 minimum number of genes changed (genes connected to a GO term), permutated p-value cutoff above 0.05, excluding terms with gene ID counts greater than 10000). Then, complete results list was used without taking into account the parameters chosen such as: Z score or number of genes changed. As we were trying to complement the biological pathway analysis, we then focused the investigation on the 1,503 GO terms (associated with 196 genes from the original 672) found in the biological processes tree. Furthermore, we used the pathway list obtained from WikiPathways in combination with the gene-variant relations retrieved from BioMart to create a SNP-gene-pathway network, using Cytoscape 3.3.0 [8]. This network contains nodes for all three types of entities. Pathways are included whenever they contain one or more genes that were found to be associated with one or more of the SNPs. These genes then become part of the network and are connected to all pathway nodes in which they occur. Finally, SNPs are connected to the gene(s) to which they were mapped.

Before building this T2DM network, we implemented procedures that allowed us to reduce the redundancy of the pathways used. The aim of this was to obtain a less crowded visualization, without losing the relevance of given networked pathways. We manually evaluated the pathway names and content and whenever two were found with similar names and overlapping content, the smallest pathway was removed. This led to 36 pathways being excluded. On the remaining 424 pathways we performed a cluster analysis

in R using the hclust function with Euclidean distance and complete linkage clustering. We obtained 81 clusters of pathways and 36 individual pathways. We displayed the resulting 117 pathway clusters as a network in Figure 4.2.

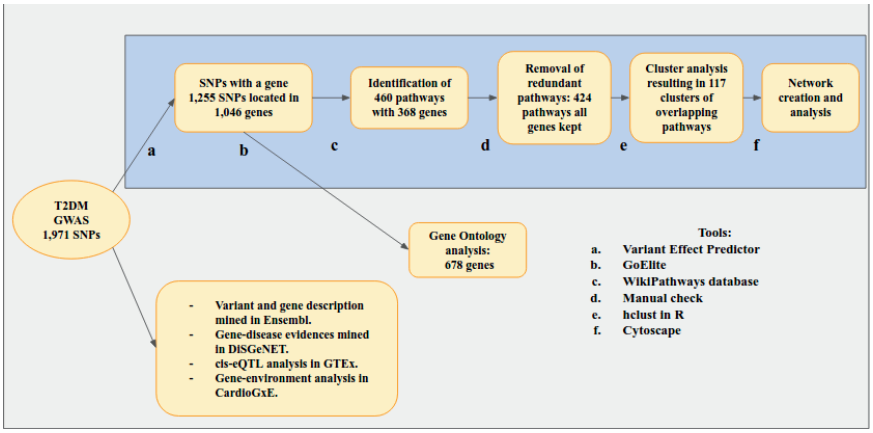

Fig. 4.1 Workflow of GWAS data analysis The data processing, online resources and tools used to perform the GWAS data analysis and visualization, as described in Materials and Methods.

Information sources for GWAS data interpretation

We used several types of online sources and performed web-based analyses to retrieve specific biological information regarding SNPs and the genes to which they are related. We include these additional steps in the workflow Figure 4.1. The information retrieved from those sources enhanced the understanding of the relations between SNPs, genes and pathways.

Regarding the SNPs the following database and analyses were performed:

1) SNP descriptions were obtained from Ensembl VEP, in which we checked the variant location and the consequences of the variation on the DNA sequence (*i.e.* intronic, missense, regulatory, etc).

2) Ensembl Variation database was consulted directly to find citations related to the variants (queried by rs IDs) and to the diabetic phenotype, and the chromatin state regarding the variant location.

3) A Pairwise Linkage Disequilibrium test on the T2DM SNPs with genes was performed for the CEU and MEX SNPs, using the SNP annotation tool SNAP (`http://archive.broadinstitute.org/mpg/snap/ldsearchpw.php`) [20]. The number of CEU SNPs is 504, the MEX SNPs is 50 and the American Indians is 28, and in Figure S1 the SNPs with genes are colored differently according to the population

in which they were detected. The parameters used were: HapMap 3 (release 2), CEU population for the 504 SNPs and MEX population for the 50 SNPs, r2 threshold 0.8, and Distance limit 500 kb. The LD test was not performed for the American Indian population because of the lack of a suitable genetic dataset.
4) The T2DM SNP list was used to query the CardioGxE database [2] to verify if any GWAS variant was previously reported to have a significant gene-environment interaction.
5) The entire set of GWAS, with data from 7,051 post-mortem samples representing 44 tissues and 449 individuals was queried at the GTEx portal[12, 33] in order to identify cis-eQTL. Particular to the GTEx data, we first queried the 716 SNPs not mapping in or near genes to retrieve eQTL data in any tissue. Then we checked for cis-QTLs for the 1,255 SNPs that map in or near genes and the genes to which they map for eQTLs in four T2DM relevant tissues: pancreas, liver, subcutaneous adipose tissue, and skeletal muscle. Finally, the results were visualized in a Venn diagram created with Venny [44].
Concerning gene information, we manually retrieved functional gene descriptions for the genes from GeneCards (`http://www.genecards.org/`) and disease associations from DisGeNET (`http://www.disgenet.org/`) [14]. The evidence obtained was sorted using "Diabetes" as a keyword and the article pertinence was evaluated using both the DisGeNET score and reading the article. DisGeNET score ranks gene-disease associations according to their level of evidence calculated by an algorithm that considers the number and type of sources present in the database, and the number of publications that support the association. For some genes we also performed an additional search using PubMed and Google Scholar in which a query of "gene name AND Diabetes" was used. From the list of articles retrieved, abstracts were scanned, and only those that reported gene name and type 2 diabetes were further analyzed.
As a last step, a search of WikiPathways was used to read the details of the Diabetes related pathway diagrams (*i.e.* Description, Ontology tag, etc) and to evaluate the role of the identified genes.

Results and Discussion

SNP analysis

The 1,971 SNPs associated with T2DM were analyzed using different tools and database information from the resources mentioned above, in order to have a detailed description of the variants, the genes and the pathways that can lead to plausible biological mechanisms regarding T2DM risk or onset or progression of this disease. 98% of T2DM SNPs were mapped to non-coding regions of which 716 SNPs (36%) were outside gene regions (called intergenic variants). The percentage of total non-coding variants in the

dataset is consistent with what other GWAS studies detect [11]. The non-coding variants are described in more detail in the section that covers the eQTL analysis.
The coding SNPs in the dataset consist of: two nonsense SNPs: rs328 and rs2499953, each with high impact on the *LPL* and *MMP26* genes respectively, five missense SNPs (rs2271586, rs5215, rs2499953, rs10494217, rs13088) with moderate impact on the gene protein function according to SIFT and PolyPhen scores, and six synonymous variants with no resulting change to the encoded amino acid, but which nonetheless may alter translation rates and protein structure and function [45].
Regarding the genes with high-impact mutations, *LPL* is a lipoprotein lipase and its mutations increase ending diabetes mellitus. Matrix metalloproteinases (MMPs) are proteolytic enzymes belonging to the family of zinc-dependent endopeptidases that are risk of hyperlipidemia, a known complication in T2DM. *LPL* is a key enzyme in human lipid metabolism that facilitates the removal of triglyceride-rich lipoproteins from the bloodstream [23, 32]. For *MMP26* a risk allele was reported to be associated with higher fasting plasma glucose [10], and the *MMP26* gene is known to have a role in diabetic nephropathy occurring after a longstcapable of degrading almost all the proteinaceous components of the extracellular matrix. It is known that MMPs play a role in a number of renal diseases, such as various forms of glomerulonephritis and tubular diseases, including some of the inherited kidney diseases [48]. The fact that *MMP26* is a T2DM GWAS hit and carries a nonsense SNPs involved in T2DM complications, is sufficient to warrant further investigation of this gene and its involvement in T2DM.
Finally, identification of proxy SNPs allowed determination of redundancy within each dataset. Strong LD (r2 ¿ 0.8) between CEU SNPs was found in 17% of the 504 variants with genes, and for the MEX SNPs strong LD was present in 24% of the 50 SNPs with genes. We also identified genes where the SNPs are in strong LD only in MEX, such as *OR51A7*. Nonetheless, pathway analysis was performed for all genes, identified in populations of European or non-European ancestry, because pathway function and disease phenotypes are highly conserved across populations.

Analysis of the SNP-gene-pathway network

The curated human WikiPathways collection was used to retrieve pathway information on the genes related to the T2DM GWAS SNPs. In WikiPathways genes and metabolites are connected by lines that show meaningful interactions and/or chemical reactions between entities present in the pathway. From the 1,046 SNP related genes identified via Ensembl BioMart, 368 were found in a total of 460 pathways of the complete 709 pathways in the WikiPathways curated collection. In order to achieve a comprehensive picture of the biological processes related to all genes and pathways detected by the T2DM SNPs, a SNP-gene-pathway network was created, following the steps explained

in the blue box of the workflow in Figure4.1. The network consists of 580 SNPs located in 368 genes present in 460 pathways, which in Figure4.2 are shown as 117 cluster nodes. Cluster nodes were created by merging redundant pathways that share the same genes as reported in the methods section. Furthermore, to better describe and discuss the biological connections of the network nodes, we conducted both a network topology investigation and an integration of information from other data sources. In particular, we studied the node degree distribution using the Cytoscape NetworkAnalyzer module [47] and we integrated additional information from other databases and sources, such as: GeneCards (gene description), Ensembl (gene and variants description), DisGeNET (evidences on gene-disease association), CardioGxE (gene-environment interaction), and PubMed and Google Scholar (for confirmation of gene function).

Focus on the gene-pathway connections

For each gene node, the number of gene-pathway connections (node degree), was calculated to detect which genes have the highest number of connections to pathways. 27 genes were found to be connected to ten or more pathways even after removal of redundant pathways. These genes are located in the core of the network. The list shows either pleiotropic genes such as transcription factors (*i.e. NFBK1, CREB1*) and serine/threonine kinases (*i.e. PRKCA, CHUK, JAK2*) or typical T2DM genes (*i.e PPARG*). Furthermore, we explored the gene-disease association in T2DM using DisGeNET, and we summarized the findings in Table S2 in supplemental material. For 18 of the 27 core genes in the network we found evidence related to T2DM phenotype in DisGeNET. The DisGeNET score for these 18 varied from 0.001 for *JAK2, MAP3K1*, and *TGFBR1* to 0.393 for the most T2DM associated gene *PPARG*. The scores rank the gene-disease associations according to their level of evidence, range from 0 to 1, with the higher score indicating greater confidence in the gene-disease association. The genes with a positive DiSGeNET score are displayed as a black triangle in figure 4.2. We explored the pathways shared by at least 2 of the 27 core genes and found 161 common pathways. These pathways were then clustered using the pathway ontology tags present on WikiPathways. The main clusters with most contributing pathways were: pathways related to immunity (*e.g.* B-cell and T-cell receptor signaling, Toll-like receptor signaling, TNF alpha, interferon type I and Interleukin 11 signaling), neuron activity (BDNF signaling and Neurotransmitter receptor binding), cell life cycle related pathways (*e.g.* apoptosis, MyD88 cascade initiated on plasma membrane and endosome), hormone signaling pathways (*e.g.* androgen and estrogen signaling), energy related pathways (leptin, insulin and AGE/RAGE signaling), heart function related pathways (*e.g.* cardiac hypertrophic response) and different types of signaling pathways (*e.g.* EGF/EGFR ErbB, MAPK signaling etc.). The variety of these pathways can be explained by the

pleiotropic action of genes such as kinases or transcription factors, and it is remarkable to observe that, according to previous knowledge, many of these pathways are relevant to T2DM pathophysiology.

All 27 genes are located in the core of the large central network which consists of the nodes with the highest connections. However, nine small network structures are disconnected from the larger network, these consist of fourteen genes (see the black frame on the left side of figure 4.2). The fact that most of the 368 genes are connected in a central network indicates that they function in pathways that are interlinked. Sharing genes between pathways can in fact point at different forms of relations between pathways. Pathways can be functionally related (one regulates the other, or metabolites move each) and can describe related processes in different ways (the strong relationships between cell cycle and cancer pathways, for example) and the shared genes can have real pleiotropic functionality where they play different roles in the two pathways. The latter typically happens for instance for transcription factors that can have multiple targets which can appear in different pathways.

The disconnected small network structures that are represented in the frame in figure 4.2 consist of a single pathway node with the associated genes. These pathways are: 1) energy related pathways (peroxisomal lipid metabolism, bile acid metabolism, glycolysis and gluconeogenesis, TCA cycle and respiratory electron transport), 2) neuron related pathways (synaptic vesicle, GABA metabolism and dopaminergic neuron), 3) pathways related to general cellular processes (RNA transcription, RNA processing and oxidative phosphorylation, and 4) the ACE inhibitor pathway. All these processes have a clear and known involvement in T2DM pathophysiology [26, 50] and there are many known connections between these processes and other parts of the larger network. Some of these pathways are separate from other similar pathways present in the larger network, because the genes known to connect them were not found to have associated SNPs. The gene-pathway isolation within the network could also be the result of incomplete knowledge representations in the pathways, where disease associated genes are not represented in each of the related pathways, not because the pathways as such are unrelated.

Our analysis helps to identify both potential mechanistic links between pathway (sub)networks and an understanding of epistasis (SNP-SNP or gene-gene interactions) [34] that is supported by SNPs mapping to different pathways. We found an example of a missing link between the unconnected pathways previously listed, for instance between the TCA cycle (WikiPathways ID: WP2766) and the glycolysis and gluconeogenesis (WP534) pathway. The two processes are clearly related but the conceptual division can be made in such a way that no genes are shared. WikiPathways in fact has a mechanism to show this: the pathway diagram can have an explicit link to another pathway diagram. However, this pathway-pathway relation is still hard to interpret

in the type of network that we presented here, because we identify a connection between a gene in one pathway and an entire other pathway, and not with a specific gene (pathway 1) to gene (pathway 2) connection. We also found an example of a missing link between the unconnected processes and the processes in the large central network, regarding again the TCA cycle and glycolysis and gluconeogenesis pathways, and a pathway that describes the transport of glucose and other sugars (WP1935). The latter pathway contain the SLC family of glucose transporters, of which seven SLC genes are present in the network, but without links to the main glucose metabolism related pathways (WP2766 and WP534). These findings are useful to improve the WikiPathways collection, where we could add pathway links to all three diagrams. Clarifying the functional connections between genes in different processes that contribute to the synergistic effects found in GWAS studies, represented in our network, helps to understand the background of epistasis.

Focus on the SNP-gene connections

Several genes have a relatively high number of SNPs associated with T2DM (*CNTN1, GRB10, PRKCA, ZNF615, SYNE1, THSD7B, NRG1, DDOST, SLC13A1, HIPK2, ATP8A1, ARHGAP26, TCF7L2, CDKAL1*). The Table S3 in supplemental material gives an overview of the number of related SNPs found for every gene. It should be noted that some SNPs are in high linkage disequilibrum (LD) and thus would point to the same causal variant. For example there are four SNP-gene network structures with a number of SNPs greater than ten. The genes at the centers of these networks are: *CDKAL1, ATP8A1, ARHGAP*, and *TCF7L2*, but considering LD only the gene *CDKAL1* remains connected to at least 10 SNPs independent of each other. *CDKAL1* variants have been reported to be associated with T2DM with highly significant p-values detected by different GWAS studies [18].

In the blue frame of Figure S1 in supplemental material 19 structures are placed, in which one or more SNPs connect with at least two genes, meaning that those SNPs are located in multiple genes according to the size of the gene regions chosen. Such genetic overlaps are well known, and they are important for the biological interpretation of the outcome from GWAS studies. In such cases the knowledge of the function of the associated genes can be used to decide which relations are more plausible [37]. For this purpose we collected gene-disease associations with T2DM and their scores from the DisGeNET database regarding the 41 genes present in the 19 structures, genes with such scores are shown as black triangles in figure 4.2, and we also report the DisGeNET score that indicate strength of the gene-disease association in Table S4 of the supplemental material. The DisGeNET analysis reveals that 13 of the 41 genes found to have SNPs associated with T2DM in the GWAS analysis already have known relationships

with T2DM (31%), a value likely the result of evaluation of the same GWAS studies. However, if when considering the SNPs in in Table S4 that overlap with multiple genes, SNPs associated with these 13 previously known T2DM genes are also associated with 8 other genes. In some cases such a dually connected SNP is the only association found for a specific gene, which reduces the likelihood that such a disease relationship is real. Finally, we checked if any of the SNPs and related genes in the network presented a known gene-environment interaction in the CardioGxE database. Finding relevant gene-environment interactions in the database adds supports for their involvement in the disease mechanism. We found six SNPs (rs9939609, rs1801282, rs7903146, rs328, rs693 and rs780094) influenced by thirteen environmental factors such as: energy intake, whole-grain intake, fiber intake, carbohydrate, fat, polyunsaturated, monounsaturated and saturated fatty acid, Vitamin E and A, normal diet, Mediterranean diet and physical activity. Those SNPs are located in six genes (*FTO, PPARG, TCF7L2, LPL, APOB*, and *GCKR* respectively), most of which are well-known to be associated with T2DM and its complications [28]. CardioGxE also provides a list of phenotypic traits related to these genes such as: body mass index (BMI), insulin, triglyceride, and cholesterol. The complete results of the gene-environment interactions are reported in Table S5 of the supplemental material.

Gene Ontology analysis of the genes without pathways

We used GOElite [7] to provide a biological description from the Gene Ontology for the 678 genes that were not found in the complete WikiPathways curated collection, and therefore not represented in the SNP-gene-pathway network. When we used all three main GO classes, molecular function, biological process and cellular component, 672 genes were annotated. For only 196 of these genes we found an annotation from the biological process tree and we focused on these. Although this resulted in less than half of the input genes being evaluated, this approach has the advantage that the annotation comes closest to a biological pathway description. The total number of biological processes identified for these 196 genes was 1,503. In order to further evaluate these annotated genes we also performed the same Gene Ontology annotation for the 368 genes from the SNP-gene-pathway network, and for these genes 4,544 biological processes GO terms were identified related to 351 of the 368 input genes. We compared these annotations of the two groups of genes. From the 1,503 annotations identified for the 362 genes not assigned by WikiPathways, 1,255 were observed previously for the 363 genes in the SNP-gene-pathway network. These 1,255 annotations are related to all the 351 genes with GO terms in the original SNP-gene-pathway network, and to the 122 out of 196 genes in the newly annotated group. This means that for the 74 (196 - 122) genes in the newly annotated group, there are biological process annotations not previously

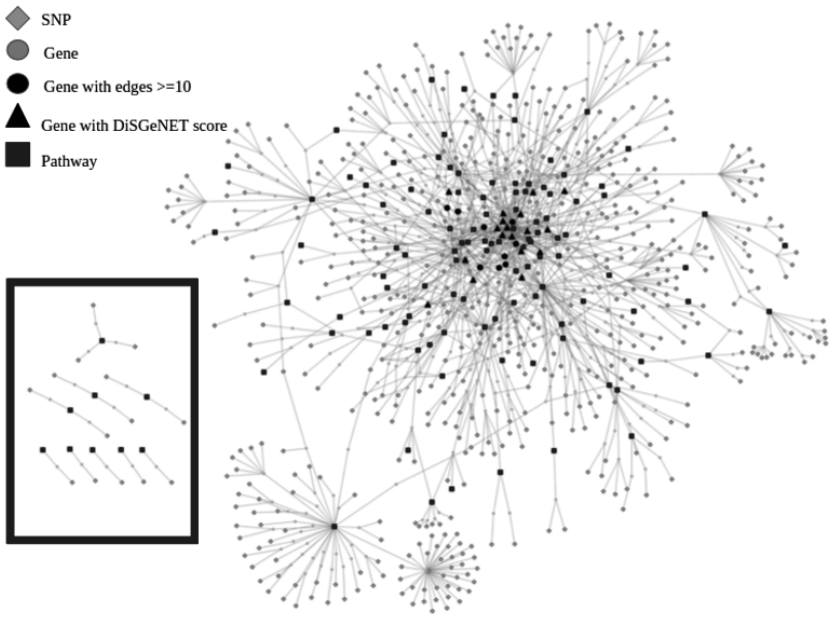

Fig. 4.2 SNP-gene-pathway network. The network displays 580 SNPs (green diamonds) located in the selected region for 365 genes (circles) present in 117 pathway clusters (blue squares). Black symbols indicate genes with ten or more connections to pathway clusters, and triangles indicate genes with a positive DisGeNET score (note that these are all black). The disconnected SNP-gene-pathway subnetworks are shown on the left, framed in black.

found. In total this added 248 (1503 - 1255) new biological process annotations to the T2DM SNP/gene set. A network visualization that illustrates the connections between these 74 genes and 248 GO terms is provided in figure 4.3.

Two points related to the Gene Ontology comparison are worth noting: first, the 1,255 common terms extracted describe most of the original SNP-gene-pathway network and the corresponding 351 genes represent most of the pathways in that network. This implies that we have identified new information for the biological processes related to most of the pathways. Exceptions include the Aminoacid conjugation and Lnc-mRNA mediated mechanism of therapeutic resistance pathways for which we did not find any new related genes. To use the information about genes discovered by GWAS analysis for diabetes that are not in current pathways with similar Gene Ontology annotations as genes that are, and the information about discovered genes that have different gene ontology annotations, expert evaluation is required. This could lead to an extension of existing pathways if the discovered genes with similar annotations are in fact known to be involved in these pathways, or to development of new pathways that would cover

processes found in the Gene Ontology not currently covered. In this way this information can be used to improve pathway curation, especially for pathways involved in the pathophysiology of T2DM. The second notable observation from the analysis of GO terms concern the connections between 248 biological processes exclusively linked to 74 genes that are not currently in any WikiPathways pathways are visualized for further analysis in figure 4.3. Of course this network is likely to contain much of spurious information. If only one gene related to a biological process was found, this result is more likely to relate to a gene-process relationships not really related to T2DM. This could be because of: a false GWAS result, or the genes is involved in multiple processes or the SNP in that gene acts at a distance.

In order to have a general overview of the different biological processes in the network, related biological processes were clustered in the same frame of the figure 4.3. Then, we constructed for each frame a GO ancestor chart of the biological terms, using the old version of QuickGO (`http://www.ebi.ac.uk/QuickGO-Old/`). In these charts GO slim terms were used to provide a summary of the results, which then gives an overview of the ontology content of the terms present in the tree. An example of the ancestor chart, with the GO slim terms related to the GO biological processes clustered in frame "l" of the figure 4.3, is reported in Supplemental material.

In figure 4.3 four frames functionalities referred to basic cell division and development such as: cytokinesis and different types of regulation of meiotic cell cycle regarding: sex determination, chromosome separation, telomere maintenance and polar body extrusion, stem cell fate determination, brain and skeletal muscle development. There is also a group of processes involved in cell communication regarding: protein targeting and transport, cell-cell junction maintenance, membrane raft assembly, and immune system processes such as: T cell antigen processing and pattern recognition of the toll-like receptor. Other groups are: DNA modification (*e.g.*: histone methylation, acetylation and ubiquitination, and G1 DNA damage checkpoint) and ATP metabolism processes such as purine ribonucleoside metabolic processes, mitochondrial metabolism such as cytochrome complex assembly, general functionalities regarding: protein modification (*i.e.* methylation, acetylation and poly- and de- glutamylation), regulation of metabolic processes related to: chitin and hydrogen peroxide catabolic processes, ubiquitin dependent protein, bile regulation, thyroid hormone generation, and glycolytic process. Finally, players of signaling cascade involved in cGMP catabolic process, epidermal growth factor-activated receptor activity, kinase A, TOR, GTPase, NIK/NF-kappaB, and keratinization. These results are an indication of relevant molecular processes in which the genes detected by the T2DM GWAS study play a role. Further literature investigation is required to expand the knowledge about the connection of these precesses and their genes in T2DM context, especially with respect to clinical measures of disease risk.

