Development of network-based analysis methods with application to the genetic component of asthma

Yuanlong Liu

Université Sorbonne Paris Cité

Thèse de doctorat de l'Université Sorbonne Paris Cité

Préparé à l'Université Paris Diderot

ÉCOLE DOCTORALE PIERRE LOUIS DE SANTÉ PUBLIQUE À PARIS ÉPIDÉMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMÉDICALE

(ED 393)

Unité de recherche: UMR 946 - Variabilité Génétique et Maladies Humaines

DOCTORAT

Spécialité: Epidémiologie Génétique

Yuanlong LIU

Development of network-based analysis methods with application to the genetic component of asthma

Thèse dirigée par Florence DEMENAIS

Présentée et soutenue publiquement à Paris le 13 Novembre 2017

JURY

M. Bertram MÜLLER-MYHSOK	Professeur, Technische Universität München	Rapporteur
Mme Kristel VAN STEEN	Professeur, Université de Liège	Rapporteur
M. Benno SCHWIKOWSKI	Directeur de Recherche, Institut Pasteur	Examinateur
M. Mohamed NADIF	Professeur, Université Paris-Descartes	Examinateur
M. Laurent ABEL	Directeur de Recherche, INSERM	Président du jury
Mme Florence DEMENAIS	Directrice de Recherche, INSERM	Directrice de thèse

ABSTRACT

Genome-wide association studies (GWAS) of asthma have been successful in identifying novel asthma-associated loci, but the genes at these loci account only for a part of the whole genetic component. One limitation of GWAS is that they rest on single-marker analyses which are underpowered to detect variants with small marginal effects but rather influence jointly on disease risk. To complement the single-marker approaches, more sophisticated strategies, which integrate biological knowledge, such as protein-protein interactions (PPI) or gene networks with GWAS outcomes to identify disease-associated gene modules, have become prominent. The objectives of this thesis were to develop network-based analysis methods, and apply them to asthma GWAS data to identify biological processes and prioritize new candidate genes related to asthma.

This thesis consists of two main studies. The first study was to extend an existing network-based method (dmGWAS) to identify novel genes associated with asthma. We used two GWAS datasets, each consisting of the results of a meta-analysis of nine childhood-onset asthma GWAS (5,924 and 6,043 subjects, called META1 and META2, respectively). We developed a novel method to compute gene-level *p*-values from SNP *p*-values (fastCGP), and proposed a bi-directional module search method to identify asthma-associated gene modules. Application of these methods to the asthma data detected a gene module of 91 genes significantly associated with asthma ($p < 10^{-5}$). This module consisted of a core network and five peripheral subnetworks including high-confidence candidates for asthma. Out of the 91 genes, 19 genes were nominally significant in both META1 and META2 datasets. They included 13 genes at 4 loci previously found associated with asthma (2q12, 5q31, 9p24.1, 17q12-q21), and six genes at six novel loci: *CRMP1* (4p16.1), *ZNF192* (6p22.1), *RAET1E* (6q24.3), *CTSL1* (9p21.33), *C12orf43* (12q24.31) and *JAK3* (19p13-p12). Functional analysis of the module revealed four functionally related gene clusters involved in innate and adaptive immunity, chemotaxis, cell-adhesion and transcription regulation, which are biologically meaningful processes underlying asthma risk.

The second study of this thesis was to develop a novel network-based method, named SigMod, to search disease-associated gene modules. SigMod takes a list of gene *p*-values and a gene network as input. It identifies a set of genes that are enriched in high association

signals and tend to have strong interconnection via the formulation of a binary quadratic optimization problem. We proposed an algorithm based on graph-cut theory to solve the optimization problem exactly and efficiently. SigMod has several advantages compared to existing methods, including the ability to find the module enriched in highest association signals, the capacity to incorporate edge weights in the network, and the robustness to background noise. Also, the emphasis of selecting strongly interconnected genes can lead to the identification of genes with close functional relevance. We applied SigMod to both simulated and real datasets. This new method outperformed existing approaches. When SigMod was applied to childhood-onset asthma data, it successfully identified a module made of 190 functionally related genes that are biologically relevant for asthma.