In conclusion, the Gene Ontology analysis performed for genes that currently were not assigned to pathways allowed us to identify 1) genes that are indeed related to several pathways identified by more fully annotated T2DM SNPs-genes, 2) genes that are functionally related in processes not covered currently in pathways, and 3) gene-process relationships that occur only occasionally (and are likely either to be of no real value to understand the disease development, or to have a conditional relationship with T2DM, such as via epistasis or gene-environment interactions).

eQTL analysis

In order to evaluate which T2DM variants influence expression of the gene to which they map or any other genes, we used the GTEx portal to search for cis-eQTL SNPs within the T2DM SNP set. We first checked if any of the 716 SNPs that did not map to a specific gene had GTEx data indicating an eQTL function. We then evaluated whether the SNPs that influence expression of a gene are known in GTEx to affect cis-eQTLs in subcutaneous adipose tissue, liver, pancreas or skeletal muscle. For each tissue we selected the genes from cis-eQTLs with a p-value minor and egual of 0.05 (Table S6 supplemental material), and we visualized the results in a Venn diagram in figure 4.4. The Venn diagram shows that most of the cis-eQTLs are tissue-specific: 36% in pancreas, 26% in adipose subcutaneous, 20% in skeletal muscle and 9% in liver. This result confirms previous findings suggesting that GWAS variants are enriched related to tissue-specific cis-eQTLs [12]. The selected tissues are relevant in T2DM, and we checked the detected genes, finding well-known T2DM genes such as: *PPARG* and *TCF7L2* in pancreas or *APOB* and *LPL* in subcutaneous adipose tissue.
Moreover, we found 45 cis-eQTLs genes shared between at least two tissues. Within this set of genes we found only one gene (*MAP3K8*) present in the core of the network of figure 4.2 because it has a high number of pathway connections, and it is present in several important pathways, such as: Insulin signaling (WP481) leading to a cell growth differentiation, and TNF-alpha signaling (WP231) ending to the activation of *NFKB*. A significant positive effect of diabetes related SNPs was found for *MAP3K8* expression in pancreas and liver, in adipose tissue there was a tendency towards a negative effect on expression but the p-value was only 0.45. The *MAP3K8* gene encodes a kinase that in adipocytes is involved in inflammatory cytokineinduced ERK1/2 activation, and deregulation of its expression suggests a role in adipose tissue dysfunction in obesity [31]. However, *MAP3K8* does not activate insulin in adipose tissue and does not trigger its effects like lipolysis. The presence of *MAP3K8* in a diabetes related eQTL in liver and pancreas could still lead to a significant alteration of the downstream insulin signaling pathway. For instance in the Angiopoietin Like Protein 8 Regulatory Pathway (WP3915), *MAP3K8* is activated via the insulin cascade together with other

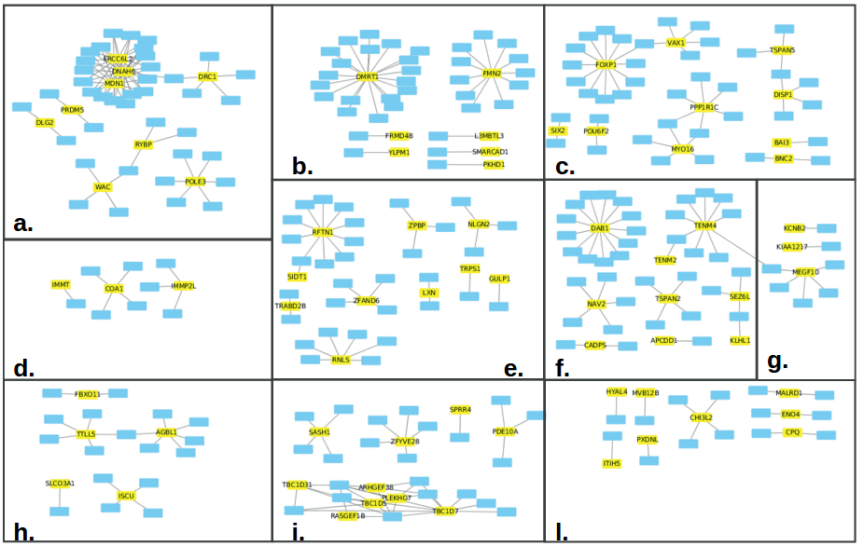

Fig. 4.3 Overview of the GO biological processes exclusively linked to genes without pathways. The image represents 248 GO biological processes (blue rectangles) linked with 74 genes without pathways (yellow rectangles). The processes are grouped in ten frames according to similar functions. Each group is identified with the highest level GO term that identifies the general action of the processes. a. DNA modification (*e.g.*: histone methylation, acetylation and ubiquitination, and G1 DNA damage checkpoint) and ATP metabolism processes such as purine ribonucleoside metabolic processes; b. Cytokinesis and different types of regulation of meiotic cell cycle regarding: sex determination, chromosome separation, telomere maintenance and polar body extrusion; c. Stem cell fate determination; d. Mitochondrial metabolism such as: cytochrome complex assembly and mithocondrial protein processing; e. Cell communication regarding: protein targeting and transport, cell-cell junction maintenance, membrane raft assembly, and immune system processes such as: T cell antigen processing and pattern recognition of the toll-like receptor 3; f. Brain development regarding: astrocyte, microglia, glial, myelin and synapse maturation; g. Skeletal muscle development; h. Protein modification *i.e.* methylation, acetylation,poly- and de-glutamylation); i. Several signaling cascades related to: cGMP catabolic process, epidermal growth factor-activated receptor activity, kinase A, TOR, GTPase, NIK/NF-kappaB, and keratinization; l. Metabolic processes especially chitin and hydrogen peroxide catabolic processes, ubiquitin dependent protein, bile regulation, thyroid hormon generation, and glycolytic process).

MEK/MAP kinases and another downstream effect, beside cell differentiation, is the activation of the *ANGPTL8* gene that is known to be a relevant regulator of glucose and lipid metabolism in liver, and it is identified as a novel drug target for treatment of T2DM and dyslipidemia [22].

Finally, we checked the common genes identified in the three approaches: the cis-eQTL

tissue-specific lists, the gene-environment analysis and the highly connected pathway based network. We identified three genes and their (partner) SNPs present in all three lists: *PPARG* in pancreas cis-eQTL (rs1801282), and *LPL* (rs328) and *APOB* (rs693) in adipose tissue cis-eQTL. *PPARG* is a transcription factor and the molecular target of the insulin-sensitizing drug, thiazolidinedione; its variant rs1801282 is robustly associated to reduced risk of T2DM in different populations [38]. In contrast, *LPL* and *APOB* variants are not (yet) considered to be associated with T2DM [28], but in the CardioGxE database they show association with diabetes related traits such as: *LDL* cholesterol and triglycerides level for *APOB* and HDL-cholesterol and triglycerides level for *LPL*. In particular, the *LPL* rs328 SNP was previously mentioned to have a high impact on sequence consequences according to VEP, because it is a nonsense mutation that truncates the *LPL* protein to 446 instead of 448 amino acid residues. Despite the high impact effect assigned to this *LPL* nonsense mutation, there are controversial results from *in vitro* studies that show either a normal enzyme activity of the *LPL* protein with the missing codon at the carboxyl-terminal [23] or that the mutation might be responsible for a defect in lipid interface recognition [32]. Yet another study examined the association of *LPL* with T2DM in a Korean population taking into account different *LPL* SNPs including rs328, that had significant associations with blood glucose-related phenotype [40]. In conclusion, our finding still supports the relevance of this and other genes like *APOB* in the context of T2DM.

Conclusions

In this Bioinformatics study we were able to re-analyze the output of previously published Type 2 Diabete Mellitus (T2DM) GWAS studies by integrating several types of biological information regarding the GWAS SNPs and the genes in which the variants are located, with the relevant findings summarized in Table 4.1. We also took advantage of network visualization to recognize different types of biological relationships, such as: genes highly connected with pathways, and overlapping genes that share the same SNP GWAS hit.

First, we identified pathways relevant in T2DM. We then analyzed a SNP-gene-pathway network that provided information about the types of biological processes in which the GWAS genes have a role. We combined this with other information about the biological roles of the genes in the network obtained from various sources (CardioGxE, GTEx, DiSGeNET). We propose these network steps as a method complementary to the standard pathway analysis, because it allows the visualization of the relationships between GWAS genes, different functional annotations, the relevant SNPs, and pathways containing these. Next, we selected a number of relevant genes that are featured in small SNP-gene network nodelets (basically the edges in these nodelets consist of

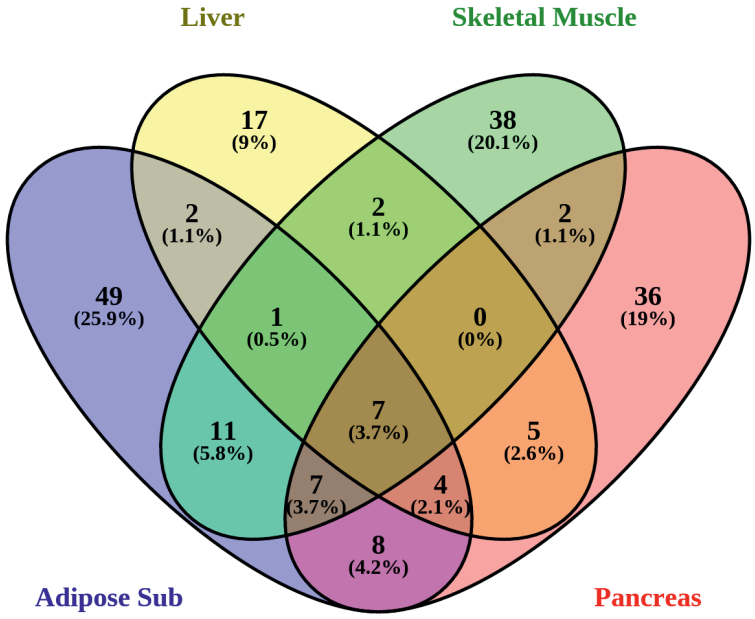

Fig. 4.4 cis-eQTL in T2DM-relevant tissues. The Venn diagram indicates the numbers of cis-eQTLs in pancreas, liver, adipose subcutaneous and skeletal muscle, and the numbers shared among these tissues.

SNPs known to be located in more than one gene). This highlights the SNPs that can affect multiples genes. Sometimes choices for one or the other interaction can be made based on the observed phenotype and the function of the affected genes.

Furthermore, we addressed the T2DM GWAS genes that were not included in the WikiPathways collection, and as a consequence were not analyzed in the SNP-gene-pathway network. For those genes the annotation tree of biological processes GO terms were analyzed, and a list of relevant terms linked to 74 T2DM GWAS genes is provided as a network visualization for further investigation. Next we created a connection map based on the GO biological process annotation shared by both the 351 genes found in pathways and the 122 genes not found in pathways. Apart from their direct utility to study the relationship between disease related genes and their annotated functions in GO, such connectivity maps are also useful to evaluate which disease relevant processes have not been captured well in pathway collections.

Table 4.1 Summary of the relevant genes detected in the secondary analysis of T2DM GWAS study.

Type of gene detection	Gene name
Genes with nonsense SNP	LPL, MMP26
Genes with missense SNPs	MMP26, KCNJ11, VSTM4, ART5, TBX15
Genes with synonimous SNPs	OR51A7, SVIL, ARF3, PLEKHG7, VSIG10, RP11-302B13.5
Genes highly connected with pathways	27 genes listed in S1 Table
Genes disconnected from the SNP-gene-pathway central network in Figure 2	VMP, TSEN, SYT, SET, RPP3, RIMS, NR3C, NFI, NDUFS, MPC, MBD, HSD17B, FADS2, CPLX2
Genes overlapping the same SNPs	41 genes listed in S3 Table
Gene highly detected by significant GWAS SNPs	CDKAL1
Genes with significant Gene-Environment interaction	35 genes listed in S4 Table
Genes detected by GTEX cis-eQTLs in adipose, liver, pancreas, and skeletal muscle tissue	264 genes listed in S5 Table
Genes with common GTEX cis-eQTLs, Gene-Environment interaction and high pathway connection	PPARG, LPL, APOB

Finally, we used three more disease targeted knowledge resources to select relevant genes in the SNP-gene-pathway network that are known to share more specific biological functionality. 1) The CardioGxE database depicted genes involved in specific environmental interactions, but also associated with T2DM related traits. 2) eQTL analysis provided an idea of the influence that the variants can have on the expression of the genes in the network at the tissue level. 3) DisGeNET shows the evidence regarding the gene-disease association and applying this suggested to us those genes in the SNP-gene-pathway network that do not have compelling evidence to confirm a T2DM role. A combination of the approaches described allowed us to identify genes such as *LPL* and *APOB*, of which the variants appear to play an important role in T2DM but have not been well studied in this context so far.

Acknowledgments

Any opinions, findings, conclusion, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the U.S. Department of Agriculture. Mention of trade names or commercial products in this publication is

solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The USDA is an equal opportunity provider and employer.

Funding statement

This project/research has been made possible with support of the Dutch Province of Limburg. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] William S. Bush and Jason H. Moore. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12), 2012.

[2] Laurence D Parnell, Britt A Blokker, Hassan S Dashti, Paula-Dene Nesbeth, Brittany Elle Cooper, and et al. CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits. *BioData Mining*, 7(1):21, dec 2014.

[3] Miguel A. Garcia-Campos, Jesus Espinal-Enriquez, and Hernandez-Lemus. Pathway analysis: State of the art. *Frontiers in Physiology*, 6(DEC):1–16, 2015.

[4] Kai Wang, Mingyao Li, and Maja Bucan. Pathway-based approaches for analysis of genomewide association studies. *American journal of human genetics*, 81(6):1278–83, dec 2007.

[5] Patrick Y.P. Kao, Kim Hung Leung, Lawrence W.C. Chan, Shea Ping Yip, and Maurice K.H. Yap. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochimica et Biophysica Acta - General Subjects*, 1861(2):335–353, 2017.

[6] Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[7] Alexander C Zambon, Stan Gaj, Isaac Ho, Kristina Hanspers, Karen Vranizan, and et al. Go-elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics*, 28(16):2209–2210, 2012.

[8] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, and et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[9] Laurence D Parnell, Britt A Blokker, Hassan S Dashti, Paula-Dene Nesbeth, Brittany Elle Cooper, and et al. CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits. *BioData Mining*, 7(1):21, dec 2014.

[10] Stéphane Cauchi, Hélène Choquet, Ruth Gutiérrez-Aguilar, Frédéric Capel, and et al. Effects of TCF7L2 polymorphisms on obesity in European populations. *Obesity*, 16(2):476–482, 2008.

[11] Lucas D. Ward and Manolis Kellis. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(D1):930–934, 2012.

[12] Francois Aguet, Andrew A Brown, Stephane Castel, Joe R Davis, Pejman Mohammadi, and et al. Local genetic effects on gene expression across 44 human tissues. *bioRxiv*, 2016.

[13] K Wang, M Li, H Hakonarson Nature Reviews Genetics, and undefined 2010. Analysing biological pathways in genome-wide association studies. *nature.com*.

[14] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, and et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839, jan 2017.

[15] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, dec 2016.

[16] Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, EgonL. Willighagen, and et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1):D488–D494, jan 2016.

[17] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, and et al. Gene Ontology: tool for the unification of biology. *Nature Genetics 2000 25:1*, may 2000.

[18] Mojgan Yazdanpanah, Chuhua Chen, and Jinko Graham. Secondary Analysis of Publicly Available Data Reveals Superoxide and Oxygen Radical Pathways are Enriched for Associations Between Type 2 Diabetes and Low-Frequency Variants. *Annals of Human Genetics*, 77(6):472–481, nov 2013.

[19] Ming Zhang, Heng Luo, Zhengrui Xi, and Ekaterina Rogaeva. Drug Repositioning for Diabetes Based on 'Omics' Data Mining. *PLOS ONE*, 10(5):e0126082, may 2015.

[20] A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell, and et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24):2938–2939, dec 2008.

[21] Andrew D Johnson and Christopher J O'Donnell. An Open Access Database of Genome-wide Association Results. *BMC Medical Genetics*, 10(1):6, dec 2009.

[22] A Siddiqa, E Cirillo, SHK Tareen, A Ali, M Kutmon Genomics, and undefined 2017. Visualizing the regulatory role of Angiopoietin-like protein 8 (ANGPTL8) in glucose and lipid metabolic pathways. *Elsevier*.

[23] F Faustinella, A Chang, J P Van Biervliet, M Rosseneu, N Vinaimont, L C Smith, and et al. Catalytic triad residue mutation (Asp156—-Gly) causing familial lipoprotein lipase deficiency. Co-inheritance with a nonsense mutation (Ser447—-Ter) in a Turkish family. *The Journal of biological chemistry*, 266(22):14418–24, aug 1991.

[24] Richa Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Richa Saxena, Benjamin F Voight, and et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (New York, N.Y.)*, 316(5829):1331–6, jun 2007.

[25] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, and et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44, apr 2005.

[26] Ali Torkamani, Eric J. Topol, and Nicholas J. Schork. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265–272, nov 2008.

[27] Valgerdur Steinthorsdottir, Gudmar Thorleifsson, Inga Reynisdottir, Rafn Benediktsson, Thorbjorg Jonsdottir, and et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature Genetics*, 39(6):770–775, jun 2007.

[28] Pierre Paul Michel Thomas, Salih Mohammed Alshehri, Henk J van Kranen, and Elena Ambrosino. The impact of personalized medicine of Type 2 diabetes mellitus in the global health context. *Personalized Medicine*, 13(4):381–393, jul 2016.

[29] Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, and et al. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, jan 2016.

[30] Evadnie Rampersaud, Coleen M Damcott, Mao Fu, Haiqing Shen, Patrick McArdle, and et al. Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: evidence for replication from diabetes-related quantitative traits and from independent populations. *Diabetes*, 56(12):3053–62, dec 2007.

[31] Jennifer Jager, Thierry Grémeaux, Teresa Gonzalez, Stéphanie Bonnafous, Cyrille Debard, and et al. Tpl2 kinase is upregulated in adipose tissue in obesity and may mediate interleukin-1beta and tumor necrosis factor-{alpha} effects on extracellular signal-regulated kinase activation and lipolysis. *Diabetes*, 59(1):61–70, jan 2010.

[32] Junji Kobayashi, Tsutomu Nishida, Detlev Ameis, Gisela Stahnke, Michael C. Schotz, and et al. A heterozygous mutation (the codon for Ser447 a stop codon) in lipoprotein lipase contributes to a defect in lipid interface recognition in a case with type I hyperlipidemia. *Biochemical and Biophysical Research Communications*, 182(1):70–77, jan 1992.

[33] Brian Jo, Yuan He, Benjamin J Strober, Princy Parsana, Francois Aguet, and et al. Distant regulatory effects of genetic variation in multiple human tissues. *bioRxiv*, page 074419, sep 2016.

[34] Clément Niel, Christine Sinoquet, Christian Dina, and Ghislain Rocheleau. A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6:285, sep 2015.

[35] Robert L Hanson, Clifton Bogardus, David Duggan, Sayuko Kobes, Michele Knowlton, and et al. A search for variants associated with young-onset type 2 diabetes in American Indians in a 100K genotyping array. *Diabetes*, 56(12):3045–52, dec 2007.

[36] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, and et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, feb 2007.

[37] Paolo Zambonelli, Roberta Davoli, Mila Bigi, Silvia Braglia, Luigi Francesco De Paolis, and et al. SNPs detection in DHPS-WDR83 overlapping genes mapping on porcine chromosome 2 in a QTL region for meat pH. *BMC Genetics*, 14(1):99, oct 2013.

[38] Kazuo Hara, Takashi Kadowaki, and Masato Odawara. Genes associated with diabetes: potential for novel therapeutic targets? *Expert Opinion on Therapeutic Targets*, 20(3):255–267, mar 2016.

[39] Jose C Florez, Alisa K Manning, Josée Dupuis, Jarred McAteer, Kathryn Irenze, and et al. A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets. *Diabetes*, 56(12):3063–74, dec 2007.

[40] YS Cho, MJ Go, HR Han, SH Cha, HT Kim ... and molecular Medicine, and undefined 2008. Association of lipoprotein lipase (LPL) single nucleotide polymorphisms with type 2 diabetes mellitus. *nature.com*.

[41] Jason Flannick and Jose C. Florez. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nature Reviews Genetics*, 17(9):535–549, sep 2016.

[42] Mojgan Yazdanpanah, Chuhua Chen, and Jinko Graham. Secondary Analysis of Publicly Available Data Reveals Superoxide and Oxygen Radical Pathways are Enriched for Associations Between Type 2 Diabetes and Low-Frequency Variants. *Annals of Human Genetics*, 77(6):472–481, nov 2013.

[43] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, and et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science (New York, N.Y.)*, 316(5829):1341–5, jun 2007.

[44] J.C. Oliveros. Venny. An interactive tool for comparing lists with Venn's diagrams.

[45] Chava Kimchi-Sarfaty, Jung Mi Oh, In-Wha Kim, Zuben E Sauna, Anna Maria Calcagno, and et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science (New York, N.Y.)*, 315(5811):525–8, jan 2007.

[46] M Geoffrey Hayes, Anna Pluzhnikov, Kazuaki Miyake, Ying Sun, Maggie C Y Ng, and et al. Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes*, 56(12):3033–44, dec 2007.

[47] Yassen Assenov, Fidel Ramírez, Sven-Eric Schelhorn, Thomas Lengauer, and Mario Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, jan 2008.

[48] X. Xu, L. Xiao, P. Xiao, S. Yang, G. Chen, and et al. A Glimpse of Matrix Metalloproteinases in Diabetic Nephropathy. *Current Medicinal Chemistry*, 21(28):3244–3260, aug 2014.

[49] J. L. Haines, Michael A Hauser, Silke Schmidt, William K Scott, Lana M Olson, and et al. Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science*, 308(5720):419–421, apr 2005.

[50] John R B Perry, Mark I McCarthy, Andrew T Hattersley, Eleftheria Zeggini, the Wellcome Trust Case Control Wellcome Trust Case Control Consortium, and et al. Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes*, 58(6):1463–7, jun 2009.

[51] E. Zeggini, M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, and et al. Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. *Science*, 316(5829):1336–1341, jun 2007.

CHAPTER 5

A genetic reference network to better understand the role of non-coding variants in obesity

Elisa Cirillo[1], Kyoko Watanabe[2], Niels Delahaije[1], Rik van Dael[1], Martina Kutmon[1,3], Michiel E Adriaens[3], Laurence D Parnell[4], Susan LM Coort[1], Chris T Evelo[1,3]

1 Department of Bioinformatics - BiGCaT, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands.
2 Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Neuroscience Amsterdam, VU University Amsterdam.
3 Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands.
4 Agricultural Research Service, USDA, Jean Mayer-USDA Human Nutrition Research Center on Aging at Tufts University, Boston, MA, USA.

Submitted to: International Journal of Obesity

Abstract

Obesity is a widespread complex trait that can lead to detrimental cardiovascular and diabetogenic diseases. Investigating the genetic background of complex traits is an essential component in the development of personalized treatments. However, the vast majority of complex trait single nucleotide polymorphisms (SNPs) identified in genome-wide association studies (GWAS) reside in non-coding regions of the genome, severely complicating their interpretation. Here, we present an integrative network based approach to study such non-coding variants SNPs in the context of obesity. The SNPs from a publicly available GWAS dataset for body mass index (BMI; n = 339,224 individuals) are connected in a network with the genes in which these SNPs are located and the genes are linked to their biological pathways. Moreover, epigenetics data such histone modifications, and expression quantitative trait loci (eQTLs) data from adipose tissue, skeletal muscle, liver and pancreas, are integrated as additional information to advance the understanding of the function of non-coding SNPs. Four tissue-specific genetic reference networks of SNPs associated with obesity are presented, showing BMI non-coding variants both located in region with epigenetically active state properties, and with an influence on gene expression in different tissues. The connections between SNPs, genes and pathways are also represented, enabling the interpretation of the SNP effects at the process level, especially concerning non-coding variants typically not so defined. The networks are available online for further exploration on the NDEx website.

Introduction

Precision medicine for obesity

Obesity, with both environmental and genetic components, is a widespread complex trait that can lead to detrimental cardiovascular and diabetogenic diseases [24]. Many studies have been performed to understand the influence of the obesogenic environment [6, 17] order to decrease the risk of obesity, but preventing the disease in high-risk individuals through lifestyle changes has proven unattainable [8]. In contrast, it is well known that obesity is partly genetically driven, and genetic factors should be explored in order to understand the obesity phenotype fully [45]. In recent studies, several genes in glucose homeostasis have been identified as risk factors for obesity [46, 47] . Aiming, to improve screening for variants known to be associated with high risk of developing type 2 diabetes mellitus (T2DM) enables immediate intervention [27, 31]. However, the identification of the complete set of the genes and proteins changes involved in a complex pathway such as energy homeostasis remains a challenge. Moreover, new genetic factors such as circulating pigment epithelium-derived factor (PEDF)[36] and

circulating lipopolysaccharide-binding protein (LBP)[5] show potential as markers for obesity-related insulin resistance.
The emerging approach of precision medicine takes an individuals variability in genes, environment, and lifestyle into account in order to treat and prevent diseases[4]. Thus, knowing the landscape of genetic variation for a given disease trait, is an essential aspect for applying precision medicine to any treatment regimen. Moreover, both genetic variants and identified biomarkers are essential components that can provide the key to understand the phenotype and the response to treatment.