RESUME

Les études d'association pan-génomiques (GWAS) ont permis d'identifier de nouveaux locus associés à l'asthme, mais ces loci n'expliquent qu'une partie de la composante génétique de cette maladie. Une limite de ces études est qu'elles sont basées sur des analyses simple-marqueurs qui manquent de puissance pour détecter des variants génétiques à effet marginal faible et influençant conjointement le risque de maladie. Des stratégies, qui intègrent des connaissances biologiques, comme les interactions protéine-protéine (PPI) ou des réseaux de gènes avec des résultats de « GWAS », ont été proposées pour identifier des modules de gènes associés aux maladies. Les objectifs de cette thèse étaient de développer des méthodes d'analyse de réseaux de gènes, et de les appliquer à des données pan-génomiques de l'asthme pour identifier de nouveaux gènes candidats et des processus biologiques potentiellement impliqués dans l'asthme.

Le premier travail de thèse a consisté à étendre une méthode de recherche de réseau de gènes à partir de données de « GWAS » (dmGWAS) pour identifier de nouveaux gènes associés à l'asthme. Nous avons utilisé deux jeux de données, chacun correspondant aux résultats d'une méta-analyse de neuf études d'association pan-génomiques de l'asthme de l'enfant (5,924 et 6,043 sujets, et appelés META1 et META2). Nous avons développé une nouvelle méthode pour calculer les *p*-valeurs de chaque gène à partir des *p*-valeurs des SNPs et proposé une stratégie de recherche bidirectionnelle à partir des deux jeux de données pan-génomiques pour identifier un module de gènes. Nous avons détecté un module de 91 gènes associé à l'asthme ($p < 10^{-5}$). Ce module est composé d'un réseau central et de cinq réseaux périphériques. Parmi les 91 gènes, 19 gènes étaient nominalement significatifs ($p < 0.05$) dans les deux jeux de données et incluaient 13 gènes à 4 loci trouvés précédemment associés à l'asthme (2q12, 5q31, 9p24.1, 17q12-q21), et six gènes à six nouveaux loci: *CRMP1* (4p16.1), *ZNF192* (6p22.1), *RAET1E* (6q24.3), *CTSL1* (9p21.33), *C12orf43* (12q24.31) et *JAK3* (19p13-p12). L'analyse fonctionnelle du module identifié a révélé quatre clusters de gènes impliqués dans l'immunité innée et adaptative, la chimiotaxie, l'adhésion cellulaire et la régulation de la transcription, qui sont des processus biologiquement pertinents pour l'asthme.

Le deuxième travail de thèse a consisté à développer une nouvelle méthode de réseau de gènes appelée SigMod. SigMod permet de sélectionner un module de gènes enrichis en

signaux d'association avec la maladie et montrant de fortes inter-connexions. Par rapport aux méthodes précédentes SigMod offre plusieurs avantages, notamment la robustesse au bruit de fond, la capacité de prendre en compte une pondération sur les liens entre gènes, et de rendre les résultats facilement interprétables. Nous avons proposé un algorithme basé sur la théorie des découpages de graphes pour résoudre le problème d'optimisation de manière exacte et efficace. Des simulations ont montré une meilleure performance de SigMod par rapport aux méthodes existantes. L'application de SigMod aux données de l'asthme a permis d'identifier un module de 190 gènes qui présentent des relations fonctionnelles et sont biologiquement pertinents pour l'asthme.

Acknowledgements

Foremost, I would like to express my deep gratitude to my supervisor Florence Demenais. I have been so lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so detailedly. I benefited enormously from her continued support, immense knowledge, enthusiasm, and encouragement. Her conscientious supervision is indispensable for me to conduct my research work and also to write this thesis.

I am also so grateful to the lad members of Inserm U946. They are my respectable colleagues and cordial friends. Emmanuelle Bouzigon is always ready to offer advices with a smiling face. Amaury Vaysse explained me time and again the biological knowledge that I needed for my projects. Chloé Sarnowski, Myriam Brossard and Martine Bothua offered countless help for both my research work and personal life. Myriam kindly assisted for the proof reading of many of my documents. It was enjoyable to discuss and work with her. Pierre Emmanuel Sugier, Anthony Herzig and Hamida Mohamdi are among those that we shared lots of discussion, amusement and happiness. I would also thank Patricia Jeannin, Nolwenn Lavielle, Marie-Hélène Dizier and all the other members in our lab for their advices and inspirations.

I would express my sincere appreciation to Karsten Borgwardt and again to Florence Demenais. They brought me to the Marie Curie Initial Training Network "Machine Learning for Personalized Medicine", so that I had the chance to connect with the leading experts in this area. I would thank Kristel Van Steen, Bertram Müller-Myhsok, who are among the PIs of this network and offered me precious advices on research work and career development. I also gained much knowledge through the discussion with Fabian Heinemann, Felipe Llinarez López, Yi Zhong, Cankut Çubuk and all other members in this network.

I would thank my friends as well. They are scattered around the world but are always reachable whenever I have happiness to share or sorrows to pour out. They are strong supporters throughout my road of study. Surely, I will not forget to thank my family: my mother Yi Qiqiong, my father Liu Xingcai, my brother Liu Hongquan, and all my relatives. Thank you for providing me a spiritual habitat throughout my long journey of exploration and pursuit for dreams. Thank you so much for always being together with me.

Lastly, I would acknowledge the FP7-Marie Curie Initial Training Network Grant "Machine Learning for Personalized Medicine" (Grant No. 316861) which supported this Ph.D work.

SCIENTIFIC PRODUCTIONS DURING THE PHD

Articles published in peer-reviewed journals

Liu, Y., Brossard, M., Roqueiro, D., Margaritte-Jeannin, P., Sarnowski, C., Bouzigon, E., & Demenais, F. (2017). SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. *Bioinformatics*, 33(10), 1536-1544.

Liu, Y., Brossard, M., Sarnowski, C., Vaysse, A., Moffatt, M., Margaritte-Jeannin, P., Llinares-López, F., Dizier, M.H., Lathrop, M., Cookson, W., Bouzigon, E. & Demenais, F. (2017). Network-assisted analysis of GWAS data identifies a functionally-relevant gene module for childhood-onset asthma. *Scientific Reports*, 7(1), p.938-947 (online open access journal)

Article in preparation

Liu, Y. & Demenais, F. SeedMod: a network-based method to identify functional gene modules under the guidance of known disease genes

Oral communications

Liu, Y., Brossard, M., Margaritte-Jeannin P., Llinares, F., Sarnowski, C., Al-Shikley, L., Lavielle, N., Vaysse, A., Dizier, M.H., Bouzigon, E., Demenais. F. Integration of network resources to optimize genetic association studies, *European Society of Human Genetics Annual Meeting, Barcelona*, Spain, 21-24 May 2016

Liu, Y., Brossard, M., Margaritte-Jeannin P., Sarnowski, C., Al-Shikley, L., Lavielle, N., Vaysse, A., Dizier, M.H., Bouzigon, E., Demenais. F. Network-assisted methods to identify genetic variants underlying asthma, *Symposium on Machine Learning for Personalized Medicine*, Barcelona, Spain, 19-20 May 2016

Liu, Y., Brossard, M., Margaritte-Jeannin P., Llinares, F., Sarnowski, C., Al-Shikley, L., Lavielle, N., Vaysse, A., Dizier, M.H., Bouzigon, E., Demenais. F. Integration of genome-wide association data and human protein interaction networks identifies a gene sub-network

underlying childhood-onset asthma, *American Society of Human Genetics Annual Meeting*, Baltimore, USA, 6-10 Oct 2015.