The challenge of data integration to improve biological interpretation

Genome wide association studies (GWAS) have been one of the most common approaches to identify genetic variants such as single nucleotide polymorphisms (SNPs) associated with the trait of interest [11]. In the context of precision medicine, association studies have limited usefulness in establishing causation and biological function; however, such SNP associations are still very important to indicate the dependence of a specific phenotype on genetic variation [39] .
Currently, one of the greatest tasks in the post-GWAS era is to describe the biological implications of non-coding SNPs, which are the majority of GWAS findings. These non-coding SNPs can regulate the expression of nearby or even distant genes by affecting cell-tissue specific regulatory elements such as enhancers and promoters, so called expression quantitative trait loci (eQTLs), rather than a direct effect on protein function [3, 15]. The histone modifications and expression measurements are used to define if there is any and which type of regulatory mechanisms are active in the position of the non-coding SNPs. Beside the type of SNP effect on genes, the function of the gene and its encoded protein(s) play a role in the larger context of biological pathways. Thus, the combination of both SNPs and gene product actions must be interpreted with respect to the phenotype. Nowadays, pathway analysis allows us to assess SNP-gene relationships in the encompassing biological landscape of genes, proteins, metabolites and other cellular entities [10, 41, 43]. The pathway context helps to investigate simultaneously: the action of multiple genes affected by body mass index (BMI) SNPs, and the changes in the biochemical and physiological process instigated by the BMI SNPs.
In this study, we aim to characterize a set of SNPs associated with BMI, including those in strong linkage disequilibrium (LD), by integrating data such as: tissue specific eQTLs, cell type specific epigenetic data and biological pathways. The relationships between the different biological levels (SNPs, genes, epigenetics, pathways) are visualized using network diagrams [1]. We investigate four relevant tissues often affected in obesity, i.e. subcutaneous and visceral adipose tissue, skeletal muscle, liver and pan-

creas. For each tissue, except for visceral and subcutaneous tissues that are combined, we generated a genetic reference network of SNPs associated with obesity relevant for both, researchers and clinicians. In particular, we focused on the interpretation of effects of non-coding SNPs on genes and their involvement pathways. The data are retrieved from multiple, freely available sources, and we provide detailed instructions, including a video tutorial, to instruct researchers how to re-apply our workflow to their own data.

Results

Construction of genetic variation networks by integrating GWAS and omics data

A workflow is presented to integrate BMI SNPs including the SNPs in LD using 1000 Genomes project European (CEU) population as a reference, and different types of linked omics data in a network representation. The network contains information on genetics (genes in which the BMI SNPs are located), epigenetics (SNPs mapped within cell-specific histone marks indicative of transcription regulation), transcriptomics (SNPs influencing gene expression; eQTL SNPs) and biological pathways (pathways containing genes with BMI SNPs), and it displays the relationships between SNPs, genes and pathways. The procedure to integrate the data is illustrated in Figure 5.1 and the key steps of the network construction are presented in a video tutorial in the supplemental material, to encourage biologists to adopt this approach with their data without prior bioinformatics skills. In this video, details of the network creation are described, such as how the epigenetic and the eQTL information related to the SNP activity are visualized in accordance with the genes that carry those SNPs. The workflow does not require advanced programming skills and researchers with basic computer knowledge can perform the analysis. At the same time, our workflow can be used by bioinformaticians to improve the steps and propose an alternative way to integrate data, in order to enhance the interpretation of non-coding SNPs.

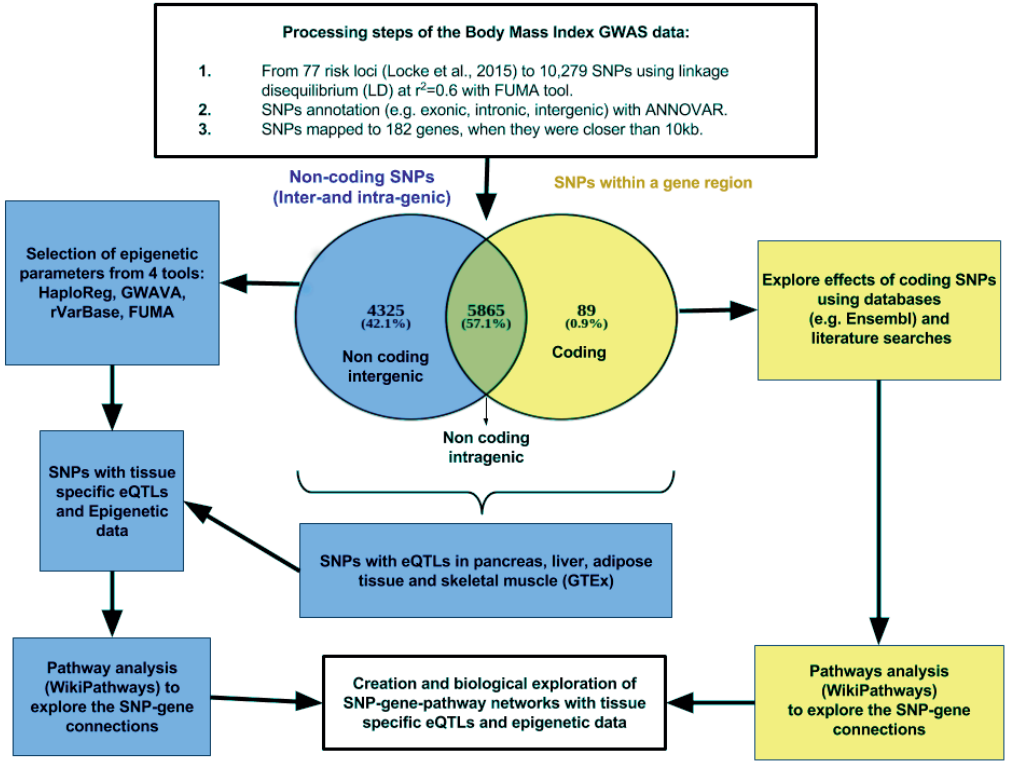

Fig. 5.1 Workflow of the data integration. The overview first lists the processing steps of the BMI GWAS data. Thereafter the workflow is divided into two parts. In yellow the pathway analysis of the 182 genes including at least one SNP located in the gene region of 10kbp up and downstream the UTR, is shown. In blue the eQTL and epigenetic data retrieved from different sources to better characterize the function of the non-coding SNPs are shown. Finally, these two parts are combined to create and explore the tissue specific genetic reference networks of SNPs associated with obesity.

The genetic reference networks of SNPs associated with obesity

The final output of our network approach consists of four tissue-centric networks that each include i) epigenetically active non-coding SNPs, that are SNPs in region with a property of an epigenetically active state, ii) genes in which those SNPs or other coding SNPs are located, iii) genes whose expression is influenced by a subset of the non-coding SNPs (eQTLs), and iv) biological pathways in which the gene products are known to be involved in. The four networks are publicly available for interactive visualization and further analysis in the Network Data Exchange (NDEx), an open-source platform to share, store and publish biological network data [42]. One reference network for each tissue is available
(adipose tissue: `http://www.ndexbio.org/#/network/ef2f83f4-4e0d-\`
`11e8-a4bf-0ac135e8bacfaccesskey=12d5dbe59f51578fa47ff424ce\`
`23aeb031176a93c14cf2ce64f8cfdca19e8`,
pancreas: `http://ndexbio.org/#/network/5955ecb3-125a-11e8-b9\`
`39-0ac135e8bacfaccesskey=5dc6d866e343f8300b3117e56187e5c2\`
`2c65ab919e01ddd9ec4f6ffa1541cd20`,
liver: `http://ndexbio.org/#/network/31f18b1d-125a-11e8-b939-\`
`0ac135e8bacf?accesskey=502a48125940b8003a5482543ab88c6ead\`
`be0528f3f56c9838bfe977f193a92a`,
skeletal muscle: `http://ndexbio.org/#/network/45350810-125a-11e\`
`8-b939-0ac135e8bacf?accesskey=0a3126fa320dc4f0635b3e96951\`
`81e41fdb99a1ad58de164d70aff0a2a918c9e`).
We designate these networks as genetic reference network of SNPs associated with obesity. The description of the node colors and node shapes used for the different entities, such as SNPs, genes and pathways, is reported in supplementary information (Figure S1). The full list of the identifiers and full names of the elements are available in the table displayed at the bottom of the network diagram. In this table, the information regarding the epigenetic activity of non-coding SNPs and eQTLs is also reported. The position of each entity in the four networks is kept in the same location, facilitating visual comparison of the differences and similarities of the nodes and their connections. Based on the number of epigenetically active non-coding SNPs, a clear variability is observable in all four networks. For example, a gene such as *HSD17B12* has epigenetically active SNPs in some tissues, but the number of those SNPs changes depending on the tissue. This confirms a well-known biological property of tissue specificity of regulatory elements. The novelty of this analysis is enabling the visual exploration of the tissue variation attributed to the SNPs and not the genes, as is commonly investigated. The epigenetic data refer to a specific genomic region such as histone modification marks and predicted open chromatin state [30] that represent enhancer, transcrip-

tion start site and other indicators of gene activity. The eQTL data retrieved from the Genotype-Tissue Expression (GTEx) project [23] via the FUMA tool [38] contains a SNP or its proxy that influences the expression of a gene highlighted in the network with light blue color. Because only tag SNPs are available, it is not possible to identify the specific causative SNP influencing the gene expression. Those SNPs in strong LD with the eQTL SNP could potentially also have that role. Table 5.1 summarizes the entities in the four genetic reference networks. The table shows that the number of genes influenced by eQTLs is significantly higher for adipose tissue and skeletal muscle. This discrepancy slightly decreases for the total number of genes with SNPs. On the other hand, the number of pathways detected in each network is quite similar with a range from 223 to 238 pathways per network.

Table 5.1 Summary of the entities present in the four tissue specific networks.he number of epigenetically active SNPs, total number of genes, eQTL genes and pathways are reported according to the four tissues explored: subcutaneous and visceral adipose tissue, liver, pancreas and skeletal muscle.

Tissue	**Num SNPs ascribed to histone marks and/or active chromatin regions**	**Num tot genes**	**Num eQTL genes**	**Num pathways**
Subcutaneous adipose	190	131	43	236
Skeletal muscle	170	112	40	238
Pancreas	126	104	20	226
Liver	66	99	8	223

A close examination of the type of processes depicted in the four tissue specific obesity networks indicates that the majority of pathways are shared in all tissues, with only adipose and skeletal muscle presenting some pathways that are unique to the respective tissue. Seven unique pathways are found in subcutaneous tissue: i) TP53 regulates transcription of cell cycle genes (WikiPathways identifier: WP3804), (ii) mitotic G1-G1/S phases (WP1858), iii) Inhibitor of DNA binding protein signaling pathway (WP53), iv) amino acid metabolism (WP3925), v) monoamine transport (WP727), vi) glycolysis and gluconeogenesis (WP534), and vii) TCA cycle and deficiency of pyruvate dehydrogenase complex(WP2453). In skeletal muscle, seven metabolism related pathways are uniquely linked to genes in the network: i) glycosaminoglycan metabolism (WP2743), ii) sphingolipid metabolism (WP2788), iii) aflatoxin B1 metabolism (WP1449), iv) benzene metabolism (WP3891), v) peptide hormone biosynthesis (WP2691), vi) the corticotropin-releasing hormone signaling pathway (WP2355), vii) regulation of toll-like receptor signaling pathway (WP1449). All other pathways are present in at least two tissue reference networks, and those pathways reflect biological processes well known in the pathophysiology of obesity, and may indicate avenues of crosstalk between tissues. One example is the signaling pathway of the well-described key regulator of metabolism, AMP-activated protein kinase (*AMPK*)[25] and its targets (WP1403, WP2748). Importantly, two *AMPK* target pathways are also present, i.e. fatty acid oxidation (WP1817) and glucose transport (WP1935). Interestingly, the signaling pathway of an *AMPK* activator, leptin (WP2034)[40] is also among the shared pathways. Leptin is already known as an important regulator of energy homeostasis [18, 34], but in our networks it is not highlighted. Interestingly, *SH2B1* and *NCOA1* genes are displayed by our data integration. These genes play a role in the downstream part of the leptin pathway, mediating the formation of complexes that either activate genes like *IRS1* and *IRS2* involved in insulin signaling for *SH2B1*, or act as chromatin remodelers and recruiters of general transcription factors, ending in controlling the energy balance between white and brown adipose tissue, for *NCOA1*. In the skeletal muscle network, both *SH2B1* and *NCOA1* show epigenetically active SNPs (rs11864107, rs7359397 for *SH2B1* and rs9309308 for *NCOA1*). In addition, in the adipose tissue network the *SH2B1* gene is an eQTL gene. This is an example of how reference networks with multiple data integrated can be used to investigate further the role of SNPs that have a regulatory influence on downstream genes in relevant disease pathway such as the leptin signaling. Importantly, other relevant pathways are those related to inflammation, including interleukin signaling pathways (WP364, WP3796, WP49, WP127), interferon signaling pathways (WP619, WP585), *TNF* alpha signaling pathway (WP231, WP3398), B cell signaling pathways (WP23, WP2746), T cell signaling pathways (WP68, WP3863) and the pathway related to development of macrophage subsets (WP3892).

In general, we see a majority of well-studied and relevant obesity related processes in

our tissue-centric networks, including energy metabolism (WP1831, WP1541), cholesterol metabolism (WP197 and WP2011), insulin signaling (WP481), circadian clock (WP3355, WP3594, WP1797), adipogenesis (WP236), neuronal activity related pathways (WP2380, WP2754, WP1871)as well as disease specific pathways (WP3407).
Among all genes defined by BMI SNPs, 64 genes contain SNPs that map to regions of active chromatin in subcutaneous and visceral adipose tissue, skeletal muscle, pancreas and liver (Figure 5.2). Skeletal muscle and adipose tissue contain the most genes with SNPs mapping within active chromatin regions, with 13 and 20, respectively. In contrast, liver and pancreas have one and three unique genes, respectively. The discrepancy between the number of genes in skeletal muscle and adipose tissue versus those in liver and pancreas was previously observed in table 5.1 for the number of eQTL genes in the respective tissues. Details regarding the 64 genes are presented in supplementary information (Table S1) including the common gene name, the Ensembl identifier, gene function and an indication of the tissue in which the gene is present. Observing the function of the genes listed in Table S1, there are several genes typically associated with or changing expression in obesity. Genes such as *NEGR1, MTCH2,*[24] and *TCF7L2*, already have been identified as important in several processes that affect obesity. Interestingly, only the *BCKDK* gene is shared by the four tissues and in all four a SNP (rs749767) located in an enhancer region is observed. The gene encodes a component of the branched-chain alpha-ketoacid dehydrogenase complex (*BCKD*), an important regulator of branched-chain amino acid (BCAA) metabolism. An increase in circulating BCAAs in obese individuals has been found to be associated with the development of insulin resistance and T2DM [44]. In an animal model, it was shown that increasing *BCKDK* protein levels led to an increase of BCAAs in plasma.

Interpreting the epigenetic activity of non-coding SNPs and their influence on gene

expression In order to interpret the eQTLs and the epigenetic variability of the BMI SNPs identified in the networks, the data related to two genes are reported as examples (Figure 5.3). The first gene is *HSD17B12*, an enzyme (E.C. 1.1.1.330) that catalyzes the second of four reactions of the long-chain fatty acid elongation cycle, allowing the addition of two carbons to the chain of long- and very long-chain fatty acids (VLCFAs) per cycle. Thereby, it may participate in the production of VLCFAs of different chain lengths that are involved in numerous biological processes as precursors of membrane lipids and lipid mediators. The protein (E.C. 1.1.1.62) also may catalyze the transformation of estrone (E1) into estradiol (E2) and participate in estrogen formation [20]. In the genetic reference network of SNPs associated with obesity, the gene is linked to three different pathways related to fatty acid, steroid hormone and vitamin D metabolism (WP1817, WP2749 and WP3836). In three of the tissue networks, epigenetically active

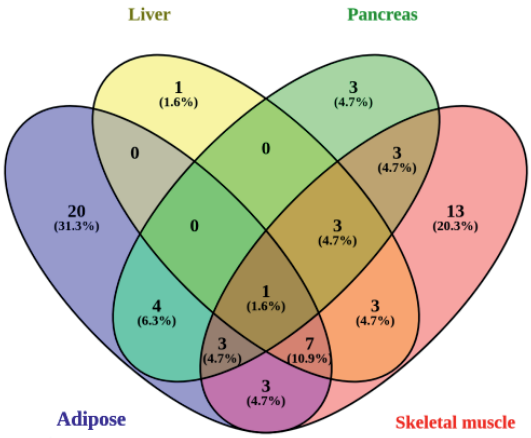

Fig. 5.2 Venn diagram of the genes with epigenetically active SNPs and eQTL data. The diagram presents the number of genes with non-coding SNPs showing epigenetic and eQTL data, distributed in the four tissues: subcutaneous and visceral adipose tissue, skeletal muscle, liver, and pancreas.

non-coding SNPs are associated to *HSD17B12*: one in liver, six in skeletal muscle and sixteen in pancreas. Close inspection of the sixteen intronic SNPs present in the pancreas network shows that they map within histone modification marks in two pancreatic cell lines (RoadMap cell code E87 and E98), and these marks identify an enhancer region. The SNPs are located in an interactive chromatin region with transcription factor (TF) binding sites. Several types of motifs are also identified in ten of the sixteen SNPs. In addition, the gene shows significant BMI association with the coding missense SNP rs11555762. Regarding the eQTL data, *HSD17B12* is significantly expressed in all four tissues and its expression is influenced by the tag SNP rs2176598 (or those SNPs in LD with it). From these data, it appears that some variants of *HSD17B12* serve to regulate its own expression and that one BMI SNP in LD with a tag SNP is responsible for this effect. Regarding the interpretation of these epigenetic and eQTL data, it is important to consider that the results refer to a non-disease condition, because the histone modification and mRNA expression measurements were performed on samples from individuals with diverse phenotypes, and not in an obesity case-control scenario. In the case of *HSD17B12*, the gene is not one of the well-known obesity genes, but it does contain one coding SNP and many non-coding SNPs that are associated with BMI. Usually, the effect of the coding SNPs, especially missense SNPs, is more commonly explored with further laboratory experiments. That coding missense SNP rs11555762 has been

demonstrated to be associated with breast cancer, and the effect has been ascribed to its alternative role in estrone conversion [9]. The exploration of the other non-coding SNPs, even if the data are related to a baseline condition, broadens the general picture and the meaning of the BMI SNPs located in the gene. In addition, it has been shown that *HSD17B12* was associated with obesity in a DNA methylation experiment, invoking its role in the fatty acid elongation pathway [9]. Our data, on *HSD17B12* show that BMI variants are located in active regulatory regions of genes, and this supports published findings and the hypothesis that non-coding SNPs can act in a regulatory function.

The other example is *HMGCR*, better known as an obese gene [32], encodes HMG-CoA reductase (E.C. 1.1.1.34), the rate-limiting enzyme for cholesterol synthesis that is regulated via a negative feedback mechanism mediated by sterols and non-sterol metabolites derived from mevalonate. The enzyme is suppressed by cholesterol derived from the internalization and degradation of low-density lipoprotein (*LDL*) via the *LDL* receptor. Competitive inhibitors of the reductase induce the expression of *LDL* receptors in the liver, which in turn increases the catabolism of plasma *LDL* and lowers the plasma concentration of cholesterol, an important determinant of atherosclerosis [20]. In the tissue-centric networks, nine pathways are linked to *HMGCR*: i) statin pathway (WP430), ii) regulation of lipid metabolism by peroxisome proliferator-activated receptor alpha (WP2797), iii) activation of gene expression by SREBF (WP2706), iv) *SREBF* and miR33 in cholesterol and lipid homeostasis (WP2011), v) integrated breast cancer pathway (WP1984), vi) sterol regulatory element-binding proteins signaling (WP1982), vii) cholesterol biosynthesis (WP197), viii) target of rapamycin signaling (WP1471), ix) *AMPK* signaling (WP1403). The epigenetic data indicate histone modification marks in two tissues in a genomic region containing the SNPs: one SNP for adipose tissue and four SNPs for skeletal muscle. In skeletal muscle, the four intronic SNPs map to a region containing histone marks retrieved from skeletal muscle cell data (cell code E108, E0107, E120, E121) indicating that those SNPs are located in an enhancer region. In addition, in this case the chromatin is active, TF binding sites are present, and location for two of four SNPs show different types of motifs (e.g. at SNP rs6453131 there are four motifs: HP1sitefactor, Pbx-1_4, SIX5_disc3, SIX5_disc4). The eQTL data indicate an mRNA-SNP association only in skeletal muscle with tag SNP rs6871667 (or those in LD with it) influencing gene expression. However, expression is not observed for *HMGCR* itself, but for the hexosaminidase B (*HEXB*) gene located 613,682 bp upstream of *HMGCR* gene. *HEXB* (E.C. 3.2.1.52) encodes the beta subunit of the lysosomal enzyme beta-hexosaminidase that, together with the cofactor GM2 activator protein, catalyzes the degradation of the ganglioside GM2, and other molecules containing terminal N-acetyl hexosamines [20]. Hence, *HEXB* becomes an additional entry in the gene list within this network-based analysis, because no BMI SNP was located in

this region and because some BMI SNPs support eQTLs for its expression. Both results relate to the epigenetic activity of *HMCGR* and the expression of *HEXB* in pathways involved in obesity pathophysiology, and strengthen previous findings [28]. For *HMCGR*, Meaney et al. [16] reported epigenetic changes in the expression of this and other genes involved in cholesterol activity. In addition, Knebel et al. [33] found in an obese mouse model that *HEXB* was significantly expressed as a protein in adipose tissue, but its role was not specifically investigated.

In summary, the *HSD17B12* example highlights the importance of non-coding BMI SNPs that can have a regulatory function influencing the expression of the gene in which they are located, and contrasts with a coding missense BMI SNP affecting the encoded protein. The *HMGCR-HEXB* example suggests that the non-coding BMI SNPs influence expression of a gene other than the one in which the epigenetic activity is found, thereby detecting genes previously not considered in the analysis.

Discussion

The genetic reference networks of SNPs associated with obesity are maps of SNPs, genes and pathways that can be used in different ways by different stakeholders interested in obesity and personalized treatments. Experts in the field of obesity can explore the networks to confirm results or generate novel hypothesis related to the functional role of BMI SNPs and their possible effect on gene regulation by influencing epigenetic marks. Indeed, we identified pathways from genes that carry SNPs associated with BMI. Our results are in line with the pathway output reported in the original study [12] that applied a gene set enrichment analysis for the BMI-associated loci found in order to identify relevant BMI biological pathways and gene sets. Moreover, the genetic reference network contains information on epigenetic marks, e.g. DNA methylation and histone modifications, which enables to explore their possible involvement in the effect of the SNP on gene expression. For example, the methylation patterns of the circadian pathway genes *CLOCK, BMAL1* and *PER2* were found to be associated with BMI in an interventional study [14]. Although that study was not designed to indicate direct causal effects to the obese phenotype, it was confirmed that epigenetic changes of the promoters at several human clock genes were altered. In our analysis *NCOA1* is the only gene linked to the circadian pathways (WP3355, WP3594, WP1797), and only in our skeletal muscle network it is shown that the downstream SNP rs9309308 is located in an enhancer region containing several transcription factor binding motifs. Such an observation both supports the work of Milagro et al.[14] and adds another potential gene in the investigation of the epigenetic changes that occur in genes involved in the same circadian pathway. Different methods for GWAS data integration using networks have been presented previously [7, 22, 26, 29] but, to our knowledge, our analysis is

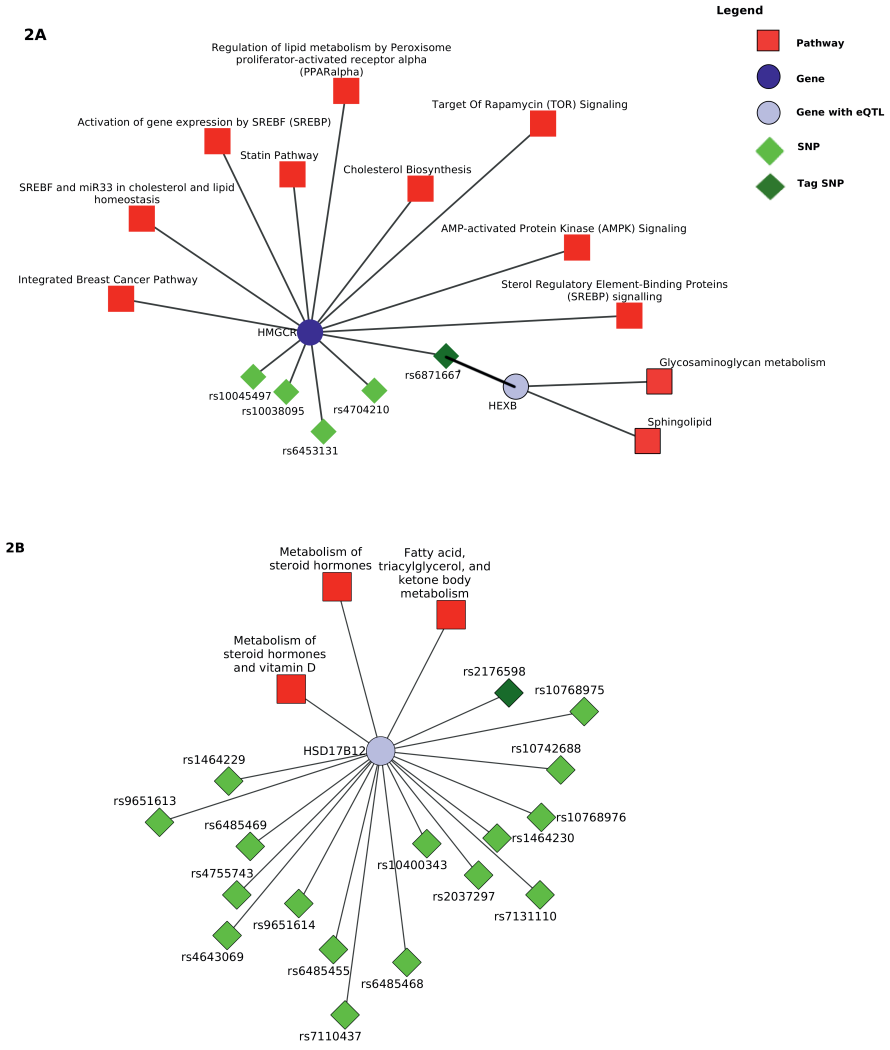

Fig. 5.3 Network detail of the HSD17B12 and HMGCR examples. 2A: SNPs (green nodes) and pathways (blue nodes) related to the HSD17B12 example are shown. The gene node is colored in light purple indicating that its expression is influenced by one of its GWS SNP. **2B:** SNPs and pathways linked to HMGCR and HEXB are presented. HEXB is colored in light purple because it is detected as an eQTL, for which expression is regulated by one of the SNPs located in HMGCR. This relation is represented by a thick black line connecting HEXB and the tag SNP of HMGCR.

a unique example of identifying genetic reference networks that present tissue specific interactions regarding the SNPs within active chromatin region, SNPs influencing gene expression, and biological pathways in obesity. Experts in obesity also can examine

these results in order to plan and improve new or different omics studies. For example, using DNA and RNA sequencing data in this type of study would an interesting alternative to improve the detection of the causative SNPs. However, nowadays such data are not easily and freely accessible, and the majority of the GWAS genotype-phenotype association tests are performed with genotypes taken from SNP arrays [24]. Moreover, the results obtained refer to the role of SNPs in individuals with a high BMI. Obesity occurs also in cases with a normal BMI, but with other parameters altered such as the level of adiposity or the waist circumference. The SNP scenario would most likely be different if those other traits were considered. The data re-analyzed in this study are related to individuals that share the same BMI parameters, but the original study [12] does not provide specific phenotypic descriptions of these individuals. This missing information would have allowed us to distinguish obese patients in phenotypic subgroups with different comorbidities.