Liu, Y., Brossard, M., Margaritte-Jeannin P., Llinares, F., Sarnowski, C., Al-Shikley, L., Lavielle, N., Vaysse, A., Dizier, M.H., Bouzigon, E., Demenais. F. Network-assisted investigation of signals from genome-wide association studies in childhood-onset asthma, *Capita Selecta in Complex Disease Analysis Conference*, Liège, Belgium, 24-26 Nov 2014

Poster communications

Brossard, M., **Liu, Y.**, Vaysse, A., Mohamdi, H., Maubec E., Avril M.F., Demenais, F. The SigMod network analysis method identifies gene modules for cutaneous melanoma and nevus count that share relevant candidates. *International Genetic Epidemiology Society Annual Meeting*, Cambridge, UK, 9-11 Sep 2017

Brossard, M., Vaysse, A., Mohamdi, H., **Liu, Y.**, Demenais, F. Genetic analysis of the telomere interactome pinpoints new candidate genes for melanoma risk. *International Genetic Epidemiology Society Annual Meeting*, Toronto, Canada, 24-26 Oct 2016

Liu, Y., Brossard, M., Roqueiro, D., Margaritte-Jeannin, P., Sarnowski, C., Bouzigon, E., & Demenais, F. A novel network method (SigMod) identifies a strongly interconnected gene module associated with childhood asthma. *European Society of Human Genetics Annual Meeting*, Barcelona, Spain, 21-24 May 2016.

Liu, Y., Brossard, M., Margaritte-Jeannin P., Llinares, F., Sarnowski, C., Al-Shikley, L., Lavielle, N., Vaysse, A., Dizier, M.H., Bouzigon, E., Demenais. F. Network-based analysis of GWAS data identifies a gene sub-network underlying childhood-onset asthma. *International Genetic Epidemiology Society Annual Meeting*, Baltimore, USA, 4-6 Oct 2015

TABLE OF CONTENTS

CHAPTER I. INTRODUCTION

1 Human genetic variation

1.1 The human genome

The human genome is the complete set of nucleic acid sequence for humans. It is encoded as DNA within chromosomes in cell nuclei and in a small molecule in individual mitochondria. The DNA molecular consists of two strands and has a "double helix" structure. Each strand consists of an assembly of basic building blocks called nucleotide or nucleotide base. Nucleotides in DNA contain four different bases: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). The pairing of two nucleotides by hydrogen bonds forms a base pair (bp). Adenine always pairs with Thymine (forming the A/T pair); Guanine always pairs with Cytosine (forming the G/C pair) (Figure 1.1).

The total length of the human genome is about 3 billion base pairs. There are 23 pairs of human chromosomes: 22 pairs of autosomal chromosomes and one pair of sex chromosomes. Somatic cells usually have one copy of chromosome 1-22 inherited from each parent, one X chromosome inherited from the mother, and another X or Y chromosome inherited from the father. These chromosomes contain the genetic blueprint for building a human being.

1.2 The genes

A gene is a sequence of nucleotides along a segment of DNA (Figure 1.2). On average, a gene is 10-15kb (1kb=1,000 base pairs) long, but this size can vary greatly from ~0.2kb (Tyrosine tRNA gene) to ~2,500kb (DMD dystrophin gene). Each person has two copies of each gene that are inherited from each parent. The number of human protein-coding genes are estimated to be 19,000 to 20,000 (Ezkurdia et al., 2014).

Every gene consists of a protein coding region, which begins with a *Start* codon and concludes with a *Stop* codon, and might be contiguous or broken up into a series of introns and exons (Figure 1.2). Every gene also contains regulatory sequences flanking the open reading frame (the part of a reading frame that has the potential to be translated), which can expand many kilobases upstream or downstream of the open reading frame. These are

stretches of DNA that do not themselves code for protein but act as binding sites for RNA polymerase and its accessory molecules as well as transcription factors. A promoter is a regulatory element that the RNA polymerase initially binds before starting the transcription of the DNA into RNA. The binding and activation of the RNA polymerase is controlled by transcription factors which bind the promoters and cis-regulatory sequences conventionally referred to as *Enhancer* and *Silencer*.

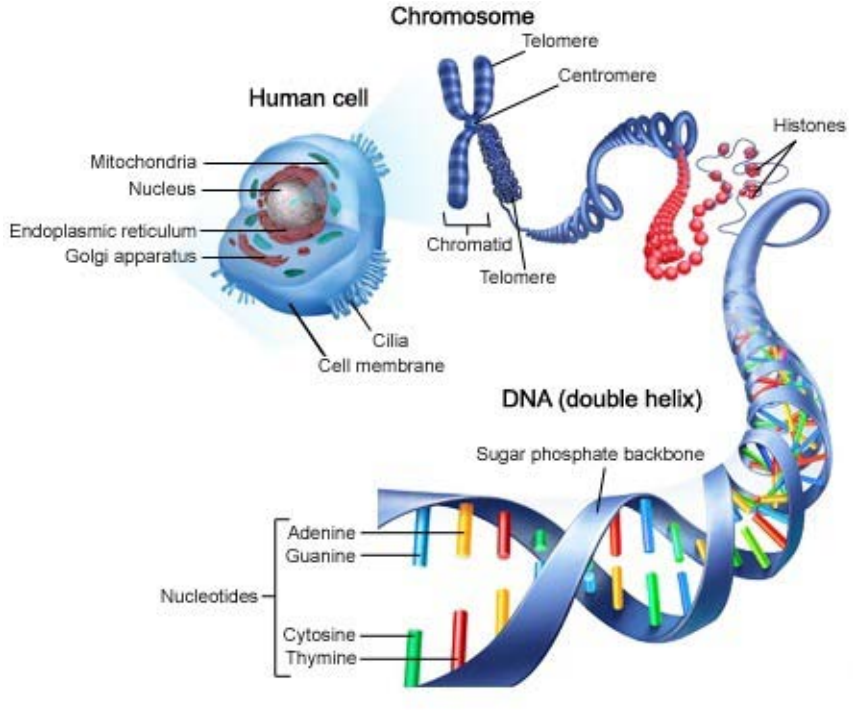


Figure 1.1: The structure of DNA. From http://sciencewithmrsb.weebly.com/genetic-variation.html.