In addition, such genetic reference networks could prove to be useful for clinical geneticists involved in precision medicine. Indeed, for those patients for whom SNP genotyping data are available, the health care team can determine susceptibility to certain diseases, by consulting the reference networks. Exploring those SNPs present both in the network and in the patient can assist interpretation of the relevance of the patient's alleles, linking them to the gene and the functional context in which they are involved. For example, the occurrence of several genotyped SNPs from the patient that indicate presence of risk or effect allele, that occur in the same or related pathways, can prompt the health care team to evaluate if those processes, in relation to the specific tissue, are relevant to the patient's condition. Precision medicine in obesity is a complex challenge because diverse factors are involved. In this study we present the integration of genetic and epigenetic factors, but also comorbidities and environmental, socio-economics aspects need to be considered to obtain the overall picture [37].

Finally, we made interactive visualization of the genetic reference networks publicly available at NDEx with a video tutorial. These are important sources to: i) encourage the application of our data integration method to other datasets, ii) reproduce our analysis step-by-step in an easy to follow manner, iii) further investigate these obesity results.

Methods

Dataset description

A publicly available GWAS summary statistics of BMI meta-analysis based on 322,154 European descent individuals [12] was obtained from here. In the study, 95 independent lead SNPs associated with BMI were detected across 77 genomic risk loci. The Func-

tional Mapping and Annotation tool (FUMA, http://fuma.ctglab.nl/)[38] was used to extend the list of the 95 lead SNPs with variants from the 1000 Genomes project European (CEU) population that were in LD at r2 above 0.6, resulting in a list of 10,367 SNPs. FUMA also annotates the SNPs with allele-specific consequences (*e.g.* intron, exon, splicing, etc.). These SNPs are mapped to 182 genes with 10 kbp window from the UTRs. Of all 10,367 SNPs, 10,190 SNPs are non-coding and 5,954 SNPs are located in the gene region (10 Kbp up and downstream the gene UTR). All SNPs are reported in Table S2 of the supplemental material with their rs-number, their functional consequences, and genes (symbol and Ensembl identifier) if the SNP was mapped to a gene.

Dataset analysis

The approach presented in this study enables network based interpretation of comprehensive genetics data derived from FUMA. The data analysis was divided in two parts as shown in the workflow of Figure 5.1. In the yellow right side, a pathway analysis was performed for the 182 genes in which at least one SNP is located. The WikiPathways human curated collection (www.wikipathways.org, April 2017 release) [2] with 747 pathways was used to identify the biological processes in which at least one of the 182 genes was present. In the blue left side of Figure 5.1, the eQTL and the epigenetic data were retrieved from different sources to better characterize the function of the non-coding SNPs. Tissue specific eQTL in subcutaneous and visceral adipose tissue, skeletal muscle, pancreas, and liver were obtained from the GTEx portal v6 [23] and annotated by FUMA on 21/02/2017. The Table S3 in the supplemental material reports the gene symbols and Ensembl identifiers of the genes, their eQTLs, the tissues, and the tag SNP(s) genotyped in relation to the gene expression data. The epigenetic data information are: chromatin state with cell type codes, five histone modification marks (H3K4me3, H3K4me1, H3K27me3, H3K9me3 and H3K36me3) with cell type codes, regulatory elements related to regions of interactive chromatin, TF binding sites, regulatory motifs, and scores indicating SNP functionality, were retrieved from four online sources: FUMA [38], GWAVA [19], HaploReg and rVarBase [21]. The main epigenetic data sources of these tools are Encode and RoadMap Epigenomics project [30]. In the selection of the epigenetic parameters that show cell type, the following codes were considered from the RoadMap Epigenomics project: E023, E025 and E063 for adipose tissue, E107, E108, E120, and E121 for skeletal muscle, E87 and E98 for pancreas, and E066 and E118 for liver. An explanation of the cell type code is accessible https://github.com/Bioconductor/BioC2015Introduction/blob/master/inst/extdata/epi_metadata.txt. The four tools were queried with the non-coding SNPs, and for each tool a table

was obtained, from which the epigenetic information previously listed was selected. Table S4 shows the combination of the epigenetic outputs from the four online tools related to the entire list of the non-coding SNPs with indications relating to the tool from where the information was retrieved.
Cytoscape v3.6.0 [13], an open-source and extendable network visualization and analysis tool, was used to visualize the four tissue-specific SNP-gene-pathway networks named genetic reference networks, where epigenetically active SNPs and eQTL genes are also visualized. The Cytoscape app DyNet [35] was used to keep the same node location of SNPs, genes and pathways in all four networks, facilitating network comparison. In addition, the Cytoscape app CyNDEX-2 was used to upload the four networks on NDEx website21, this step enables to share them with the network community. Finally, a video provided in the supplemental material was created with Screencast-O-Matic (`https://screencast-o-matic.com/`) to explain the network generation workflow and network visualization.

Acknowledgement

Any opinions, findings, conclusion, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the U.S. Department of Agriculture. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The USDA is an equal opportunity provider and employer.

References

[1] Albert-László Barabási. Network Medicine From Obesity to the Diseasome. *New England Journal of Medicine*, 357(4):404–407, jul 2007.

[2] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, and et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, jan 2018.

[3] Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[4] Ben van Ommen, Suzan Wopereis, Pepijn van Empelen, Hilde M. van Keulen, Wilma Otten, and et al. From Diabetes Care to Diabetes CureThe Integration of Systems Biology, eHealth, and Behavioral Change. *Frontiers in Endocrinology*, 8:381, jan 2018.

[5] J M Moreno-Navarrete, F Ortega, M Serino, E Luche, A Waget, G Pardo, J Salvador, and et al. Circulating lipopolysaccharide-binding protein (LBP) as a marker of obesity-related insulin resistance. *International Journal of Obesity*, 36(11):1442–1449, nov 2012.

[6] George Osei-Assibey, Smita Dick, Jennie Macdiarmid, Sean Semple, John J Reilly, and et al. The influence of the food environment on overweight and obesity in young children: a systematic review. *BMJ open*, 2(6):e001538, jan 2012.

[7] Nirmala Akula, Ancha Baranova, Donald Seto, Jeffrey Solka, Michael A. Nalls, and et al. A network-based approach to prioritize results from genome-wide association studies, 2011.

[8] Angela C. Estampador and Paul W. Franks. Precision Medicine in Obesity and Type 2 Diabetes: The Relevance of Early-Life Exposures. *Clinical Chemistry*, 141:clinchem.2017.273540, 2017.

[9] Marie Plourde, Alexandra Ferland, Penny Soucy, Yosr Hamdi, Martine Tranchant, and et al. Analysis of 17β-hydroxysteroid dehydrogenase types 5, 7, and 12 genetic sequence variants in breast cancer cases from French Canadian Families with high risk of breast and ovarian cancer. *The Journal of Steroid Biochemistry and Molecular Biology*, 116(3-5):134–153, 2009.

[10] Elisa Cirillo, Laurence D. Parnell, and Chris T. Evelo. A Review of Pathway-Based Analysis Tools That Visualize Genetic Variants. *Frontiers in Genetics*, 8:174, nov 2017.

[11] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I Mccarthy, and et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101:5–22, 2017.

[12] Adam E. Locke, Bratati Kahali, Sonja I. Berndt, Anne E. Justice, Tune H. Pers, and et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[13] Paul Shannon, Andrew Markiel, 2 Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, and et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, (13):2498–2504, 2003.

[14] Fermín I. Milagro, Purificación Gómez-Abellán, Javier Campión, J. Alfredo Martínez, and et al. CLOCK, PER2 and BMAL1 DNA methylation: Association with obesity and metabolic syndrome characteristics and monounsaturated fat intake. *Chronobiology International*, 29(9):1180–1194, 2012.

[15] Yu Gyoung Tak and Peggy J. Farnham. Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics and Chromatin*, 8(1):1–18, 2015.

[16] Steve Meaney. Epigenetic regulation of cholesterol homeostasis. *Frontiers in Genetics*, 5(AUG):1–10, 2014.

[17] M. A. Papas, A. J. Alberg, R. Ewing, K. J. Helzlsouer, T. L. Gary, and A. C. Klassen. The Built Environment and Obesity. *Epidemiologic Reviews*, 29(1):129–143, may 2007.

[18] Gema Frühbeck and Javier Gómez-Ambrosi. Rationale for the existence of additional adipostatic hormones. *The FASEB Journal*, 15(11):1996–2006, sep 2001.

[19] Graham R.S. Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nature Methods*, 11(3):294–296, 2014.

[20] D. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet. GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics*, 1997.

[21] Liyuan Guo, Yang Du, Susu Qu, and Jing Wang. rVarBase: An updated database for regulatory features of human variants. *Nucleic Acids Research*, 44(D1):D888–D893, 2016.

[22] Li Liu, Jing Lei, and Kathryn Roeder. Network assisted analysis to reveal the genetic basis of autism. *Annals of Applied Statistics*, 9(3):1571–1600, 2015.

[23] Francois Aguet, Andrew A Brown, Stephane Castel, Joe R Davis, Pejman Mohammadi, and et al. Local genetic effects on gene expression across 44 human tissues. *bioRxiv*, 2016.

[24] Rodrigo Alonso, Magdalena Farías, Veronica Alvarez, and Ada Cuevas. The Genetics of Obesity. *Translational Cardiometabolic Genomic Medicine*, (October):161–177, 2015.

[25] Daniel Garcia and Reuben J. Shaw. AMPK: Mechanisms of Cellular Energy Sensing and Restoration of Metabolic Balance. *Molecular Cell*, 66(6):789–800, jun 2017.

[26] Mark D M Leiserson, Jonathan V. Eldridge, Sohini Ramachandran, and Benjamin J. Raphael. Network analysis of GWAS data. *Current Opinion in Genetics and Development*, 23(6):602–610, 2013.

[27] Adem Yesuf Dawed, Kaixin Zhou, and Ewan Robert Pearson. Pharmacogenetics in type 2 diabetes: influence on response to oral hypoglycemic agents. *Pharmacogenomics and personalized medicine*, 9:17–29, 2016.

[28] Peter Arner, Indranil Sinha, Anders Thorell, Mikael Rydén, and et al. Dahlman-Wright, Karin and.

[29] Louise B. Thingholm, Lars Andersen, Enes Makalic, Melissa C. Southey, Mads Thomassen, and et al. Strategies for integrated analysis of genetic, epigenetic, and gene expression variation in cancer: Addressing the challenges. *Frontiers in Genetics*, 7(FEB):1–13, 2016.

[30] Inderpreet Sur and Jussi Taipale. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Reviews Cancer*, 28(10):1045–1048, 2016.

[31] Valeriya Lyssenko, Cristina Bianchi, and Stefano Del Prato. Personalized Therapy by Phenotype and Genotype. *Diabetes care*, 39 Suppl 2(Supplement 2):S127–36, aug 2016.

[32] Masahiro Yamasaki, Shinya Hasegawa, Masahiko Imai, Noriko Takahashi, and Tetsuya Fukui. High-fat diet-induced obesity stimulates ketone body utilization in osteoclasts of the mouse bone. *Biochemical and Biophysical Research Communications*, 473(2):654–661, apr 2016.

[33] Birgit Knebel, Simon Goeddeke, Gereon Poschmann, Daniel F. Markgraf, Sylvia Jacob, and et al. Novel insights into the adipokinome of obese and obese/diabetic mouse models. *International Journal of Molecular Sciences*, 18(9), 2017.

[34] Javier Gómez-Ambrosi, Javier Salvador, Jose A Páramo, Josune Orbe, Jokin de Irala, and et al. Involvement of leptin in the association between percentage of body fat and cardiovascular risk factors. *Clinical Biochemistry*, 35(4):315–320, jun 2002.

[35] Ivan H. Goenawan, Kenneth Bryan, and David J. Lynn. DyNet: visualization and analysis of dynamic molecular interaction networks. *Bioinformatics*, 32(17):2713–2715, sep 2016.

[36] Mònica Sabater, Jose M. Moreno-Navarrete, Francisco José Ortega, Gerard Pardo, Javier Salvador, and et al. Circulating Pigment Epithelium-Derived Factor Levels Are Associated with Insulin Resistance and Decrease after Weight Loss. *The Journal of Clinical Endocrinology and Metabolism*, 95(10):4720–4728, oct 2010.

[37] Gema Frühbeck, Dimitrios N Kiortsis, and Victoria Catalán. Precision medicine: diagnosis and management of obesity. *The lancet. Diabetes and endocrinology*, 6(3):164–166, mar 2018.

[38] Kyoko Watanabe, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, 8(1):1826, dec 2017.

[39] Editorial. Genome variation in precision medicine. *Nature Genetics*, 48(7):701, 2016.

[40] Hyeong-Kyu Park and Rexford S Ahima. Leptin signaling. *F1000prime reports*, 6:73, 2014.

[41] Aharon Brodie, Johnathan Roy Azaria, and Yanay Ofran. How far from the SNP may the causative genes be? *Nucleic Acids Research*, 44(13):6046–6054, 2016.

[42] Dexter Pratt, Jing Chen, David Welker, Ricardo Rivas, Rudolf Pillich, and et al. NDEx, the Network Data Exchange. *Cell Systems*, 1(4):302–305, oct 2015.

[43] Sujoy Ghosh, Juan C. Vivar, Mark A. Sarzynski, Yun Ju Sung, James A. Timmons, and et al. Integrative pathway analysis of a genome-wide association study of V o _{2max} response to exercise training. *Journal of Applied Physiology*, 115(9):1343–1359, nov 2013.

[44] M. D. Adams. The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195, mar 2000.

[45] Blanca M Herrera and Cecilia M Lindgren. The genetics of obesity. *Current diabetes reports*, 10(6):498–505, dec 2010.

[46] D S Sinasac, J D Riordan, S H Spiezio, B S Yandell, and J H Nadeau. Genetic control of obesity, glucose homeostasis, dyslipidemia and fatty liver in a mouse model of diet-induced metabolic syndrome. *International Journal of Obesity*, 40:346–355, 2015.

[47] Jill M Norris and Stephen S Rich. Genetics of glucose homeostasis: implications for insulin resistance and metabolic syndrome. *Arteriosclerosis, thrombosis, and vascular biology*, 32(9):2091–6, sep 2012.

CHAPTER 6

Biological pathways leading from *ANGPTL8* to Diabetes Mellitus - A co-expression network based analysis

Amnah Siddiqa[1,2], Elisa Cirillo[2], Samar H. Tareen[3], Amjad Ali[4], Martina Kutmon[2,3], Lars M.Eijssen[2], Jamil Ahmad[1], Chris T. Evelo[2,3], Susan L. Coort[2]

1 Research Center for Modeling and Simulation, National University of Sciences and Technology, Pakistan.
2 Department of Bioinformatics - BiGCaT, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands.
3 Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands.
4 Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology, Pakistan.

Submitted to: Frontiers in Physiology

Abstract

Angiopoietin like protein 8 (*ANGPTL8*) is a newly identified hormone with unique nature due to its ability to regulate both glucose and lipid metabolic pathways. It is characterized as an important molecular player of insulin induced nutrient storage and utilization pathway during fasting to re-feeding metabolic transition. Several studies have contributed to increase our knowledge regarding its function and mechanism of action. Moreover, its altered expression levels have been observed in Insulin Resistance, Diabetes Mellitus (Types I and II) and Non Alcohlic Fatty Liver Disease emphasizing its assessment as a drug target. However, there is still a great deal of information that remains to be investigated including its associated biological processes, partner proteins in these processes, its regulators and its association with metabolic pathogenesis. In the current study, the analysis of a transcriptomic data set was performed for functional assessment of *ANGPTL8* in liver. Weighted Gene Co-expression Network Analysis coupled with pathway analysis tools was performed to identify genes that are significantly co-expressed with *ANGPTL8* in liver and investigate their presence in biological pathways. Gene ontology term enrichment analysis was performed to select the gene ontology classes that over-represent the hepatic *ANGPTL8*-co-expressed genes. Moreover, the presence of diabetes linked SNPs within the genes set co-expressed with *ANGPTL8* was investigated. The co-expressed genes of *ANGPTL8* identified in this study (n=460) provides narrowed down list of molecular targets which are either co-regulated with it and/or might be regulation partners at different levels of interaction. These results are coherent with previously demonstrated roles and regulators of *ANGPTL8*. Specifically, thirteen co-expressed genes (*MAPK8, CYP3A4, PIK3R2, PIK3R4,PRKAB2, G6PC, MAP3K11, FLOT1, PIK3C2G, SHC1, SLC16A2*, and *RAPGEF1)* are also present in the literature curated pathway of *ANGPTL8* (WP3915[1]). Moreover, the gene-SNP analysis of highly associated biological processes with *ANGPTL8* revealed significant genetic signals associated to Diabetes Mellitus and similar phenotypic traits. It provides meaningful insights on the influencing genes involved and co-expressed in these pathways. Findings of this study have implications in functional characterization of *ANGPTL8* with emphasis on the identified genes and pathways and their possible involvement in the pathogenesis of Diabetes Mellitus and Insulin Resistance.

Introduction

Diabetes Mellitus (DM) is a pathological condition which is often characterized by hyperinsulinemia and hyperglycemia and has become a global health challenge for both developed and developing countries [33]. It is estimated to affect 642 million peo-

[1]https://www.wikipathways.org/index.php/Pathway:WP3915

ple by 2040 according to the International Diabetes Federation [34]. The underlying pathogenic mechanisms are well studied and encompass deregulated glucose and lipid homeostasis involving inter-organ crosstalk of substrates and hormones. However, the suboptimal effectiveness of current diabetic therapies to control pathological glycemic conditions necessitates the identification of novel molecular players involved in regulation of lipid and glucose homeostasis for the development of better pharmacological interventions [32]. An emerging novel molecular target for treatment of DM and related metabolic disorders is Angiopoietin like protein 8 (*ANGPTL8*) due to its unique nature in regulating both lipid and glucose metabolism [31]. Recent studies have demonstrated the upregulation of *ANGPTL8* gene expression in various related metabolic disorders including insulin resistance, obesity, DM (type I and II), Metabolic Syndrome, Non Alcoholic Fatty Liver Disease (NAFLD) and Hepatocellular Carcinoma (HCC) emphasizing its assessment as a potential drug target [41, 44, 45, 65].

ANGPTL8 is a newly identified member of angiopoietin like protein (ANGPTL) family and is also known as lipasin, refeeding induced in fat and liver (*RIFL*), betatrophin, C19orf80 and TD26 [39, 43, 53]. It is induced upon feeding in liver and adipose tissue (both white adipose tissue (WAT) and brown adipose tissue (BAT)) whereas fasting suppresses its expression [39, 43, 53]. It has been recognized as one of the essential molecular players involved in the metabolic transition of fasting to re-feeding through both *in vivo* and *in vitro* studies [37, 39, 53]. It has been demonstrated to play a role in triglyceride (TG) metabolism by regulating postprandial lipid traffic via inhibition of lipoprotein lipase (LPL) activity [37–39, 53]. *LPL* is a hydrolytic enzyme which generates free fatty acids (FFA) from hydrolysis of TGs for subsequent uptake by heart, skeletal muscles and WAT. According to the molecular mechanism demonstrated by [37], *ANGPTL8* inhibits the postprandial *LPL* activity of cardiac and skeletal muscles which allows the uptake of FFA by WAT for storage. On the other hand, fasting decreases the expression of *ANGPTL8* and in turn the *LPL* activity in cardiac and skeletal muscles which allows the uptake of FFA by them for energy expenditure. Thus, *ANGPTL8* exhibits a significant role in lipid metabolism being a part of lipid partitioning machinery according to nutritional levels. *ANGPTL8* has also been demonstrated to play role in other lipid metabolic pathways including adipogenesis and autophagy [36, 53]. Its role in glucose metabolism was reported in several studies individually [41, 43, 44]. However, Guo and colleagues demonstrated the mechanism of *ANGPTL8* mediated glucose regulation via *AKT*/*GSK3beta* and *AKT*/*FOXO* arms of insulin signaling pathway [44]. *AKT*/*GSK3beta* and *AKT*/*FOXO* signaling regulates the activation of glycogen synthesis and inhibition of gluconeogenesis, respectively.

Recently, we have designed and published an up-to-date literature curated pathway of ANGPTL8 regulation based on its reported regulators and pathways in liver [31]. The

pathway model is available on WikiPathways[2] [31]. The pathway allows to clearly visualize the regulatory interactions between different regulators of *ANGPTL8* including insulin in presence of glucose, thyroid hormone receptors (*THR-alph/beta*), sterol regulatory element-binding protein (*SREBPs*), carbohydrate response element binding protein (ChREBP), mitogen-activated protein kinases (*MAPKs*), and 5' AMP-activated protein kinase (*AMPK*) for its regulation (reviewed in [31]). Moreover, the presence of *ANGPTL8* can be visualized in a broader spectrum with respect to other linked pathways including insulin signaling pathway [41, 43, 44, 53], postprandial TG partitioning [37], adipogenesis [53], autophagy [36], and CD45+ hematopoietic-derived cell proliferation [35].

Despite new insights, there is still a great deal of information that remains to be investigated regarding *ANGPTL8*'s functions, regulation and physiological mechanism of action. For example, different studies have indicated the biological processes (such as autophagy, adipogenesis and CD45+ hematopoietic-derived cell proliferation) in which *ANGPTL8* is involved but the underlying mechanism of action, associated receptors and signaling molecules (genes/proteins/metabolites) still remain elusive [35, 36, 53]. Besides, the role of *ANGPTL8* might not be limited to the already associated biological processes and transcription factors and hence needs further investigation from this point of view as well. Moreover, already identified transcription factors of *ANGPTL8* and their coordinated role in initiating its expression during refeeding/fasting metabolic transition also needs further investigation because they have been reported in individual studies.

Briefly, the investigation of predominant functional roles, biological processes, and associated signaling molecules (receptors/cofactors/genes) of *ANGPTL8* is of immense importance for its assessment as a molecular target for the treatment of DM and related metabolic disorders. Therefore, in the current study, we specifically aimed to identify the significantly co-expressed genes with *ANGPTL8* and their presence in known pathways (present in WikiPathways), in order to gain mechanistic insights regarding its function. The genes exhibiting similar expression pattern have been demonstrated to be involved in similar functions and/or biological processes besides being co-regulated [46, 66–68]. We further explored the co-expressed genes present in a selection of identified pathways to scrutinize their significant Single Nucleotide Polymorphism (SNP) based association with DM and/or other metabolic disorders. The investigation of the effect of SNPs associated with DM can enhance and redefine the gene role in the identified pathways [4].

Weighted gene co-expression network analysis (WGCNA) is an established method for identification of modules (cluster of genes with similar co-expression patterns) of biologically related genes [46–48]. In the present study, we performed WGCNA uti-

[2]https://www.wikipathways.org/index.php/Pathway:WP3915

lizing a human liver transcriptomics data set retrieved from gene expression omnibus (GEO). The selection of the gene expression data is based on the facts that *ANGPTL8* is a predominantly liver expressed gene in humans besides being up-regulated in insulin resistance [42–44], obesity [41] and DM type II [45]. Overall, the data set consisted of 21 human liver samples from lean, obese and type II diabetic patients. WGCNA [46] coupled with pathways analysis [1, 8, 9] as demonstrated in sections below was performed to: i) identify the genes that are significantly co-expressed with *ANGPTL8* in liver, ii) select Gene Ontology classes that over-represent the hepatic *ANGPTL8*-co-expressed genes, iii) identify biological pathways in which the hepatic *ANGPTL8*-co-expressed genes are present and iv) investigate whether DM linked SNPs are present in the *ANGPTL8* co-expressed genes. The study focused on the analysis of *ANGPTL8* co-expression genes module to increase our knowledge regarding its functions, its pathways based interactions (with co-expressed genes) and its relationship with the other DM related genes. To the best of our knowledge, this is the first instance to perform a transcriptomics data based analysis for functional assessment of *ANGPTL8* in liver.