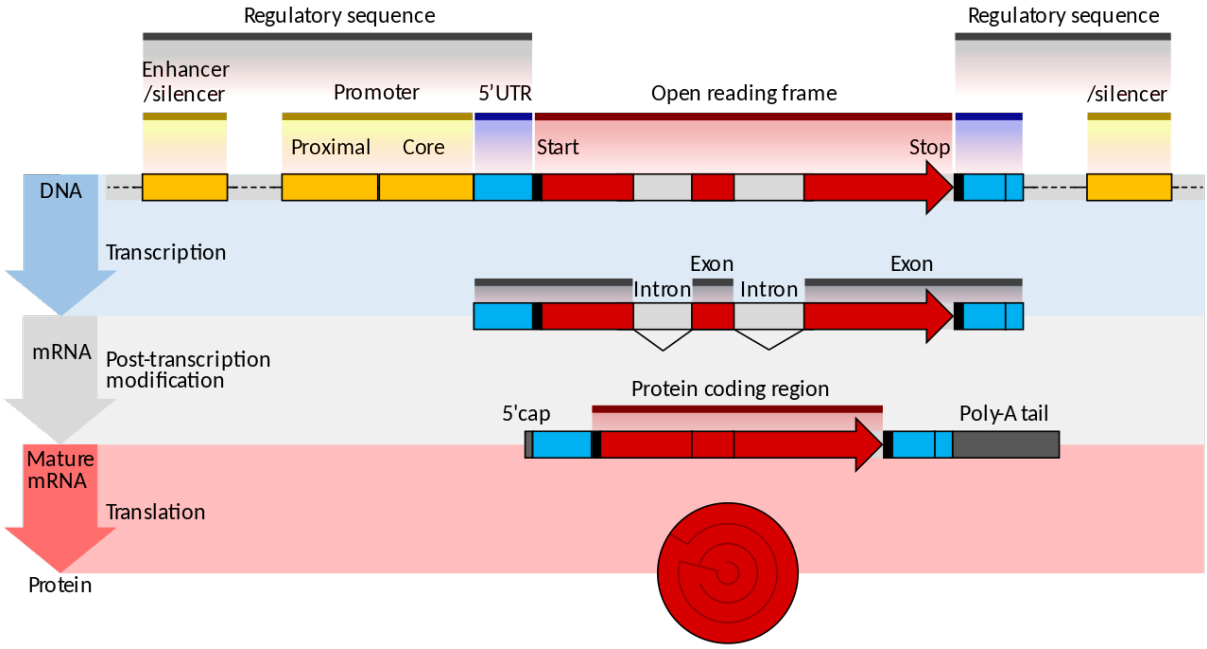


Figure 1.2: The structure of a eukaryotic protein-coding gene. Regulatory sequence controls when and where expression occurs for the protein coding region (red). Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to remove introns (light grey) and add a 5' cap and poly-A tail (dark grey). The mRNA 5' and 3' untranslated regions (blue) regulate translation into the final protein product. From https://en.wikipedia.org/wiki/Gene.

1.3 Human genetic variation

Variation of human genome arises from point mutation (single base modification), base pair insertion/deletion (indel), chromosome rearrangement and gene copy-number variation. Each form of variation at a given point in the genome is called an allele. The two alleles at the same position on homologous chromosomes form the genotype of an individual.

The most common type of genetic variation among people is single nucleotide polymorphism, abbreviated as SNP. A SNP is defined as a variation in a single nucleotide that occurs at a specific position in the genome. For example, a SNP may replace the nucleotide Cytosine with the nucleotide Thymine in a certain stretch of DNA, as depicted in Figure 1.3. SNPs occur frequently throughout a person's DNA. On average, there is one SNP in every 300 nucleotides, which indicates there are around 10 million SNPs in the human genome.

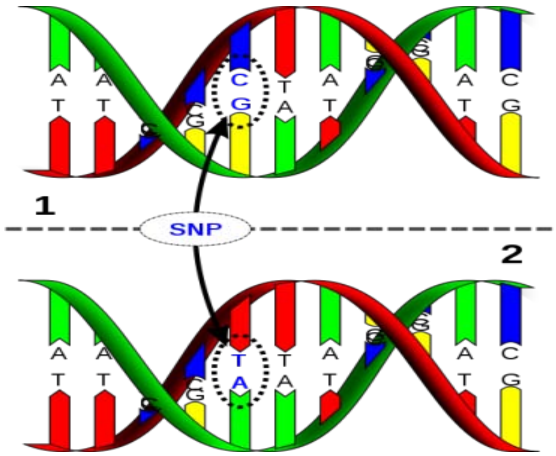


Figure 1.3: An illustrative example of SNP. The upper DNA molecule differs from the lower DNA molecule at a single base-pair location (a C/T polymorphism). From http://isogg.org/wiki/Single-nucleotide_polymorphism.

On average, the proportion of nucleotides that differ between two individuals is estimated to be 0.1% (Jorde et al., 2004) to 0.4% (Tishkoff et al., 2004) of the whole base pairs. The 1000 Genomes Project was set out to provide a comprehensive description of common human genetic variations by applying whole-genome sequencing to 2,504 individuals from 26 populations (The 1000 Genomes Project Consortium, 2015). The completion of the project has characterized in total over 88 million variants, including 84.7 million SNPs, 3.6 million short insertions/deletions, and 60,000 structural variants.