Methodology

The complete work flow deployed in the present study is illustrated in Figure 6.1.

Selection of Transcriptomics Data Set

Liver is the predominant expression site of *ANGPTL8* that is also over-expressed during insulin resistance [42–44], obesity [41] and DM type II [45]. Therefore, the selection of a data set in which all of these conditions are present could aid in the identification of highly corelated genes with *ANGPTL8* based on similar expression pattern observed across all the samples. A systematic and thorough check of GEO database [3] was performed for the selection of a suitable data set as described above. The gene expression profiles of human liver samples with GEO ID: GSE64998 was selected out of the identified data sets (GSE15653, GSE23343, and GSE64998) based on the best quality and appropriate sample size for performing co-expression network analysis. It consists of six healthy control samples, eight obese non-diabetic and seven type 2 diabetic patient samples and was performed in GPL11532 (Affymetrix Human Gene 1.1 ST Array Platform). This data set had been already analyzed with a different approach and aims than ours by Kirchner. H and colleagues [50]. Several clinical parameters associated with the samples are also provided comprehensively by [50].

?

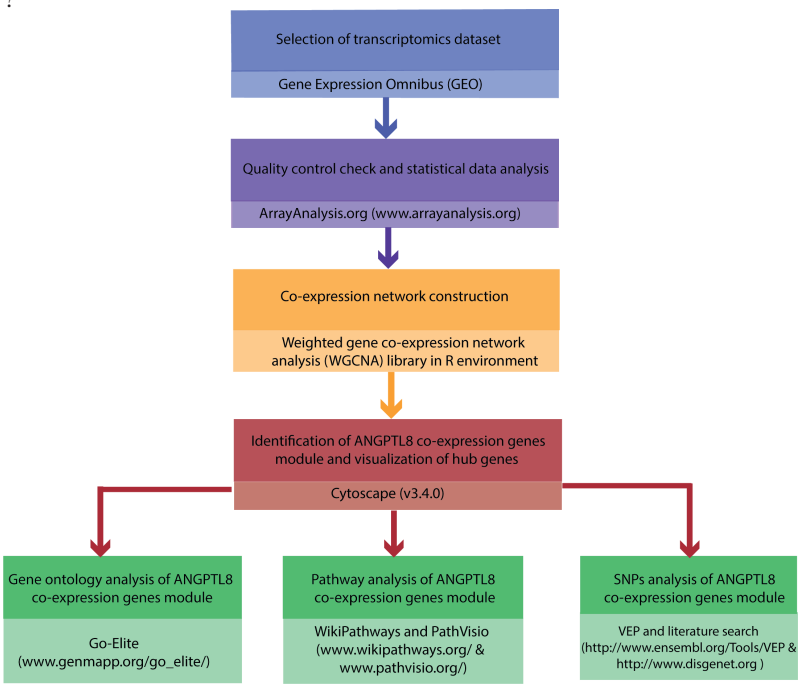

Fig. 6.1 Integrated workflow deployed for functional assessment of *ANGPTL8*: The steps and the tools/softwares used for the the quality control assessment, construction of *ANGPTL8* co-expression network, Gene Ontology analysis, pathway analysis and variant effect prediction analysis (through SNPs identification) are described.

Quality Control Check and Statistical Data Analysis

The raw data of GSE64998 was downloaded and reanalyzed using ArrayAnalysis.org [49]. ArrayAnalysis.org is a web server to perform quality control, preprocessing and statistical analysis of microarray data. We selected Entrez IDs for gene annotation of microarray probe IDs via ArrayAnalysis.org. The quality control and preprocessing report obtained is provided as supplementary file 1. The data was normalized using Robust Multi-array Average (RMA) method and is provided as data sheet 1 of supplementary material. All the samples of GSE64998 were included for the subsequent analysis as there were no outliers. Average expression of less than 5 was selected as cutoff value to remove the genes with low expression values from the data set which resulted in selection of 10869 genes (data sheet 2 of supplementary material).

ANGPTL8 Co-expression Network Construction

The weighted gene co-expression network analysis (WGCNA) is an established systems biology method for construction of correlation networks based on similar gene expres-

sion patterns observed across microarray samples [47]. It allows the identification of co-expression genes modules (set of genes observed with similar correlation pattern) from gene expression data through unsupervised learning methods. The method was implemented using the R package "WGCNA [48] in order to identify the *ANGPTL8* co-expression genes module. The preprocessed normalized data of all samples (data sheet 2 supplementary material) obtained in previous step was used as input. We selected automatic network construction and module detection method to perform WGCNA [48]. The complete R code utilized to perform the analysis is provided in data sheet 3 of supplementary material.

As a first step, a similarity matrix was constructed by measuring Pearsons correlation for all gene pairs. Next, an adjacency matrix was constructed by raising the similarity matrix to the soft thresholding power beta (Equation 1) [46].

$$a(i,j) = |cor(x(i), x(j)|^{\beta} \tag{6.1}$$

where *x(i)* and *x(j)* corresponds to expression values of gene *i* and gene *j*, respectively. The soft thresholding power beta is selected in order to achieve the approximate scale-free network topology as described in [48]. We selected power of $beta = 14$ to fulfill the scale free topology criterion. This Adjacency matrix was converted into Topological Overlap Measure (TOM) matrix where TOM is a highly robust network proximity measure [47, 48] (equation 2). Next, TOM matrix was converted into dissimilarity TOM matrix (equation 3) which was subsequently used to create a dendrogram through average hierarchical clustering method. Lastly, the dynamic branch cutting algorithm was applied on the dendrogram in order to obtain the clusters (modules) of highly correlated genes.

$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \tag{6.2}$$

where a_{iu}, a_{uj} and a_{ij} represents adjacency function based values between gene pairs *(i,u) (u,j)* and *(i,j)*. k_i, k_j represents connectivity of genes *i* and *j*, respectively.

$$DistTOM_{ij} = 1 - TOM_{ij} \tag{6.3}$$

The co-expression genes module detected with *ANGPTL8* was selected for further analysis and it was exported in Cytoscape [52] network format using the WGCNA R function "exportNetworkToCytoscape". This function allows to remove the edges with lower TOM values based on the value of the parameter named "threshold". We used a threshold value equal to 0.02 for removing the low weighted edges from the *ANGPTL8* genes co-expression module. The module-trait relationship was not assessed because we were not interested to relate the modules with a single phenotype as already described in section 6.

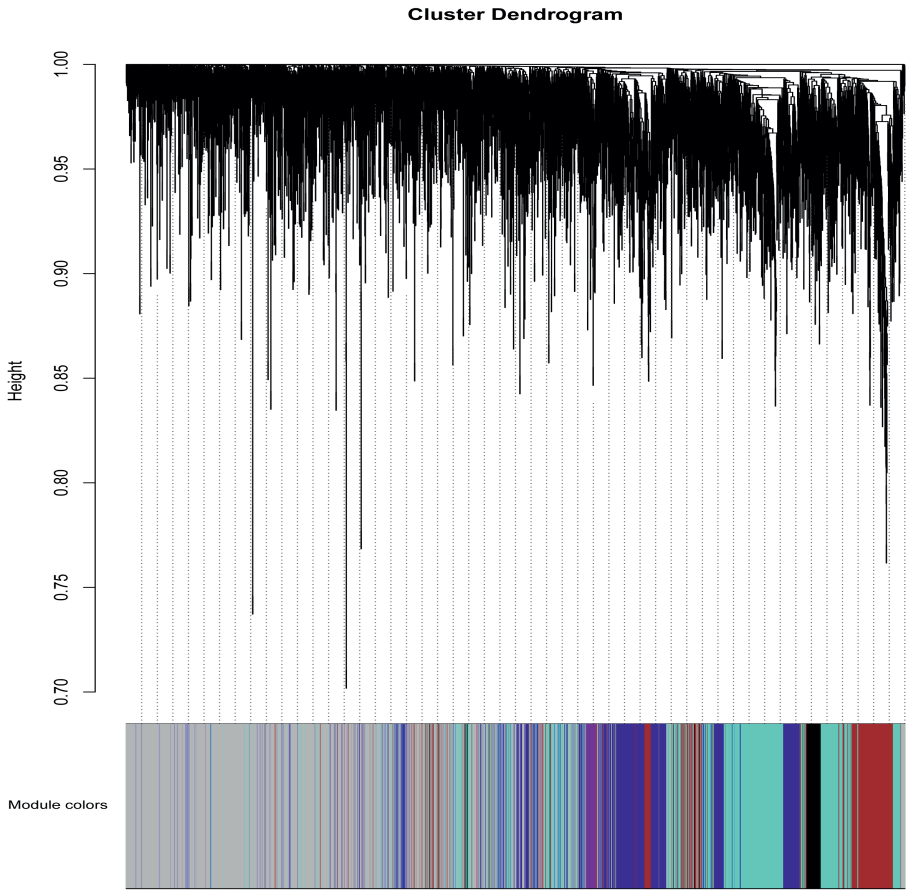

Fig. 6.2 The dendrogram of dissimilarity TOM Matrix constructed using hierarchical clustering. Each vertical line corresponds to the gene. Module colors are provided at the bottom. Each color corresponds to the separate genes module.

Identification of Hub Genes of Co-expression Genes Module of *ANGPTL8*

Next, the co-expression genes module of *ANGPTL8* (identified in the previous step) was visualized as a network using the network visualization and analysis software Cytoscape (version 3.4.0) [52]. For the identification of hub genes in the *ANGPTL8* related co-expression genes module, we used the connectivity (degree centrality) as described by Langfelder. P and colleagues [48]. In an undirected network, the degree centrality of a node (gene/protein/metabolite etc) can be defined as the total number of the edges incident on the node. Genes of *ANGPTL8* co-expression module with a degree greater

than or equal to 80^{th} percentile were considered as the hub genes [40]. The hub genes are the most representative genes in a co-expression network due to the presence of maximum number of co-expressed genes linked with them [48].

Gene Ontology (GO) Analysis of Co-expression Genes Module of *ANGPTL8*

Gene Ontology (GO) analysis allows us to infer the gene properties from the controlled vocabulary (defined terms) maintained by GO project [51]. Every gene product is classified based on three types of ontologies i.e. biological process (BP), molecular function (MF) and cellular compartment (CC). We used GO-Elite [7] version 1.2.5 to perform GO analysis [7]. GO-Elite is a software which identifies minimal non-redundant set of GO terms describing a given set of genes. We compared the genes present in the genes module identified with *ANGPTL8* with all the measured genes. We used the following settings for GO analysis: (i) 2000 permutations, (ii) Z-score threshold > 1.96, (iii) p-value threshold < 0.05 and (iv) minimum number of changed genes is three. We used Cytoscape (version 3.4.0) [52] for intuitive visualization of the results in order to analyze the connections between the genes and identified GO terms (Figure 1 of supplementary material).

Pathway Analysis of Co-expression Genes Module of *ANGPTL8*

We investigated the presence of genes identified within the *ANGPTL8* co-expression network in the complete curated human pathway collection ($n = 710$) of WikiPathways [1]. All pathways were scrutinized for the presence of at least one of the *ANGPTL8* co-expression module genes. PathVisio, was used for the visualization of the selected pathways [8]. The pathway analysis was performed in order to allow us to (i) determine the biological processes that might be the part of physiological mechanisms associated with *ANGPTL8* and the significant genes co-expressed with it; (ii) determine the unknown genes/proteins co-expressed with *ANGPTL8* from biological processes already known or associated with it.

Single Nucleotide Polymorphism (SNP) Analysis of Selected Co-expressed Genes with *ANGPTL8*

We identified SNPs associated to DM and other metabolic disorders that are located in 72 *ANGPTL8* co-expressed genes, present in a selection of ten pathways with the highest number of co-expressed genes. The analysis was performed using DisGeNET database [54] version 4.0. The names of 72 genes were provided as input in the gene

search panel of the DisGeNET website, in which the top 10 disease-association list and the top 10 disease-associated variants list for each gene, were further consulted. We extracted the names and IDs of the diseases associated to the genes that reported the highest DisGeNET score (Data sheet 4 of supplemental material). However, if in the top 10 disease-associations, a disease related to DM and other metabolic disorders was listed with a lower score, its name, ID and score was also included in Data sheet 4 of supplemental material. Moreover, in this table SNPs associated to DM and other metabolic disorders are also reported for several genes, with the name and ID of the associated disease and the DisGeNET score related to the strength of the association. The DisGeNET score ranges from 0 to 1 and it ranks the gene-disease associations taking into account the number and type of sources (level of curation, organism), and the number of publications supporting the association. The effect of the variants in the genes and the pathways were further investigated with literature search in Google and consultation of several databases such as: Ensembl [2] and SNPedia (`https://www.snpedia.com/`).

Results

Identification of *ANGPTL8* co-expression genes module and visualization of hub genes

WGCNA [46–48] is applied to gain insights into the functional organization of *ANGPTL8* and its associated co-expressed genes in human liver utilizing a transcriptomics data set of lean, obese and DM type II subjects (available online at [3]) [50]. *ANGPTL8* is a predominantly liver expressed gene in humans which has been found up-regulated in insulin resistance [42–44], obesity [41] and DM type II [45]. Therefore, a dataset expressing all these conditions was selected in order to allow the selection of highly corelated genes with *ANGPTL8* across all the samples and conditions. The expression profile of 10869 unique genes (data sheet 2 of supplementary material) obtained after normalization and filtering off the probes with low intensities were used to construct the gene co-expression network by applying the steps described in the section 6. Twelve gene modules (clusters of highly co-expressed genes) other than grey module (unclustered genes) were obtained by applying automatic module detection and dynamic tree cutting algorithm with minimum cluster size of 30. The graphical illustration of the resultant dendrogram, obtained from the hierarchical clustering based on the dissimilarity TOM matrix, is given in Figure 6.2. The number of genes in the corresponding modules with the respective color codes are provided in Table 6.1. The complete list

[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64998

of the genes (with respective Entrez ids) identified in each module is provided in data sheet 5 of supplementary material.

Table 6.1 Modules resulted from the hierarchical clustering: Module names are assigned colors and module size corresponds to the number of genes clustered in each module.

Module name	Module Size	Module name	Module Size
Turquoise	999	Black	371
Blue	783	Pink	314
Brown	600	Magenta	183
Yellow	498	Purple	76
Green	462	Greenyellow	64
Red	461	Tan	43

The red module was identified with *ANGPTL8* and 460 other genes and was then exported in cytoscape [52] network format for subsequent visualization and analysis. This network is composed of 447 nodes and 1781 interactions due to the filtering criteria used for removing the edges with the lower TOM values. It will be referred as co-expression network of *ANGPTL8* from here onwards. The graphical illustration of the entire co-expression network is provided as Figure 6.3. The topological analysis of this network revealed that 97 genes were greater than or equal to 80^{th} percentile according to the degree (Figure 6.3). These genes are classified as hub genes and represent the most highly connected nodes in the entire network. The hub genes in a co-expression network are important to gain insights into the associated functional roles (phenotypic outcomes) related to majority of the genes. It is because these genes show highly similar co-expression patterns and often are part of similar biological functions, biological process and/or are co-regulated [48].

Ubiquitin protein ligase E3 component n-recognin 2 (*UBR2*) is the gene with the highest degree in the entire co-expression network of *ANGPTL8*. It is a part of the N-end rule pathway which regulates proteolysis of intracellular proteins on the basis of identity of their N-terminal amino acids [69]. This pathway is found conserved from yeast to eukaryotes and is important determinant of half-life of diverse set of proteins. It has been previously demonstrated to serve various developmental and physiological processes including fidelity of chromosome segregation, apoptosis, autophagy, cardiovascular development in animals, regulation of cellular check point controls (by degradation of regulatory proteins involved in cellular differentiation, division and programmed cell death) quality control of cytosolic proteins, controlling the redox dynamics of stress related cellular compounds (such as nitric oxide, thiols, heme, oxygen and others) and leaf senescence in plants (reviewed in [29]). Additionally, it has been demonstrated to play an inhibitory role in mTOR signaling pathway [28]. It is interesting to observe

that *ANGPTL8* is a part of insulin and glucose mediated signaling pathway which also includes downstream regulators of mTOR signaling arm [31]. Moreover, one of the outstanding questions that remained elusive regarding *ANGPTL8* was the identity of its degradation pathway pointed out by Ren Zhang [37]. The results of current analysis demonstrate the possible involvement of N-end rule pathway (through UBR2) in degradation of *ANGPTL8* which should be further investigated through wet-lab studies. Other top nine hub genes based on the degree are *KANSL1L, ORC2, AGL, BNIP2, MET, MBTD1, TFPI, ALDH6A1*, and *SLC16A4* (Figure 6.3). The role of *KANS1L, ORC2, MBTD1, BNIP2, MET, TFPI* is mainly associated with DNA replication and/or cellular division; *AGL* and ALDH6A1 are enzymes involved in metabolic pathways and *SLC16A4* is a solute transporter protein [10–14, 26].

ANGPTL8 itself is connected with nine other genes in its co-expression network which means they are the most strongly co-expressed genes with it (Table 6.2). Two of these genes are also identified as hub genes i.e. tissue factor pathway inhibitor (TFPI) and insulin-like growth factor-binding protein 1 (IGFBP1). TFPI plays an important role in the regulation of blood coagulation pathway [27]. It is an inhibitor of tissue factor (TF) which is a glycoprotein present on surface of macrophages and other extravascular cells. TF is involved in positive induction of inflammatory cytokines (such as TNFα, IL-1 and IL-6) and coagulation signaling cascade. Thus TFPI plays a protective role in maintaining cellular and systemic homeostasis of immune system. Additionally, TFPI has been demonstrated to be involved in three interdependent biological processes that is coagulation, angiogenesis and lipid metabolism [6]. Excess cellular lipid forms lipotoxic metabolites (such as cholesterol crystals) which on one hand induce inflammatory cytokine production and on the other induce TFPI [6, 26].TFPI not only regulates the inflammatory processes through their inhibition but also reduces cholesterol concentration (through stimulation of internalization and degradation of VLDLs through HSPG-dependent pathway) [6]. *ANGPTL8* has also been previously demonstrated as an integral component of lipid metabolism. Therefore, these results imply that TFPI and *ANGPTL8* represent interesting multifunctional molecular players which might be mutually involved in maintaining the interconnected physiological feedback mechanisms between angiogenesis, coagulation and lipid metabolism. These feedbacks should be investigated further to understand the role of these genes in the integrated physiological pathways for maintaining homeostasis.

IGFBP-1 is a plasma carrier protein which binds to insulin-like growth factors (IGFs) I and II and increases their half-life [55, 56]. *IGFI* and *IGFII* are ligands of IGF signaling system involved in cell proliferation, differentiation, migration and metabolic processes. These ligands (*IGF I* and *II*) can bind with *IGF-I* and *II* receptors, isoforms of insulin-receptors and their hybrid receptors [25]. IGFBP-1 has also been demonstrated to improve whole body glucose regulation through its role in integrin mediated

Table 6.2 Neighbors of *ANGPTL8* in its co-expression network. Degree represents the number of connected nodes (genes) with respective genes in the *ANGPTL8* co-expression network.

Gene Symbol	Full Gene Name	Degree
TFPI	Tissue factor pathway inhibitor	42
IGFBP1	Insulin like growth factor binding protein 1	21
YKT6	YKT6 v-SNARE homolog	10
PPARGC1A	PPARG coactivator 1 alpha	6
MID1IP1	MID1 interacting protein 1	4
C10orf10	chromosome 10 open reading frame 10	3
BHLHE40	basic helix-loop-helix family member e40	3
VPS18	VPS18, CORVET/HOPS core subunit	3
SDF2L1	stromal cell derived factor 2 like 1	2

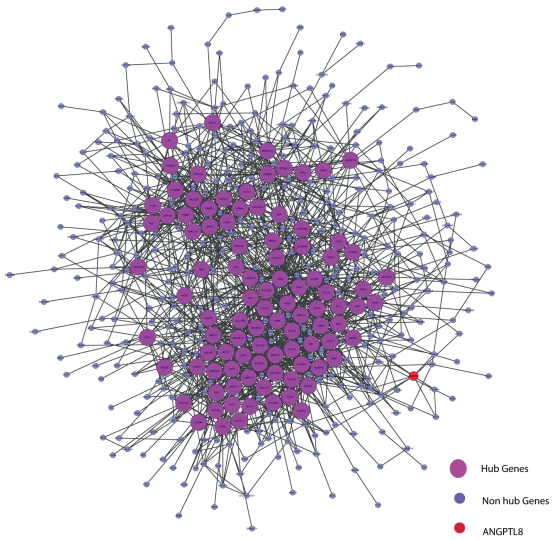

Fig. 6.3 The complete co-expression network of the red module containing *ANGPTL8* is provided. The nodes are annotated with their gene symbols. *ANGPTL8* is shown with red color. The hub genes are shown with bigger size as compared to the non-hub genes which are shown in smaller size.

signaling cascade [24]. *IGFBP1* can bind to integrins (transmembrane cellular adhesion proteins/receptors) through its Arg-Gly-Asp (RGD) domain and activate focal adhesion kinase (*FAK*) [23, 24]. FAK signalling converges with insulin/insulin like growth factors signaling at *IRS-1* phosphorylation signalling point. Previous studies have shown that *IGFBP-1* induced improved glucose regulation and increased insulin sensitivity is a part of protective mechanism induced upon insulin resistance in the body [24]. These results are crucial for further investigation with reference to the similar role of *ANGPTL8* demonstrated for improved glucose tolerance in insulin resistance condition by Guo and colleagues [44]. Their data indicated that *ANGPTL8* is increased in the presence of both glucose and insulin which subsequently induces the phosphorylation of AKT involved in improving glucose tolerance through inhibition of gluconeogenesis (via phosphorylation of *FOXO*) and induction of glycogen synthesis (via phosphorylation of GSK3beta). The signalling events leading to induction of *ANGPTL8* were verified in several other studies as well and can be visualized in the recently curated *ANGPTL8* regulatory pathway present in WikiPathways [31]. However, the mechanism of action of phosphorylation induced activation of *AKT* via *ANGPTL8* being direct or indirect (involving other genes/proteins than *ANGPTL8*) is still a quest. Therefore, it would be interesting to investigate the connection between *IGFBP-1* and *ANGPTL8* in improving glucose tolerance in insulin resistance since both are highly co-expressed with each other and are part of overlapping signalling pathways (focal adhesion pathway and insulin/*IGF* signalling pathway).

Among other neighbors, peroxisome proliferator-activated receptor gamma coactivator 1-alpha (*PPARG- C1A*) is a key regulator of mitochondrial biogenesis. It integrates vast set of physiological stimuli (including growth factors, stress, cold exposure, cytokines, exercise etc.) into respective metabolic responses involving fat and glucose metabolism [20]. It is a key co-activator of several transcription factors including NRFs, peroxisome proliferator-activated receptor (PPAR), thyroid hormone, glucocorticoid, oestrogen and oestrogen-related receptors (ERRs) alpha and gamma [19]. This result is in line with previously demonstrated regulators of ANGPTL8 including thyroid hormone receptor alpha and beta, Liver X receptor (*LXR*) and PPAR (reveiewd in [31]). Synaptobrevin homolog *YKT6* (*YKT6*) and vacuolar protein sorting-associated protein 18 homolog (VPS18) are two other neighbors of ANGPTL8 in the co-expression network which are involved in vesicular transport of cytoplasmic proteins within different cellular locations. VPS18 is specifically involved in the vesicles transport of endosome/lysome pathway [22]. *ANGPTL8* itself is a secreted protein which was demonstrated to reside in lysosomal vesicles like compartments in cytoplasm and these proteins (YKT6 and VPS18) might represent its associated partner molecules during vesicular transport process [36]. Among other neighbors, MID1 Interacting Protein 1 (*MID1IP1*) is involved in hepatic lipogenesis [18] and microtubule stabilization during cell division [17], stro-

mal cell-derived factors 1-alpha and 1-beta (*SDF2L1*) is a chemokine protein playing role in hematopoiesis [16, 70], Class E basic helix-loop-helix protein 40 (BHLHB2) plays role in cell differentiation and control of circadian rythm, and chromosome 10 open reading frame 10 (C10orf10) plays role in regulation of autophagy [15]. These results (identified co-expressed genes) are in line with the roles associated with *ANGPTL8* including lipid metabolism [37, 39], hematopoiesis [35], autophagy [36] and circadian rhythm [5]. Overall, the results of the co-expression network analysis revealed the genes with similar roles observed for *ANGPTL8* in previous studies. Therefore, these genes represent a focused and tremendous knowledge body for further investigations regarding functional insights of *ANGPTL8*.

Gene Ontology Analysis

Gene Ontology (GO) analysis was performed to find the significant GO terms associated with the genes present in co-expression genes module of *ANGPTL8* using GO-Elite [7]. Thirty Two biological processes, seven molecular function and six cellular components GO terms were identified to be associated with co-expression genes module of *ANGPTL8* (data sheet 6 of supplementary material). The graphical illustration of the significant GO terms along with the associated genes is provided as Figure 1 of supplementary material. Overall, the significant GO terms of biological processes were identified related to different metabolic processes. Top five biological processes on the basis of degree (number of connections of a node) include carbohydrate metabolic process (GO:0005975), monocarboxylic acid metabolic process (GO:0032787), regulation of small GTPase mediated signal transduction (GO:0051056), lipid modification (GO:0030258) and phosphatidylinositol biosynthetic process (GO:0006661). *ANGPTL8* is found associated with carbohydrate metabolic process (GO:0005975), that is the largest connected metabolic process in the entire network (Figure 1 in supplementary material). Previous studies have demonstrated the role of *ANGPTL8* in different metabolic processes including carbohydrate and lipid metabolism (reviewed in [21, 31, 71]). Other than the metabolic processes, several other biological processes were also identified including leukocyte migration (GO:0050900), extracellular matrix disassembly (GO:0022617), blood vessel development (GO:0001568), regulation of epithelial cell migration (GO:0010632) and regulation of epithelial to mesenchymal transition (GO:0010717). These biological processes are especially relevant to a recently revealed role of *ANGPTL8* in stimulation and proliferation of CD45+ hematopoietic derived cells demonstrated by Cox.A and colleagues [35]. *CD45* is a glycoprotein also known as receptor-type tyrosine-protein phosphatase C (*PTPRC*) which is present at the surface of leukocytes and their progenitor hematopoietic stem cells [30]. It plays important role in different hematopoiesis related processes including cellular differen-

tiation, migration and proliferation of hematopoietic stem cells (HSCs). It is a key signalling component of B- and T-cell activation. Since the underlying signalling pathways and genes of *ANGPTL8*s role in proliferation of CD45+ derived cells remained elusive, the co-expressed genes of *ANGPTL8* identified in these biological processes (provided in data sheet 6 of supplementary material) represent peculiar molecular targets for future investigation. The associated cellular compartment related GO terms includes mitochondrial matrix , (GO:0005759), microtubule (GO:0005874), actomyosin (GO:0042641), cell-cell junction (GO:0005911), cytoskeleton (GO:0005856) and microtubule organizing center part (GO:0044450). Mitochondrial matrix is a cellular site involving fatty acid oxidation and other energy expenditure related processes whereas the other identified compartments (microtubule, actomyosin, cytoskeleton, microtubule organizing center part) are involved in the cellular motility, maintaining cellular shape and cell division. These results are in line with the biological processes identified with the genes present in the co-expression module of *ANGPTL8* as discussed above. Finally, the main molecular functions identified in *ANGPTL8* co-expression network are phospholipase activity (GO:0004620), monocarboxylic acid transmembrane transporter activity (GO:0008028), ion channel binding (GO:0044325), protein binding, bridging (GO:0030674), nucleoside-triphosphatase regulator activity (GO:0060589), protein homodimerization activity (GO:0042803) and transferase activity (GO:0016740).

Pathways Analysis

The genes in the *ANGPTL8* co-expression network were further investigated for their presence in the complete curated WikiPathways collection. WikiPathways is a public repository of curated and dynamic models of biological processes [1]. A total of 474 human pathways were identified to contain at least one of the genes from the co-expression network of *ANGPTL8*. Whereas, 258 genes from co-expression network of *ANGPTL8* were identified to be present and 189 genes were identified to be not present in any of these identified pathways. The complete list of pathways along with the respective genes found in them is provided as data sheet 7 supplementary material and entire gene to pathway network of these results is graphically illustrated in Figure 6.4. Ten of these pathways identified with maximum (above 9) number of genes are listed in Table 6.3 representing highly associated biological processes with *ANGPTL8*. The gene-pathway network of these ten pathways (subset derived from the complete gene to pathway network in Figure 6.4 is graphically illustrated in Figure 6.5. The network is composed of a total of 72 genes and ten pathways connected with shared genes among them. Twenty two hub genes identified in this network are shown with large size as compared to non-hub genes in the network. Two pathways including the angiopoetin like protein 8 regulatory pathway (WikiPathways ID: WP3915) and Focal

Table 6.3 Top ten pathways in gene-pathway co-expression network: The maximum number of *ANGPTL8* associated co-expressed genes identified in WikiPathways along with respective gene symbols are listed.

PID	PathwayName	Gene Count	Genes
WP3915	Angiopoietin Like Protein 8 Regulatory Pathway	13	MAPK8, CYP3A4, PIK3R2, PIK3R4, ANGPTL8, PRKAB2, G6PC,MAP3K11, FLOT1, PIK3C2G, SHC1, SLC16A2, RAPGEF1
WP306	Focal Adhesion	13	MAPK8, COL1A1, ACTN1, COL5A1, MET, PIK3R2, PIK3R4, ZYX, SHC1, COL3A1, PIP5K1C, ARHGAP5, RAPGEF1
WP2882	Nuclear Receptors Meta-Pathway	12	FGD4, SLCO1B1, UGT1A9, CYP3A4, ABCB11, BHLHE40, GCLM, PRDX6, BAAT, PPARGC1A, SLC19A2, IGFBP1
WP3888	VEGFA-VEGFR2 Signaling Pathway	11	MAPK8, FRS2, ATF6, PFN1, SHC1, PLCG1, PIK3R2, MYH9, GIPC1, RAPGEF1, CYP2C8
WP481	Insulin Signaling	10	PIK3C2G, MAPK8, INPP4A, SHC1, GYS2, PIK3R2, PIK3R4, RAPGEF1, MAP3K11, FLOT1
WP3362	Chromatin modifying enzymes	10	KDM6A, CARM1, SETD1A, SMARCA4, ELP2, TADA1, SMARCD1,CHD4, GATAD2A, JADE3
WP702	Metapathway biotransformation	10	UGT1A10, UGT1A9, CYP3A4, FMO4, GLYAT, BAAT, CYP4V2, HNMT, CYP2C8, NAA40
WP2857	Mesodermal Commitment Pathway	9	BMPR2, MBTD1, KDM6A, DIP2A, BHLHE40, EPB41L5, C9orf72, HPRT1, AXIN1
WP3932	Focal Adhesion-PI3K-Akt-mTOR-signaling pathway	9	COL1A1, COL5A1, COL3A1, MET,GYS2, PIK3R2, PIK3R4, GNG7,PPARGC1A
WP3925	Amino Acid metabolism	9	CTH, MAOA, FH, GCLM, MCCC1, MUT, BHMT, HNMT, AUH

Adhesion pathway (WikiPathways ID: WP306) were identified with the presence of thirteen genes (maximum number of genes per pathway in this analysis) each. Both of

these pathways share several genes among them including *PIK3R2, PIK3R4, MAPK8, SHC1, RAPGEF1*. Previous studies have demonstrated the role of both of these pathways in improving glucose tolerance and insulin sensitivity in insulin resistance condition [23, 24, 44]. Besides, IGFBPI has been identified as a highly co-expressed gene with *ANGPTL8* (as mentioned in sections above) which induces focal adhesion pathway through its RGD domain [23, 24]. Therefore, further studies are required to investigate the interdependence of *ANGPTL8* signaling pathway and focal adhesion signaling pathway in regulating glucose homeostasis especially in pathological conditions like insulin resistance and DM. These results emphasizes the association of revealed molecular players with *ANGPTL8* which should be further investigated especially in these identified pathways.

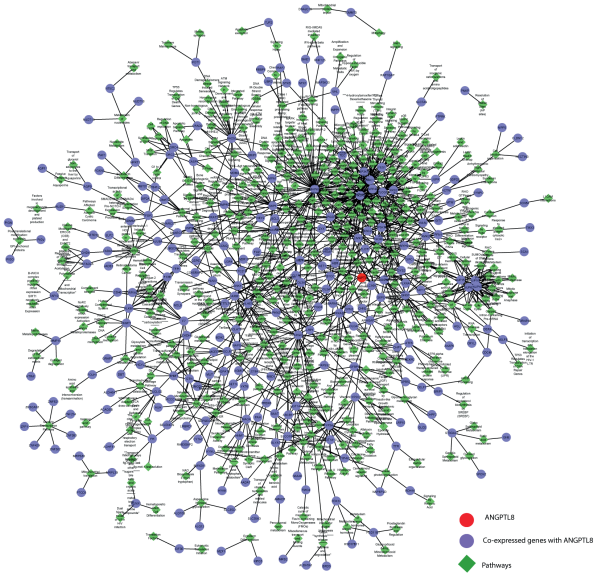

Fig. 6.4 The complete gene to pathway network. The nodes correspond to the genes and the pathways. The purple color nodes represent the genes and the green color nodes correspond to the pathways. The connections between the nodes indicate the presence of the genes in the corresponding pathways.

Mainly Angiopoietin Like Protein 8 Regulatory Pathway, Focal Adhesion pathway, VEGFA-VEGFR2 signaling pathway and Focal Adhesion-PI3K-Akt-mTOR-signaling pathway are part of overlapping signaling pathways. Other five pathways (Amino Acid metabolism, Mesodermal Commitment pathway, Metapathway biotransformation, Chromatin modifying enzymes and Nuclear Receptors Meta-Pathway) also share several genes among them and are in line with the previously demonstrated roles of *ANGPTL8* in metabolism and cell differentiation/division. Overall, the results of the

pathway analysis identifies important signaling pathways and associated co-expressed genes with *ANGPTL8* which should be investigated further for their mutual and individual role in pathogenesis of DM and related metabolic disorders.

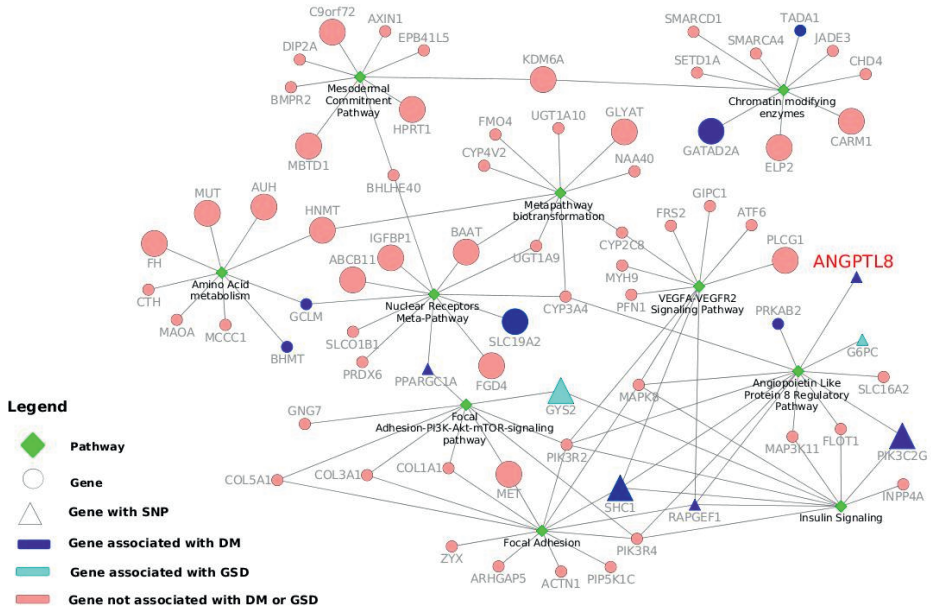

Fig. 6.5 Gene-pathway network with information of genes and variants associated to DM and other metabolic disorders. The top ten pathways identified with maximum number of genes in WikiPathways human curated collection are illustrated as green diamonds. The pathways are connected with seventy two genes (circles and triangles) and the size of the gene nodes indicate that the gene is either a hub gene (large nodes) or not (small nodes) in the co-expression network of *ANGPTL8*. The dark blue nodes indicate genes associated with DM. The aqua color nodes are the genes associated with GSD1. The triangle shapes represent the genes with a SNPs associated to DM or other metabolic disorders. *ANGPTL8* gene is highlighted in red.

Single Nucleotide Polymorphism (SNP) Analysis

We performed a SNP analysis on the 72 genes present in the ten highly associated biological processes with *ANGPTL8*. In data sheet 6 of the supplemental material we listed both the gene names queried in the DisGeNET database associated with the top diseases, and the SNPs, located in those genes, that reported the highest DisGeNET score with the disease association. Ten genes were found associated with DM Non-Insulin Dependent (*ANGPTL8, BHMT, GATAD2A, GCLM, PIK3C2G, PPARGC1A, PRKAB2, RAPGEF1, SLC19A2*, and *TADA1*). In particular, *RAPGEF1* with the intronic SNP rs11243444 [57] and *PIK3C2G* with the two intronic SNPs rs10841048 and rs12816270 [58] reported association with DM Non-Insulin Dependent, *PPARGC1A*

has the upstream SNP rs590183 associated with blood pressure [59] and *ANGPTL8* shows the missense variant rs2278426 associated with high density lipoprotein measurement [60]. In addition, the SHC1 gene does not show a gene association with DM, but it has the SNP rs8191979 associated to DM [61]. Moreover, *GYS2* with its missense SNP rs121918420 [62], and G6PC with another missense SNP rs1801175 [63] are associated with Glycogen storage disease type 1 (*GSD1*) that is a disorder characterized by severe fasting hypoglycaemia. Although, *GSD1* seems completely the opposite disorder of DM, they share similar metabolic pathways leading to nephropathy and fatty liver [64].The genes involved in the control of glucose and energy homeostasis are the same and for this reason investigating their variants effect can help to a better understanding of the role of these genes. In the Figure 6.5 ten genes reporting DM association and two genes associated with GSD1 are highlighted in dark blue and light blue, respectively. Moreover, when they present a SNP associated with DM or related phenotypic condition, the genes nodes are represented with triangles. The network in the Figure 6.5 visualizes the gene-pathway relationships, allowing to investigate deeply the roles of the seven genes already mentioned with a relevant SNP-disease association. We observed that those seven genes were grouped around five processes: Angiopoietin like protein 8 regulatory pathway (WP3915), Insulin signaling pathway (WP481), Focal adhesion-PI3K-Akt-mTOR signaling pathway (WP3932), VEGFA-VEGFR2 signaling pathway (WP3888) and Nuclear receptor meta-pathway (WP2882). The first three pathways not only contain at least one of the seven genes, but also share one or more of them. This is also due to the fact that Angiopoietin like protein 8 regulatory pathway diagram present subpaths of the other two processes, confirming the tightly biological interconnectivity within the three pathways.

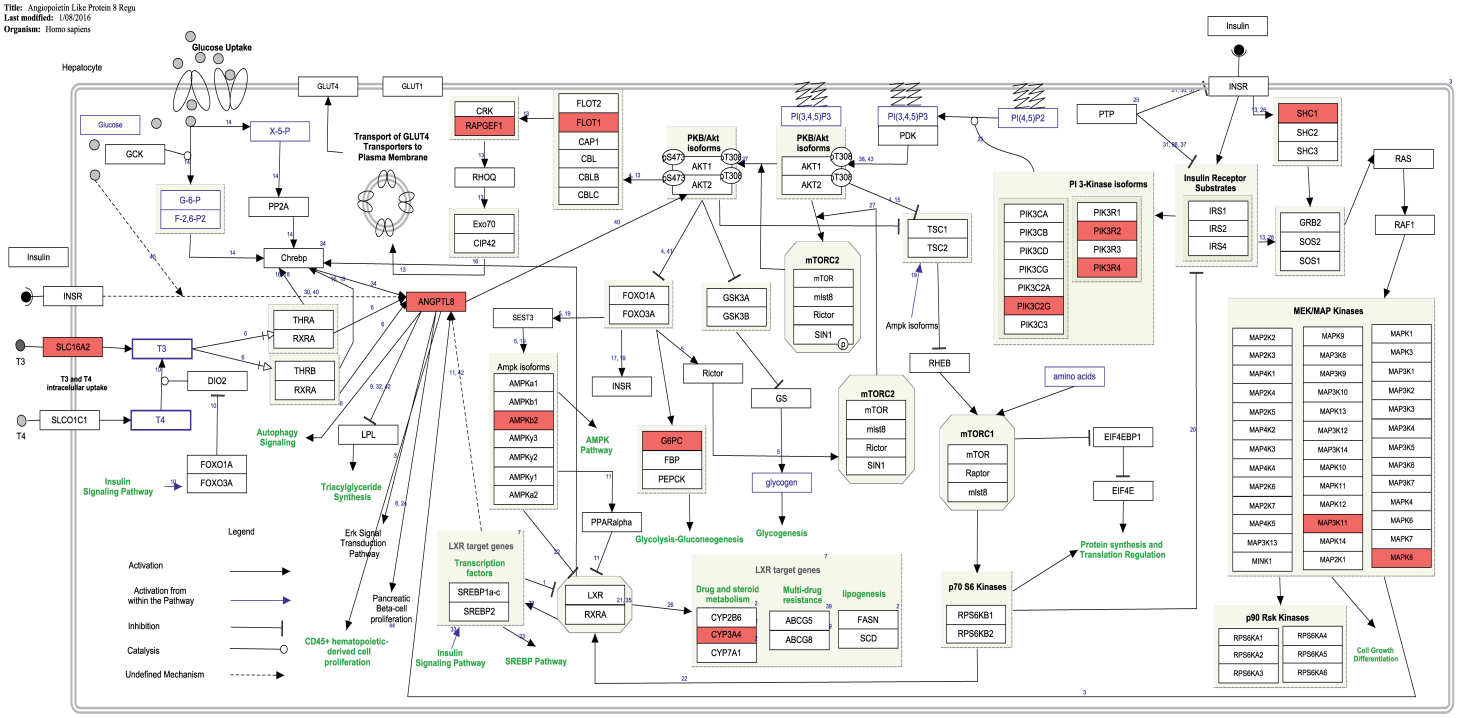

Fig. 6.6 Presence of co-expressed genes of *ANGPTL8* in the *ANGPTL8* Regulatory Pathway (pathway id:WP3915 (www.wikipathways.org/instance/WP3915)). Thirteen genes in the *ANGPTL8* regulatory pathway which are also identified to be co-expressed with it are marked with red color. These genes (*MAPK8, CYP3A4, PIK3R2, PIK3R4, PRKAB2, ANGPTL8, G6PC, MAP3K11, FLOT1, PIK3C2G, SHC1, SLC16A2, RAPGEF1*) belong to different signaling arms of the pathway. Among these genes, *SHC1* and *PIK3C2G* are also identified as hub genes.

Discussion

Several studies have demonstrated the role of *ANGPTL8* in lipid metabolism through LPL inhibition, regulation of autophagy and adipogenesis [36, 37, 39, 53]. It has also been demonstrated to regulate a crucial gene circuit required for maintenance of glucose homeostasis [44]. These unique features of *ANGPTL8* in regulation of different aspects of metabolism is driving the notion of its potential as a molecular target for treatment of DM. However, due to the lack of knowledge regarding its gene/protein partners, the associated biological processes and its mechanism of action, it has remained elusive to understand its role in pathogenesis of DM and subsequent assessment as molecular target. In this study, an integrated network analysis work flow especially suitable for such problems was designed to allow the extraction of relevant information with several regulatory levels. It helped us to identify the co-expressed genes with *ANGPTL8*, their identification as hub/nonhub genes, their presence in pathways and their co-occurrence in DM.

The current study provides the first instance of identification of co-expressed genes of *ANGPTL8* by utilizing a liver transcriptomics data set with the outcomes which are in line with previous literature (Figure 6.6) and also unfolds several regulatory findings which could present an important resource for future investigations (Figure 6.5). The co-expressed genes of *ANGPTL8* identified in this study $(n = 460)$ provides narrowed down list of molecular targets which are either co-regulated with it and/or might be regulation partners at different levels of interaction. Current analysis revealed the co-expression of thirteen genes with *ANGPTL8* in the literature curated pathway of *ANGPTL8* (WP3915) which was designed in our previous work [31]. These findings provides support to the current analysis and also emphasizes the association of the thirteen revealed molecular players with *ANGPTL8* in its pathway due to shared co-expression pattern (Figure 6.6).

Previous studies demonstrated the role of *ANGPTL8* in several biological processes such as carbohydrate and lipid metabolism, adipogenesis, autophagy and CD45+ hemtopoietic cell proliferation with none and/or partially identified partner proteins. The GO analysis performed in this study revealed several biological processes (in line with these previous literature findings) and the associated genes from co-expression network of *ANGPTL8*. Thus, the revealed genes in each biological process have implications for future investigation as being co-regulated with *ANGPTL8* or mutual engagement in these processes. The findings of the pathway analysis in the current study provides another level of information on the role of *ANGPTL8* in the identified biological processes. It allows us to view the interactions between *ANGPTL8* and the co-expressed

genes based on the previously identified pathway diagrams present in WikiPathways. The gene-pathway network represented in Figure 6.5 helped to identify visually the relationships between significant pathways and co-expressed genes with SNPs associated to DM and similar phenotypic traits. It is remarkable how the seven genes identified with a relevant SNPs association, happen to be clustered around processes linked with the Angiopoietin like protein 8 regulatory pathway. Although in some studies the variants-disease associations were detected in different populations than the Caucasian, such as Korean [57] and Aborigen [58], the literature regarding the gene-disease associations of those genes included Caucasian individuals as well. The effect of the SNPs is not always well characterized except for the missense variant of the *ANGPTL8* [60]. For this reason exploring the possible SNP effects in the pathways identified by those genes is not feasible with the literature information retrieved. However, from this genetic investigation it is possible to observe that there are significant genetic signals associated to DM and similar traits, influencing genes involved and co-expressed in *ANGPTL8* pathways. Thus, further experimental studies on those genes need to take the genetic background into account or under control in case of mice studies. Moreover, upcoming or existing GWAS studies for DM, could be checked for signals related to the co-expressed *ANGPTL8* genes, to properly assess their relevancy in the pathophysiology of the disease.

The key findings of this study provide focused information on molecular players co-expressed with *ANGPTL8* and associated pathways with implications for follow up experimental studies which could aid in identifying the exact mechanism of action and signaling events leading to pathogenesis of DM and metabolic disorders. Moreover, the integrated systems biology workflow deployed in this study provides a way to assess the gene-centric insights and to elucidate different levels of regulation from a transcriptomics data, in contrast to the typical -omics workflows which less directedly target the systems level knowledge.

Conclusion

In this study, an integrated systems biology workflow is deployed to analyze a hepatic transcriptomics dataset. The co-expression network analysis coupled with pathways analysis of this data aided in identification of the genes associated with *ANGPTL8* at different levels of regulation. The findings of GO analysis provided the complete annotation of the *ANGPTL8* co-expression genes module. Moreover, the genes already associated with DM in *ANGPTL8* genes co-expression network were identified which increased our knowledge regarding the possible mutual engagement of these genes in the pathogenic mechanism. All of the findings of this study have implications for follow up experimental studies which could aid in identifying the exact mechanism of action

and signaling events leading to pathogenesis of DM and metabolic disorders. Moreover, the integrated analysis workflow based on different methods and tools employed in the current study allows to assess a previously less characterized or uncharacterized gene/protein in a systematic way which may aid future studies.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

SC, CE, AS conceived the experiments, AS and EC conducted the experiments, AS wrote the paper. All the authors took part in discussions, analysis and layout of results and reviewed the manuscript.

Funding

No funding was received for this research study.

Acknowledgments

This work is a part of the published PhD thesis of Amnah Siddiqa.
We would like to thank Higher Education Commision (HEC) of Pakistan for providing the IRSIP fellowship to Ms Amnah Siddiqa as a visiting PhD Student at BiGCaT, UM. We would also like to thank Kirchner.H et al. without their work and data available at GEO (acession no: GSE64998), this paper would not have been possible.

Supplemental Data

The transcriptomic data set used for analysis was already available in the Gene Expression Omnibus database under accession no GSE6498. All the subsequent data generated and analyzed during this study are included in this published article [and its supplementary information files].

References

[1] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, and et al. Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, 2017.

[2] Xosé M Fernández and Ewan Birney. Ensembl genome browser. In *Vogel and Motulsky's Human Genetics*, pages 923–939. Springer, 2010.

[3] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data setsupdate. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[4] Elisa Cirillo, Laurence D Parnell, and Chris T Evelo. A review of pathway-based analysis tools that visualize genetic variants. *Frontiers in Genetics*, 8:174, 2017.

[5] Fabin Dang, Rong Wu, Pengfei Wang, Yuting Wu, Md Shofiul Azam, and et al. Fasting and feeding signals control the oscillatory expression of angptl8 to modulate lipid metabolism. *Scientific reports*, 6:36926, 2016.

[6] Eric W Holroyd and Robert D Simari. Interdependent biological systems, multifunctional molecules: the evolving role of tissue factor pathway inhibitor beyond anticoagulation. *Thrombosis research*, 125:S57–S59, 2010.

[7] Alexander C Zambon, Stan Gaj, Isaac Ho, Kristina Hanspers, Karen Vranizan, and et al. Go-elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics*, 28(16):2209–2210, 2012.

[8] Martina Kutmon, Martijn P van Iersel, Anwesha Bohler, Thomas Kelder, Nuno Nunes, and et al. Pathvisio 3: an extendable pathway analysis toolbox. *PLoS computational biology*, 11(2):e1004085, 2015.

[9] Martina Kutmon, Chris T Evelo, and Susan L Coort. A network biology workflow to study transcriptomics data of the diabetic liver. *BMC genomics*, 15(1):971, 2014.

[10] Yong Bao, Thomas L Dawson Jr, and Yuan-Tsong Chen. Human glycogen debranching enzyme gene (agl): complete structural organization and characterization of the 5 flanking region. *Genomics*, 38(2):155–165, 1996.

[11] Luigi Naldini, Elisa Vigna, Riccardo Ferracini, Paola Longati, Lucia Gandino, and et al. The tyrosine kinase encoded by the met proto-oncogene is activated by autophosphorylation. *Molecular and cellular biology*, 11(4):1793–1803, 1991.

[12] Yi-Bo Luo, Jun-Yu Ma, Qing-Hua Zhang, Fei Lin, Zhong-Wei Wang, and et al. Mbtd1 is associated with pr-set7 to stabilize h4k20me1 in mouse oocyte meiotic maturation. *Cell Cycle*, 12(7):1142–1150, 2013.

[13] Phillip B Carpenter, Paul R Mueller, and William G Dunphy. Role for a xenopus orc2-related protein in controlling dna replication. *Nature*, 379(6563):357, 1996.

[14] A Norling, AL Hirschberg, KA Rodriguez-Wallberg, E Iwarsson, A Wedell, and et al. Identification of a duplication within the gdf9 gene and novel candidate genes for primary ovarian insufficiency (poi) by a customized high-resolution array comparative genomic hybridization platform. *Human Reproduction*, 29(8):1818–1827, 2014.

[15] S Salcher, M Hermann, U Kiechl-Kohlendorfer, MJ Ausserlechner, and P Obexer. C10orf10/depp-mediated ros accumulation is a critical modulator of foxo3-induced autophagy. *Molecular cancer*, 16(1):95, 2017.

[16] Conrad C Bleul, Robert C Fuhlbrigge, Jose M Casasnovas, Alessandro Aiuti, and Timothy A Springer. A highly efficacious lymphocyte chemoattractant, stromal cell-derived factor 1 (sdf-1). *Journal of Experimental Medicine*, 184(3):1101–1109, 1996.

[17] Caterina Berti, Bianca Fontanella, Rosa Ferrentino, and Germana Meroni. Mig12, a novel opitz syndrome gene product partner, is expressed in the embryonic ventral midline and co-operates with mid1 to bundle and stabilize microtubules. *BMC cell biology*, 5(1):9, 2004.

[18] Nikolas G Tsatsos, Lance B Augustin, Grant W Anderson, Howard C Towle, and Cary N Mariash. Hepatic expression of the spot 14 (s14) paralog s14-related (mid1 interacting protein) is regulated by dietary carbohydrate. *Endocrinology*, 149(10):5155–5161, 2008.

[19] Renée Ventura-Clapier, Anne Garnier, and Vladimir Veksler. Transcriptional control of mitochondrial biogenesis: the central role of pgc-1α. *Cardiovascular research*, 79(2):208–217, 2008.

[20] François R Jornayvaz and Gerald I Shulman. Regulation of mitochondrial biogenesis. *Essays in biochemistry*, 47:69–84, 2010.

[21] Ren Zhang and Abdul B Abou-Samra. Emerging roles of lipasin as a critical lipid regulator. *Biochemical and biophysical research communications*, 432(3):401–405, 2013.

[22] Marjan Huizing, Aaron Didier, Jason Walenta, Yair Anikster, William A Gahl, and et al. Molecular cloning and characterization of human vps18, vps 11, vps16, and vps33. *Gene*, 264(2):241–247, 2001.

[23] Patricia Lebrun, Isabelle Mothe-Satney, Laurent Delahaye, Emmanuel Van Obberghen, and et al. Insulin receptor substrate-1 as a signaling molecule for focal adhesion kinase pp125fak and pp60 src. *Journal of Biological Chemistry*, 273(48):32244–32253, 1998.

[24] Natalie J Haywood, Paul A Cordell, Kar Yeun Tang, Natallia Makova, Nadira Y Yuldasheva, and et al. Insulin-like growth factor binding protein 1 could improve glucose regulation and insulin sensitivity through its rgd domain. *Diabetes*, 66(2):287–299, 2017.

[25] Antonino Belfiore, Francesco Frasca, Giuseppe Pandini, Laura Sciacca, and Riccardo Vigneri. Insulin receptor isoforms and insulin receptor/insulin-like growth factor receptor hybrids in physiology and disease. *Endocrine reviews*, 30(6):586–623, 2009.

[26] Sandra Espada, Benedicte Stavik, Sverre Holm, Ellen Lund Sagen, Vigdis Bjerkeli, and et al. Tissue factor pathway inhibitor attenuates er stress-induced inflammation in human m2-polarized macrophages. *Biochemical and Biophysical Research Communications*, 2017.

[27] Xia Dong, Li-ping Song, Dun-wan Zhu, Hai-ling Zhang, Lan-xia Liu, and et al. Impact of the tissue factor pathway inhibitor gene on apoptosis in human vascular smooth muscle cells. *Genetics and molecular biology*, 34(1):25–30, 2011.

[28] Kanako Kume, Yosuke Iizumi, Masashi Shimada, Yuki Ito, Tsutomu Kishi, and et al. Role of n-end rule ubiquitin ligases ubr1 and ubr2 in regulating the leucine-mtor signaling pathway. *Genes to Cells*, 15(4):339–349, 2010.

[29] Jung Hoon Lee, Yanxialei Jiang, Yong Tae Kwon, and Min Jae Lee. Pharmacological modulation of the n-end rule pathway and its therapeutic implications. *Trends in pharmacological sciences*, 36(11):782–797, 2015.

[30] Ian S Trowbridge and Matthew L Thomas. Cd45: an emerging role as a protein tyrosine phosphatase required for lymphocyte activation and development. *Annual review of immunology*, 12(1):85–116, 1994.

[31] Amnah Siddiqa, Elisa Cirillo, Samar HK Tareen, Amjad Ali, Martina Kutmon, and et al. Visualizing the regulatory role of angiopoietin-like protein 8 (angptl8) in glucose and lipid metabolic pathways. *Genomics*, 2017.

[32] Amy K Rines, Kfir Sharabi, Clint DJ Tavares, and Pere Puigserver. Targeting hepatic glucose metabolism in the treatment of type 2 diabetes. *Nature Reviews Drug Discovery*, 15(11):786–804, 2016.

[33] Arun Nanditha, Ronald CW Ma, Ambady Ramachandran, Chamukuttan Snehalatha, Juliana CN Chan, and et al. Diabetes in asia and the pacific: implications for the global epidemic. *Diabetes Care*, 39(3):472–485, 2016.

[34] Paul Zimmet, K George Alberti, Dianna J Magliano, and Peter H Bennett. Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. *Nature Reviews Endocrinology*, 12(10):616–622, 2016.

[35] Aaron R Cox, Ornella Barrandon, Erica P Cai, Jacqueline S Rios, Julia Chavez, and et al. Resolving discrepant findings on angptl8 in β-cell proliferation: A collaborative approach to resolving the betatrophin controversy. *PloS one*, 11(7):e0159276, 2016.

[36] Yi-Hsin Tseng, Po-Yuan Ke, Chia-Jung Liao, Sheng-Ming Wu, Hsiang-Cheng Chi, and et al. Chromosome 19 open reading frame 80 is upregulated by thyroid hormone and modulates autophagy and lipid metabolism. *Autophagy*, 10(1):20–31, 2014.

[37] Ren Zhang. The angptl3-4-8 model, a molecular mechanism for triglyceride trafficking. *Open biology*, 6(4):150272, 2016.

[38] Amnah Siddiqa, Jamil Ahmad, Amjad Ali, Rehan Zafar Paracha, Zurah Bibi, and et al. Structural characterization of angptl8 (betatrophin) with its interacting partner lipoprotein lipase. *Computational biology and chemistry*, 61:210–220, 2016.

[39] Ren Zhang. Lipasin, a novel nutritionally-regulated liver-enriched factor that regulates serum triglyceride levels. *Biochemical and biophysical research communications*, 424(4):786–792, 2012.

[40] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[41] Zhiyao Fu, Feven Berhane, Alemu Fite, Berhane Seyoum, Abdul B Abou-Samra, and et al. Elevated circulating lipasin/betatrophin in human type 2 diabetes and obesity. *Scientific reports*, 4, 2014.

[42] Zhiyao Fu, Abdul B Abou-Samra, and Ren Zhang. An explanation for recent discrepancies in levels of human circulating betatrophin. *Diabetologia*, 57(10):2232, 2014.

[43] Peng Yi, Ji-Sun Park, and Douglas A Melton. Betatrophin: a hormone that controls pancreatic β cell proliferation. *Cell*, 153(4):747–758, 2013.

[44] Xing Rong Guo, Xiao Li Wang, Yun Chen, Ya Hong Yuan, Yong Mei Chen, and et al. Angptl8/betatrophin alleviates insulin resistance via the akt-gsk3β or akt-foxo1 pathway in hepg2 cells. *Experimental cell research*, 345(2):158–167, 2016.

[45] Hodaka Yamada, Tomoyuki Saito, Atsushi Aoki, Tomoko Asano, Masashi Yoshida, and et al. Circulating betatrophin is elevated in patients with type 1 and type 2 diabetes. *Endocrine journal*, 62(5):417–421, 2015.

[46] Wei Zhao, Peter Langfelder, Tova Fuller, Jun Dong, Ai Li, and Steve Hovarth. Weighted gene coexpression network analysis: state of the art. *Journal of biopharmaceutical statistics*, 20(2):281–300, 2010.

[47] Bin Zhang, Steve Horvath, et al. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.

[48] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[49] Lars MT Eijssen, Magali Jaillard, Michiel E Adriaens, Stan Gaj, and Philip J andet al. de Groot. User-friendly solutions for microarray quality control and pre-processing on arrayanalysis. org. *Nucleic acids research*, 41(W1):W71–W76, 2013.

[50] Henriette Kirchner, Indranil Sinha, Hui Gao, Maxwell A Ruby, Milena Schönke, and et al. Altered dna methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. *Molecular metabolism*, 5(3):171–183, 2016.

[51] David Botstein, J Ms Cherry, M Ashburner, CA Ball, JA Blake, and et al. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–9, 2000.

[52] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, and et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[53] Gang Ren, Ji Young Kim, and Cynthia M Smas. Identification of rifl, a novel adipocyte-enriched insulin target gene with a role in lipid metabolism. *American Journal of Physiology-Endocrinology and Metabolism*, 303(3):E334–E351, 2012.

[54] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, and et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839, jan 2017.

[55] Briony E Forbes, Peter McCarthy, and Raymond S Norton. Insulin-like growth factor binding proteins: a structural perspective. *Frontiers in endocrinology*, 3, 2012.

[56] Sue M Firth and Robert C Baxter. Cellular actions of the insulin-like growth factor binding proteins. *Endocrine reviews*, 23(6):824–854, 2002.

[57] Kyung-Won Hong, Hyun-Seok Jin, Ji-Eun Lim, Ha-Jung Ryu, Min Jin Go, and et al. Rapgef1 gene variants associated with type 2 diabetes in the korean population. *Diabetes research and clinical practice*, 84(2):117–122, 2009.

[58] Makoto Daimon, Hidenori Sato, Toshihide Oizumi, Sayumi Toriyama, Takafumi Saito, and et al. Association of the pik3c2g gene polymorphisms with type 2 dm in a japanese population. *Biochemical and biophysical research communications*, 365(3):466–471, 2008.

[59] Christopher J O'donnell, L Adrienne Cupples, Ralph B D'Agostino, Caroline S Fox, Udo Hoffmann, and et al. Genome-wide association study for subclinical atherosclerosis in major arterial territories in the nhlbi's framingham heart study. *BMC medical genetics*, 8(1):S4, 2007.

[60] Daphna Weissglas-Volkov, Carlos A Aguilar-Salinas, Elina Nikkola, Kerry A Deere, Ivette Cruz-Bautista, and et al. Genomic study in mexicans identifies a new locus

for triglycerides and refines european lipid loci. *Journal of medical genetics*, pages jmedgenet–2012, 2013.

[61] K Almind, MG Ahlgren, T Hansen, SA Urhammer, JO Clausen, and O Pedersen. Discovery of a met300val variant in shc and studies of its relationship to birth weight and length, impaired insulin secretion, insulin resistance, and type 2 diabetes mellitus. *The Journal of Clinical Endocrinology and Metabolism*, 84(6):2241–2244, 1999.

[62] Marju Orho, Nils U Bosshard, Neil RM Buist, Richard Gitzelmann, Albert Aynsley-Green, and et al. Mutations in the liver glycogen synthase gene in children with hypoglycemia due to glycogen storage disease type 0. *Journal of Clinical Investigation*, 102(3):507, 1998.

[63] Roseline Froissart, Monique Piraud, Alix Mollet Boudjemline, Christine Vianey-Saban, François Petit, and et al. Glucose-6-phosphatase deficiency. *Orphanet journal of rare diseases*, 6(1):27, 2011.

[64] F Rajas, P Labrune, and G Mithieux. Glycogen storage disease type 1 and diabetes: learning by comparing and contrasting the two disorders. *Diabetes and metabolism*, 39(5):377–387, 2013.

[65] Yanhua Yin, Xiaoying Ding, Liang Peng, Yanqiang Hou, Yunxia Ling, and et al. Increased serum angptl8 concentrations in patients with prediabetes and type 2 diabetes. *Journal of diabetes research*, 2017, 2017.

[66] Genevieve Konopka, Tara Friedrich, Jeremy Davis-Turak, Kellen Winden, Michael C Oldham, and et al. Human-specific transcriptional networks in the brain. *Neuron*, 75(4):601–617, 2012.

[67] Michael C Oldham, Genevieve Konopka, Kazuya Iwamoto, Peter Langfelder, Tadafumi Kato, and et al. Functional organization of the transcriptome in human brain. 11(11):1271–1282, 2008.

[68] Ezra Y Rosen, Eric M Wexler, Revital Versano, Giovanni Coppola, Fuying Gao, and et al. Functional genomic analyses identify pathways dysregulated by progranulin deficiency, implicating wnt signaling. *Neuron*, 71(6):1030–1042, 2011.

[69] Daniel J Gibbs, Jaume Bacardit, Andreas Bachmair, and Michael J Holdsworth. The eukaryotic n-end rule pathway: conserved mechanisms and diverse functions. *Trends in cell biology*, 24(10):603–611, 2014.

[70] Toshiaki Ara, Yuri Nakamura, Takeshi Egawa, Tatsuki Sugiyama, Kuniya Abe, and et al. Impaired colonization of the gonads by primordial germ cells in mice lacking a chemokine, stromal cell-derived factor-1 (sdf-1). *Proceedings of the National Academy of Sciences*, 100(9):5319–5323, 2003.

[71] Yi-Hsin Tseng, Yung-Hsin Yeh, Wei-Jan Chen, and Kwang-Huei Lin. Emerging regulation and function of betatrophin. *International journal of molecular sciences*, 15(12):23640–23657, 2014.

CHAPTER 7

General discussion

The current thesis addresses two bioinformatics challenges related to analyze in a biological framework a set of genetic variants measured on a large scale. The first challenge relates to how the current technologies are designed for visualizing and analyzing single nucleotide polymorphisms (SNPs) in the pathway context. The second challenge focuses on how to explore and describe SNP effects by integrating multiple data types with pathway and network analysis methodologies.

Technological requirements for the analysis of genetic variation in pathways

The analysis of genetic variants requires the support of various bioinformatics tools: from the start when processing the DNA sequences to the end where the variants effect on a gene are placed in a broader context of the affected biological processes. Nowadays, an enormous variety of online software, web-based services and programmatic tools is available that can be used for the different stages of the genetic analysis [5, 9]. Chapter 2 presents an inventory and evaluation of tools performing analysis and visualization of variants linked to genes in pathways. Different beneficial characteristics showing the interactions between genes, metabolites and variants are identified in those tools. However, features including the integration of more specific methods for GWAS pathway analysis, and better visualization strategies for combining the genetic data with other omics data in pathways are recommended requirements for future tool developments. The first aspect can improve the accuracy and reproducibility of the analysis and the second one can support the interpretation of the biology behind the data [13, 14]. In addition, a more robust genetic analysis strategy must consider other types of genetic interactions, including edgetics [7, 12], gene environment (G - E) interactions [1], and epistatic interactions [10, 15]. Edgetics refers to network perturbation models focusing on specific alterations of the molecular interactions resulting from genetic variants [12]. This perturbation model might improve understanding of how mutations associating with complex diseases affect biological networks or interactome properties [7]. With network visualization already developed in some of the presented tools, it would be exciting to see this model implemented as a new feature. Another area in which pathway visualization of genetic associations can be improved involves G - E, where the genotypephenotype association exists only under certain environmental conditions. A recently published catalog of G - E interactions for numerous cardiometabolic phenotypes showed the wide extent under which the genotypephenotype association can be modified by factors such as diet, exercise, sleep, and many other exposures and lifestyle factors [1]. Third, epistasis is yet another manner in which connections within a pathway are different in different individuals, where two alleles mapping to differ-

ent loci associate in concert with a phenotype, but where those two alleles individually show no phenotype association [10, 15]. Consider, for example, that pathway endpoints are a phenotype, clinical indicator of health or disease status, or disease itself. Then, the epistatic relationships can be indicated by epistatic- or e-edges that serve to connect distinct pathways or different nodes within a single pathway in this conditional relationship. The pathways linked by such e-edges would give support to co-function and/or co-regulation with regard to the given phenotype of interest.

In summary the evaluation of tools in Chapter 2 has the scope to facilitate the work of both bioinformatics developers, in order to improve and provide better tools, and biologists who can become informed of the tools' potentialities and limitations.

Another aspect related to the applicability of the bioinformatics tools is the need for the biological entities used in the tools to be interoperable [4, 6, 11, 16]. This means that such entities are recognized and mapped with identifiers, enabling the use of the same element in different online sources and in a consistent manner. Currently, the interoperability is an application feature that has become very important in data analysis. Projects like the FAIR data principles, for instance, aim to increase awareness of this and other related issues in the data community, with the purpose to promote good practices in managing and re-using data [8]. Regarding the requirements to make genetic variant data interoperable, an upfront requirement is the proper identification and retrieval of the different variant types stored in online sources. In this thesis only SNPs were considered, but other variant forms exist: small or short-length variants like indels (insertion and deletion), and the structural variants like copy number variants. Large and well known databases like Ensembl (`http://www.ensembl.org`) and NCBI (`https://www.ncbi.nlm.nih.gov/`) store or provide the link of almost all types of variants using a specific nomenclature or identification. However, such identifications can change depending on the genome build considered and sometimes even if a new variant is discovered. Thus, a continuous update of the variant IDs and chromosome positions is required. Moreover, another controversial aspect is the mapping of variant to gene. Variants can be located in multiple genes or sometimes not even close to a single gene. Usually, when variants are annotated in databases, the chromosome position of the variant is the reference to check if and how many genes are located at that locus. However, if the interest of the user is to ascertain the biological effect of the variant, and this variant is located in multiple genes, usually focus is placed on the functionality of the chosen that relates most with the phenotype of interest. In addition, specialized researchers or physicians not always are able to share their databases, and often those specialized databases are very relevant for the biological interpretation. Despite such limitations, the amount of variants is growing daily at the Ensembl and NCBI repositories. For this reason, it is possible to use these data in connection with other sources and tools, in order to increase the understanding of variant function in a

biological system. A database identifier mapping service can make database information interoperable, and in Chapter 3 an improvement of the identifier mapping database called BridgeDb is proposed [17]. A gene-to-variant and variant-to-gene mapping using SNPs stored in Ensembl is integrated in the BridgeDb repository and is updated according to the regular BridgeDb mapping update schedule. This new feature empowers the user with the ability to analyze variants in different applications like R, PathVisio, and Cytoscape [3]. Tools that perform several types of tasks such as statistical analysis, pathway analysis or network analysis can now easily process the variant-to-gene (and gene-to-variant) identification, which then supports extensive genetic variant analysis.

Data analysis for interpretation of genetic variants

The improvement of the tools for genetic analysis is one of several basic steps towards the more impactful achievement of the interpretation of the effects of genetic variation. Once the tools and the databases are in place, the genetic analysis and the data interpretation can be performed. In this respect Chapters 4 and 5 show a workflow designed to extend analysis of a GWAS output, in which several data sources are integrated in tools that perform pathway and network analysis. In particular pathway analysis provides pathway data related to the genes where the variants are located, and network analysis is a supporting application to visualize and analyze the SNP-gene-pathway connections in the context of phenotypes. One major advantage of displaying these connections in a network is the easy detection of several biological relationships such as genes highly connected with pathways, and overlapping genes that share the same GWAS-identified SNP. The major difficulty is to discern if and how the intricate map represented in the network can be translated in a biological meaning. The features that can facilitate this interpretation step emerge from an examination of the major differences between the studies presented in Chapters 4 and 5. In those chapters the type of data chosen to describe the SNP-gene-pathway network in more detail are different. In both cases the starting point is a list of significant SNPs derived from GWAS studies, those SNPs are mapped to genes, and the genes are linked to their pathways. However, the selection of the data sources chosen to clarify the SNPs interpretation is different, mainly because the focus of the research question of the study differs as well.
In Chapter 4 the goal is to discover the role of variants that are located in the gene or very close to the gene, with the specific focus on SNPs that present a protein coding effect. For this reason, gene-environment relationships, eQTLs and laboratory experiments selected from literature were linked to the genes that show missense and nonsense SNPs. In Chapter 5 the goal is to understand if and which type of effect the GWAS variants show in the non-coding area of a gene. In this case epigenetics and eQTL data are used to identify clear signals of potential regulatory activities of the variants in specific tissue

types. The take of message of the interpretation of this data integration methodology based on pathway and network analysis is the possibility to combine and find meaningful relationships using different publicly available data. A key aspect of the analysis is the selection of the type of data sources, and it is the responsibility of the researchers to choose suitable sources that describe the biological meaning of the experimental data used and to explore the research hypothesis.

However, there are also technical bottlenecks of the analysis: data quality and the understanding of the network connections in relation to the biological question. The first limitation refers to the fact that re-using existing publicly available data is difficult because of the lack of information and/or metadata in the original data source. Mostly this refers to missing phenotypic descriptions of the samples and only in some cases a poor documentation of the methodology used to retrieve the data. The second limitation regards how the obtained network connections are interpreted. Usually there are typical network analysis algorithms [2] that facilitate the identification of hub nodes (e.g related to parameters like betweenness and centrality). However, these algorithms are not suitable in networks that integrate multiple entities such as the SNP-gene-pathway network, reported in Chapter 4 and 5. In this case, our analysis shows that interpreting of the network is facilitated by using extra information related to the data (such as Gene Ontology terms in Chapter 4 or eQTLs in Chapter 5). Those additional descriptors of the data nodes help to cluster the items in the network based on biological features and to find meaningful biological connections.

Finally, Chapter 6 reports another example of how re-using and combining prior knowledge from different data sources allows elucidation of the biological mechanism of a specific gene. In this case the *ANGPTL8* gene and its encoded protein are the focus because it is a potential drug target for T2DM. However, its specific regulatory mechanism is not known. A new and more comprehensive *ANGPTL8* pathway is designed using different literature sources from human and mouse experiments. In addition, results from independent transcriptomics and genomics studies of T2DM patients are visualized on the genes present in the *ANGPTL8* pathway. The outcome of this data integration is the identification of genes directly involved in the regulation of *ANGPTL8*, which also show significance in co-expression analysis and carry significant GWAS variants for T2DM. This result confirms the correctness of the pathway interactions drawn from literature that can be used for further computational experiments supporting validation tests in the wet laboratory. At the same time, the experimental data provide the strength of gene-gene interactions that could be further validated in the laboratory with ad hoc study.

Conclusion

The data in this thesis show that in order to conduct analysis that supports the interpretation of the effect of genetic variants in the pathway context, the analysis software must meet certain specific requirements. In these tools a combination of both algorithms suitable for GWAS analysis in pathways and a dynamic data visualization to display the analyzed data, are relevant in order to provide accurate results and to facilitate their biological interpretation. Currently, pathway applications still require the implementation of these two characteristics. In addition, enabling the analysis and visualization of new types of variant interactions (e.g. edgetics, gene - environment (G - E), and epistatic) would provide for conditional links within the SNP-centric network, and these would be an added improvement for the field. In terms of tool efficiency, the development of the interoperability in the genetic resources, is another technological aspect to consider. For this reason, in this thesis a new variant-to-gene and gene-to-variant mapping is implemented in the BridgeDb mapping database. Raising awareness about this aspect in the life science community is a responsibility of the bioinformaticians, computational biologists and data scientists. Lastly, several workflows for data integration using network analysis and pathway information are evaluated, resulting successful in two aspects. Firstly, network visualization allows depicting a general overview of the potential biological effect that significant SNPs associated with a certain disease could have on the phenotype. Secondly, using pathway knowledge enables the construction and understanding of specific gene-gene interactions not previously known. Further, development in using these methodologies with the individual DNA sequence of patients can lead to their application at the clinical level, in particular in the field of precision medicine.

References

[1] Laurence D Parnell, Britt A Blokker, Hassan S Dashti, Paula-Dene Nesbeth, Brittany Elle Cooper, and et al. CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits. *BioData Mining*, 7(1):21, dec 2014.

[2] Nadezhda T Doncheva, Yassen Assenov, Francisco S Domingues, and Mario Albrecht. Topological analysis and interactive visualization of biological networks and protein structures. *Nature Protocols*, 7(4):670–685, mar 2012.

[3] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, and et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[4] Sean Bechhofer, Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble, and

et al. Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*, jul 2010.

[5] Katherine Wolstencroft, Stuart Owen, Olga Krebs, Quyen Nguyen, Natalie J Stanford, and et al. SEEK: a systems biology data and model management platform. *BMC Systems Biology*, 9(1):33, dec 2015.

[6] Damien Lecarpentier, Peter Wittenburg, Willem Elbers, Alberto Michelini, Riam Kanso, and et al. EUDAT: A New Cross-Disciplinary Data Infrastructure for Science. *International Journal of Digital Curation*, 8(1):279–287, jun 2013.

[7] Florian Markowetz. How to Understand the Cell by Breaking It: Network Analysis of Gene Perturbation Screens. *PLoS Computational Biology*, 6(2):e1000655, feb 2010.

[8] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, and et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, mar 2016.

[9] Angela Bauch, Izabela Adamczyk, Piotr Buczek, Franz-Josef Elmer, Kaloyan Enimanev, and et al. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, 12(1):468, 2011.

[10] Rishika De, Ting Hu, Jason H. Moore, and Diane Gilbert-Diamond. Characterizing gene-gene interactions in a statistical epistasis network of twelve candidate genes for obesity. *BioData Mining*, 8(1):45, jun 2015.

[11] Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, and Steve Pettifer. API-centric Linked Data integration: The Open PHACTS Discovery Platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 29:12–18, dec 2014.

[12] Quan Zhong, Nicolas Simonis, Qian-Ru Li, Benoit Charloteaux, Fabien Heuze, and et al. Edgetic perturbation models of human inherited disorders. *Molecular systems biology*, 5(1):321, jan 2009.

[13] Alejandra González-Beltrán, Peter Li, Jun Zhao, Maria Susana Avila-Garcia, Marco Roos, and et al. From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing: The Complementary Roles of Data Models and Workflows in Bioinformatics. *PLOS ONE*, 10(7):e0127612, jul 2015.

[14] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10):e1003285, oct 2013.

[15] Wen-Hua Wei, Gibran Hemani, and Chris S. Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733, nov 2014.

[16] Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, and et al. Toward interoperable bioscience data. *Nature Genetics*, 44(2):121–126, feb 2012.

[17] Martijn P van Iersel, Alexander R Pico, Thomas Kelder, Jianjiong Gao, Isaac Ho, and et al. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1):5, jan 2010.

SUMMARY

In this thesis solutions related to two aspects of the analysis of SNP data are presented. The first aspect regards the type of tools and technologies available for the analysis of SNPs in the context of biological pathways. The second aspect concerns the identification of potential biological function of the SNPs associated to a certain phenotype.

Currently, a variety of tools are used to perform analysis of genetic variants. **Chapter 2** presents an inventory and evaluation of tools that combine the analysis and visualization of variants linked to genes in pathways. We identified the advantages and the limitations of those tools. The purpose was to facilitate the work of both bioinformaticians and experimental biologists that are the first users of such tools. In addition, we propose new analytical features to add to software, such as the inclusion of new types of genetic interactions: edgetics, gene-environment interactions and epistasis relationships. **Chapter 3** presents an implementation of the BridgeDb identifier mapping database aimed at the improvement of interoperability in pathway analysis through mapping of genetic variants to genes. A gene-to-variant and variant-to-gene mapping using publicly available SNPs, is available in the BridgeDb repository. This has expanded the mapping domains available in that repository, that already included gene-products, metabolites, and reactions. The addition of the variants to gene mapping enables building a more complex data analysis in the different applications in which BridgeDB is integrated, such as R, PathVisio, and Cytoscape or can be used as a webservice e.g. in R or any programming language.

The second part of the thesis presents methods that are able to integrate multiple data in order to explain the genetic variation effect in diseases such as T2DM or in the development of obesity. In this respect **Chapters 4** and **5** show a workflow designed to further analyze a GWAS output, combining multiple data sources using pathway and network analysis. The result is a workflow that ends with a network construction, used as support for the interpretation of the data in a biological point of view. With the distinction that in **Chapter 4** the focus is on the SNPs that cause a protein coding effect, and in **Chapter 5** the main interest is on the understanding of the variants effects in non-coding area of the genome. The lesson learnt is pathway and network analysis are both valid methodologies that support data integration, and thereby allow us to dive deeper in the biological context. However, bottlenecks are still present concerning data re-usability. In addition, researchers need to select carefully the data that they wish to integrate, based on relevant biological connections with the genes affected by genetic variants.

Finally, **Chapter 6** reports an example in which re-using open data of different types such as co-expressed genes in diabetes, significant associations of SNPs with diabetes and pathways, provided biologically relevant outcomes. The purpose was to explore in more detail the molecular interactions of the ANGPTL8 gene with other genes and proteins. The results confirm that re-using data with a different purpose than the one for

which the original studies were designed, is possible. Indeed, we were able to identify what types of relationships exist between ANGPTL8 variants and T2D phenotypes in our populations, which can be validated in follow-up laboratories studies.

RIASSUNTO

In questa tesi vengono presentate soluzioni relative a due aspetti dell'analisi delle varianti genetiche SNPs. Il primo aspetto riguarda il tipo di tecnologie disponibili per l'analisi degli SNPs nel contesto dei pathway biologici. Il secondo aspetto riguarda l'identificazione della potenziale funzione biologica degli SNPs, associati ad un determinato fenotipo.

Il **Capitolo 2** presenta un inventario e una valutazione di software che combinano l'analisi e la visualizzazione di SNPs, associati a geni presenti in specifici pathways. Sono stati identificati vantaggi e limiti di tali software, con lo scopo di facilitare il lavoro di bioinformatici e biologi sperimentali. Inoltre, sono state proposte nuove funzionalitá analitiche da aggiungere ai software, come l'inclusione di nuovi tipi di interazioni genetiche tipo: edgetics, interazioni gene-ambiente e relazioni epistasiche. Il **Capitolo 3** presenta un'implementazione del database di mappatura chiamato BridgeDb, finalizzata al miglioramento dell'interoperabilitá nell'analisi dei pathway in cui sono presenti geni contenenti rilevanti varianti genetiche. La mappatura da gene a variante e da variante a gene, e' disponibile pubblicamente nella repository di BridgeDb. Questa implementazione ha ampliato i domini di mappatura disponibili nel database, che includeva giá geni, metaboliti e reazioni. L'aggiunta delle varianti alla mappatura dei geni, consente di costruire un'analisi dei dati piú complessa nelle diverse applicazioni in cui BridgeDb é integrato come: R, PathVisio e Cytoscape. Inoltre BridgeDb puó essere utilizzato come servizio web, ad esempio in R o in altri linguaggi di programmazione.

La seconda parte della tesi presenta metodi che sono in grado di integrare piú dati, al fine di spiegare l'effetto di varianti genetiche in malattie come il Diabete Mellito o l'obesitá. A tale riguardo, i **Capitoli 4** e **5** mostrano un workflow progettato per analizzare piú in dettaglio gli SNPs risultanti da studi di associazione genetica chiamati GWAS. In tale workflow gli SNPs sono combinati in un network con diversi tipi di dati ad esempio geni e pathways. Tale network é utilizzato come una mappa di supporto per l'interpretazione degli SNPs da un punto di vista biologico. Con la differenza che nel **Capitolo 4** il focus é sugli SNPs che causano un effetto funzionale sulla proteina, e nel **Capitolo 5** l'interesse principale é comprendere gli effetti delle varianti localizzate nell'area non codificante del genoma. La lezione appresa che l'analisi di SNPs usando networks e' una valida metodologia che supporta l'approfondimento dello studio degli SNPs nel contesto biologico. Tuttavia, sono ancora presenti dei problemi relativi alla riutilizzabilitá dei dati. Inoltre, i ricercatori devono selezionare attentamente i dati che desiderano integrare nel network, sulla base di connessioni biologiche rilevanti con i geni interessati dalle varianti genetiche. Infine, il **Capitolo 6** riporta un esempio in cui il riutilizzo di dati pubblici di diversa tipologia come: i geni co-espressi in individui con il diabete, gli SNPs associati con il Diabete e i pathway biologici, ha fornito esiti biologicamente rilevanti. I risultati confermano che é possibile riutilizzare i dati con uno scopo diverso da quello per cui sono stati progettati gli studi originali. Infatti, sono

state esplorate in piú dettaglio, le interazioni molecolari del gene ANGPTL8 con altri geni e proteine e il tipo di relazioni tra le varianti di ANGPTL8 e quelle delle proteine che interagiscono con ANGPTL8. Lo studio si e' focalizzato su individui con il diabete e i dati ottenuti possono supportare e direzionare la scelta di nuovi studi di laboratorio.

VALORIZATION

From researchers to researchers

Nowadays, the internet provides an enormous number of tools that scientists can use to analyze, consult and store data obtained from the wet laboratory. However, the time has not yet come where ”one tool does everything”. Thus, often researchers spend time to search for tools that potentially provide the service required, check if and how the tool performs the tasks desired, investigate if the way to run it fits with the user competencies and only then are able to test that analysis tool for his/her research. These operations require investment of time, effort and expertise that often biologists do not have. For this reason, the review article presented in this thesis is a valuable resource to those researchers who wish to interpret genetic variants using the power of the biological pathway knowledge. The fact that the applications reviewed have the peculiarity of providing visualization features and a user interface makes the tool evaluation even more appealing for those researchers that are not very familiar with programmatic tools. Moreover, the relevance of such a tool inventory is not only for those that will use the tools, but also for bioinformaticians or computational biologists that develop such applications. These researchers need the opinions of the users and experts of the field, in order to understand if the tools work properly, and what type of improvements are needed to perfect them.

Another effort in raising awareness on using bioinformatics methodologies for data interpretation is presented in Chapter 5, where we make publicly available a video tutorial and a web-session of the genetic networks obtained by combining multiple data types. Sharing with the scientific community such resources is important to: i) encourage the application and validation of the data integration method in other datasets, ii) reproduce the analysis step-by-step in an easy to follow manner, iii) further investigate the results with different perspectives.

Finally, presenting an inventory and evaluation of tools, sharing data and explaining the methodology are all efforts that improve communication between researchers. This is a key aspect in modern science, where due to the technological advancements expertise becomes more and more specialized, but at the same time where interdisciplinary skills are an essential requirement.

Towards an improvement of data interoperability

An important concept in data science is to enable computer systems and software to exchange and easily make use of information, this concept is also known as interoperability. This is an essential aspect for improving the understanding of data in life sciences. Nowadays, it is clear to every researcher that the advancement of the new technologies create the ”big data” issue. This is not only a problem related to data

storage, but also to data use and interpretation. In this regard, putting the effort of increasing and perfecting data interoperability worldwide contributes to building a better structure of the body of human knowledge. Moreover, data analysis tools can see improved performances if the data are properly linked, and researchers smoothly can run workflows that comprehend the usage of several tools and environments. Such implementation results in saving time and money for complex and specialized data analysis. In line with this value, Chapter 3 includes an implementation of a mapping protocol for existing identifiers called BridgeDb, which is an existing resource that makes data interoperable because it matches the different types of identifiers related to the same biological entity. The map already stores genes, proteins and metabolites of different species including human. The implementation described here is a gene-to-variant and variant-to-gene mapping that adds a new dimension in the database related to the molecular world/dogma. The tangible benefit of such an implementation is evident in the application of the mapping database to the tools that analyze biological data. Indeed, the map can be used in pathway and network applications (e.g. PathVisio and Cytoscape) that allow analysis and visualization of different data types.

Looking forward to the future of precision medicine

SNPs are the common genetic variations in the human genome, and currently there is growing business activity around the concept of precision medicine, much of which relies on the possibility to perform genetic tests to predict diseases or improve physical or health conditions. Numerous companies, rather than hospitals or clinics, are providing genetic tests for several types of purposes. In those tests specific variants are assessed and related to risk of certain diseases, food intolerances, or even improvement in physical performance. For this reason, tools and methods that are able to support and improve the interpretation of the biological effects of the genetic variant are in high demand. In this thesis, a workflow is presented that combines multiple data types primarily in order to understand the effect of the genetic variant in specific clinical conditions, namely T2DM and obesity. Moreover, genetic reference networks of SNPs associated with obesity are proposed, in an attempt to provide a visual instrument to elucidate the biological and medical fucntion of the variants. These maps of SNPs, genes, pathways, and their relationships to each other can be used in different ways by different stakeholders who are interested in obesity and personalized treatments. Experts in the field of obesity can explore the networks to generate novel hypotheses or confirm results related to the functional role of BMI SNPs and their possible effects on gene regulation by influencing epigenetic marks. Clinicians involved in precision medicine also can benefit from such networks. For patients with available SNP genotyping data, the health care team, in theory, can determine susceptibility to certain diseases by consulting the reference

networks. Exploring those SNPs present in the network and the patients genotype can assist interpretation of the impact of the patient's alleles, linking them to the gene and the functional context in which they are involved. For example, the occurrence of several genotyped SNPs from the patient that indicate presence of risk or effect alleles that occur in the same or related pathways, can prompt the health care team to evaluate if those processes, in relation to the specific tissue, are relevant to the patient's current or future condition.

Conclusion

Current fields of bioinformatics and systems biology have developed new technologies and methodologies to further explore life science data. However, often the technical specialization within this field increase the gap of biological understanding, mostly due to a jargon issue. This is the reason why in the Bioinformatics and System biology communities despite the technical advances, researchers need to be able to be good translators within the biological and computer science area of knowledge. This role itself has a strategic influence in terms of societal and economic value, and this thesis contributes to highlight it. The societal value of the work reported relate to the aspect of improving communication within researchers of the same discipline (like bioinformatics), but also stakeholders of different fields, (e.g. bioinformaticians, biologists and clinicians). On the other side, the economic value is less tangible to the public because the advancements presented, such as the workflow for data integration and interpretation and an example of improvement in tool interoperability, directly benefit the area of basic research. Improved methods to perform analysis and interpret results is a key to innovation, without wasting public funds. The work performed in this thesis definitely contributes to this purpose.

ACKNOWLEDGEMENTS

These four years and half were wonderful and intense and I need to thank many people.
First of all thank you Chris, I liked a lot the way that you guided and taught me how to become an independent professional, and strengthen my scientific skills. I enjoyed our scientific discussions on SNPs, pathways and networks, and more important I am proud of the human and professional relationship that we were able to built.
Larry, you have the role of second supervisor in this project, and despite the 5,666 km of separation, you were always able to be present in my path, and keep the constant support throughout the years. Thank you a lot to be involved, to be supportive and also funny at the time.
Susan, I think you know already that I am grateful to the moment that you stepped in my PhD project. I think that your guidance was crucial for many aspects: time management, scientific advices, life advices, etc. All of this gave me a lot of confidence, support and energy to arrive till the end. Thank you, thank you, thank you.
For all the BiGCaT people, my biggest fear when I had to find a new job, was that I might not be able to find again a friendly and supportive work environment as it is present in the department.
Tina and Mirella, I choose you as paranymphs, because you embody very well my PhD path. First Tina, thank you very much for your time and patience spent at the beginning of the PhD. You introduced me into the work environment, you helped me when I had no idea on how to move, and you also stand by me in several projects, always with kindness and professionality.
Mirella, you came in the second stage of my PhD path. In that moment I needed a confident friend, that would be eager to share with me good and bad things, between work breaks, or just be supportive when the work was a bit intense. Thank you a lot for being all of this.
Then, Anwesha an old BiGCaT, I keep very nice memories of our office life. Especially during our pregnancy time, for which we were able to synchronize in a wonderful way. Thank you for being spontaneous and always ready to share a good tip, from a Linux code to the nicest website for shopping. Ryan, after Anwesha you became my new office mate, and I could not ask for a better one. Thank you for sharing music, political opinions, general societal conversations and more and more things, during the work breaks. This definitely made my work life happier. Nuno, you deserve a big thank you for the patience, kindness, and competence. You helped me to solve smoothly all the technical issues that once in a while came. We miss you at the department!
Then, Jonathan, the ”blonde guy”, you took over very nicely Nuno’s tasks. Especially for me, when the technical issues came again, you were always able to give it a try and patiently find a solution. Thank you a lot, the workflows explained in this thesis are published also because of your hands! Lars, thank you very much to be so light and fresh at any time even if the clock is running far away in front of you. We might

have not done a full project together, but I admire your knowledge that range from the genetics to the mathematics and beyond! Egon you always were able to share some tips or experiences, revealing interesting and helpful points of views. Freddie, I told you already that I admire your discipline in performing tasks, deliver papers and following multiple projects. Thank you for being such a good example to me, and thank you even more for your delicious cakes, that quite often sweetens the work day! Then, the new PhDs: Denise, Nasim and Marvin. All of you came at the last period of my path, but was enough to be able to share good moments. Thank you Denise for your energy, thank you Nasim for your kindness and thank you Marvin for your funny spirit! I will miss you very much!

Then, the BiGCaT secretaries: Jos, Ria and Myrtle, from all of you I received some help together with the smile! Thank you very much for your work, and especially to Myrtle that followed the last years of my PhD. I am grateful for all your support with byteMAL, invoices and thesis bureaucracy. It meant a lot to me!

Through the years, people at BiGCaT has came and gone, and some of them let me a very good memory. Amnah, I feel that you never went back to Pakistan, because we were able to maintain a great collaboration. Thank you for sharing with me your precision to being a scientist, for the discussions on the white board and for sharing the life troubles. This thesis contains pieces of you not only because of the paper that there is inside! Amadeo, you actually are still half a BiGCaT, thank you a lot for your Italian-Spanish spirit. Linda, Yokathama, Zahara, each of you is in my mind, and you influenced in different ways my work.

During these PhD years I had the wonderful opportunity to be part of several groups, that were dedicated to strengthen the communication within PhD students in different areas.

Thank you to the members of the RSG. First Annika that introduced me to the group and then the RSG board members (Summer 2016 - Summer 2018) : Joske, Mirella, Kyoko and Gosia. I felt to belong to a professional group with whom I also shared several crazy moments! An extra thanks to Kyoko that was eager to collaborate with me since the first meeting at BioSB in front of her poster.

Then, the NUTRIM PhD Council was also a great experience, mainly because of the people that took part of it, thank you: Rianne, Lotte, Bernard, Pauline, Charlotte, Max, Mirijam, Eva, Mattea, Jacqueline, Martijn and Nijnke. It has been a pleasure to organize social projects with you!

Finally, there is byteMAL, or better the byteMAL people. A big thank to Lisa and Manuel that two years ago blindly decided to follow me and Ryan in this adventure of creating an international conference in Limburg. Thank you to all the members that followed: Pejam, Ali, Mirella, Shauna, and Bob. We achieved more than what I had expected initially! I wish you good luck for all the upcoming conferences and I hope

that byteMAL can continue for long!
Maastricht University gave me the opportunity to meet wonderful people. The list is long to be written in this acknowledgement, but I still want to mention some names: Georgina, Pierre, Dries, Lauren, Monika, Nicola, Sarah, Dennie, Michiel, Samar. For Monika and Nicola a special thank you, because of the Italian fresh air and support that you gave constantly.
Finally, the family or better the families. In these years next to my Italian family I also received a precious support and love from my Egyptian family. Thank you Nadia, Samir, Hadeer, and Nurhan. Thank you Ahmed, Mona and Malek. Thank you Salah, Ghadra, Mahmoud and Asmah. Thank you Abir, Nagua and their families. I feel constantly your love and encouragement despite the distance.
Thank you to my ”mamma”, that helps me concretely and constantly in everything that I am doing, even if (and especially if) I am far away from her. Thank you to my ”papa’”, that stands by my choices, although they are difficult. Thank you to my brother Simone that makes me remember where I come from, and how I was, and my sister Sara, that wherever I go she comes! Francesco, thank you to be present in the important moments! Then, thank to all the members of Cirillo’ s and Pecoraro’ s families. You are really the best gift that life could give me, I am very lucky to have you behind me for support!
Finally, I let the last lines for you Wael and my girls Nadine and Mila. I am grateful to share my life with all of you!
Wael you showed me the love and support in so many ways, that sometimes I even do not realize it. Starting with the choice to put in a side your career and life in Utrecht, for facilitating me in doing and finishing the PhD. This thesis is for you!
Nadine and Mila, thank you to be patient especially in these last moments of changes, and thank you even more to bring me everyday in the beautiful world of the dreams, where magic happens!

ABOUT THE AUTHOR

Elisa Cirillo was born on 4th June 1989 in Salerno, Italy. A city located in one of the most popular gulfs of the Mediterranean sea. However, she grew up in Feltre located in the valleys of the italian north-east Dolomites. Here she performed and concluded her high school studies at Liceo Scientifico G. Dal Piaz in 2008. Then, she moved to Bologna for attending a Bachelor study in Biotechnology, and she performed the thesis research at the James Hutton Institute in Dundee, Scotland. In 2011 she graduated with the maximum mark. In 2011 she started to attend a Master study in Bologna, in Pharmaceutical Biotechnology. Also at this time, she performed the thesis research abroad, at TNO in Zeist, The Netherlands. She received a graduation with honors in November 2013. Finally, she engaged for a PhD project at the department of Bioinformatics (BiGCaT) in Maastricht University, having the possibility to improve both her scientific attitude and the professional soft skills. Currently, she is working as Business Analyst and Product Owner at the Hyve in Utrecht, an IT company that supports researchers with open source software.

LIST OF PUBLICATIONS

1. Cirillo E,Kutmon M,Gonzalez Hernandez M,Hooimeijer T,Adriaens ME,Eijssen LMT,Parnell LD,Coort SL,Evelo CT (2017) From SNPs to pathways:Biological interpretation of type 2 diabetes(T2DM)genome wide association study(GWAS)results. PlosOne. DOI:10.1371/journal.pone.0193515.

2. Cirillo E,Parnell LD,Evelo CT (2017) A Review of Pathway-Based Analysis Tools That Visualize Genetic Variants. Frontiers in Genetics. DOI: 10.3389/fgene.2017.00174.

3. Grimaldi KA,van Ommen B,Ordovas JM,Parnell LD,Mathers JC,Bendik I,Brennan L,Celis-Morales C,Cirillo E,Daniel H,de Kok B,El-Sohemy A,Fairweather-Tait SJ,Fallaize R,Fenech M,Ferguson LR,Gibney ER,Gibney M,Gjelstad IMF,Kaput J1,Karlsen AS,Kolossa S,Lovegrove J,Macready AL,Marsaux CFM8,Alfredo Martinez J,Milagro F,Navas-Carretero S,Roche HM,Saris WHM,Traczyk I,van Kranen H,Verschuren L,Virgili F,Weber P,Bouwman J. (2017) Proposed guidelines to evaluate scientific validity and evidence for genotype-based dietary advice. Gene and Nutrition. DOI: 10.1186/s12263-017-0584-0.

4. Slenter DN,Kutmon M,Hanspers K,Riutta A,Windsor J,Nunes N,Mlius J,Cirillo E,Coort SL,Digles D,Ehrhart F,Giesbertz P,Kalafati M,Martens M,Miller R,Nishida K,Rieswijk L,Waagmeester A,Eijssen LMT,Evelo CT,Pico AR,Willighagen EL. (2017) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. DOI: 10.1093/nar/gkx1064.

5. Siddiqa A,Cirillo E,Tareen SHK,Ali A,Kutmon M,Eijssen LMT,Ahmad J,Evelo CT,Coort SL. (2017) Visualizing the regulatory role of Angiopoietin-like protein 8 (ANGPTL8) in glucose and lipid metabolic pathways. Genomics. DOI: 10.1016/j.ygeno.2017.06.006.

6. Ehrhart F, Coort SL, Cirillo E, Smeets E, Evelo CT, Curf L (2016) Rett syndrome biological pathways leading from MECP2 to disorder phenotypes. Orphanet Gen Rare Dis. DOI: 10.1186/s13023-016-0545-5.

7. Ehrhart F, Coort SL, Cirillo E, Smeets E, Evelo CT, Curf L (2016) New insights in Rett syndrome using pathway analysis for transcriptomics data. Wien Med Wochenschr. DOI: 10.1007/s10354-016-0488-4.

8. Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, Malius J, Waagmeester A, Sinha SR, Miller R, Coort S, Cirillo E, Smeets B, Evelo CT, Pico AR. (2016) WikiPatways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. DOI:10.1093/nar/gkv1024.

9. Houston K, McKim SM, Comadran J, Bonar N, Druka I, Uzrek N, Cirillo E, Guzy-Wrobelska J, Collins NC, Halpin C, Hansson M, Dockter C, Druka A, Waugh R. (2013) Variation in the interaction between alleles of HvAPETALA2 and microRNA172 determines the density of grains on the barley inflorescence. Proc Natl Acad Sci USA. DOI:10.1073/pnas.1311681110.