1.4 Linkage disequilibrium

Linkage disequilibrium (LD) describes the non-random association of alleles at different loci in a population. Consider two biallelic loci $Locus_1$ and $Locus_2$, the two alleles at these loci are *a/A* and *b/B* respectively. The relationship between the frequencies of gametes carrying each allele and allele pairs (known as the haplotype) is summarized in Table 1.1, where $f_{(\bullet)}$ represents the frequency of an allele or a haplotype. The level of LD between allele A and allele B can be quantified by a statistic defined as

$$D_{AB} \coloneqq f_{AB} - f_A \times f_B.$$

Table 1.1: 2 × 2 table of allele and haplotype frequencies at two loci. $f_{(\bullet)}$ represents the frequency of an allele or a haplotype.

$Locus_1$ / $Locus_2$	a	A	Total
b	f_{ab}	f_{Ab}	f_b
B	f_{aB}	f_{AB}	f_B
Total	f_a	f_A	1

Through the relationship among the frequencies described in Table 1.1, it can be deduced that

$$D_{AB} = -D_{aB} = -D_{Ab} = D_{ab}.$$

Therefore any of these four statistics is sufficient to characterize the LD between the alleles at the two loci. The indications of a value of D_{AB} are given as below

- $D_{AB} = 0$, i.e., $f_{AB} = f_A \times f_B$: A and B are in complete linkage equilibrium
- $D_{AB} \neq 0$, i.e., $f_{AB} \neq f_A \times f_B$: A and B are in linkage disequilibrium
- $D_{AB} > 0$, i.e., $f_{AB} > f_A \times f_B$: A and B are preferentially associated
- $D_{AB} < 0$, i.e., $D_{Ab} > 0$: A and b are preferentially associated

It is of note that although D_{AB} is a measure of the extent to which two alleles are associated, it is not always the best statistic to be used because the range of its possible values are constrained by the allele frequencies. The smallest possible value of D_{AB} is $\max\{-f_A f_B, -(1-f_A)(1-f_B)\}$, while its largest possible value is $\min\{f_A(1-f_B), f_B(1-f_A)\}$. This makes it less favorable to compare the LD between different loci. Two alternate measures have been proposed. They are the D' (Slatkin, 2008):

$$D' = \begin{cases} \dfrac{D_{AB}}{\min(f_A \times f_B, f_a \times f_b)} & \text{if } D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(f_a \times f_B, f_A \times f_b)} & \text{if } D_{AB} > 0 \end{cases}$$

and the r^2 (also called Δ^2) (Hill et al., 1968):

$$r^2 = \frac{D_{AB}{}^2}{f_a \times f_A \times f_b \times f_B}.$$

The definition of D' has the convenient property that it indicates at least one of the four possible haplotypes is absent when $|D'|=1$, a situation commonly described as complete linkage disequilibrium. The r^2 statistics is a measure of the correlation between allele A and allele B (ranges from 0 to 1). When $r^2 = 1$, there is perfect linkage disequilibrium, which means only two of the four possible genotypes are present in the population. As a result, the two loci have the same allele frequencies.

2 Exploring the genetic component of human diseases

Researchers are learning that many human diseases have a genetic component. Some diseases, such as sickle-cell anemia, Tay-Sachs disease, xeroderma pigmentosa and cystic fibrosis, arise from the change or alteration in a single gene, and are inherited according to Mendel's law (Riordan et al., 1989). These diseases often cluster in families and can be predicted based on the medical history with the help of a family tree. The causes of many other diseases, however, are much more complex. Common medical problems such as inflammatory bowel disease (IBD), diabetes, Alzheimer's disease, asthma, and many chronic disorders do not have a single genetic cause—they are likely to be associated with the effects of multiple genes in combination with lifestyle and environmental exposure. This complex mechanism is illustrated in an example given in Figure 1.4. The diseases that are caused by many contributing factors are called complex or multifactorial diseases. Most of the multifactorial diseases are common in the population and represent a major challenge for public health.

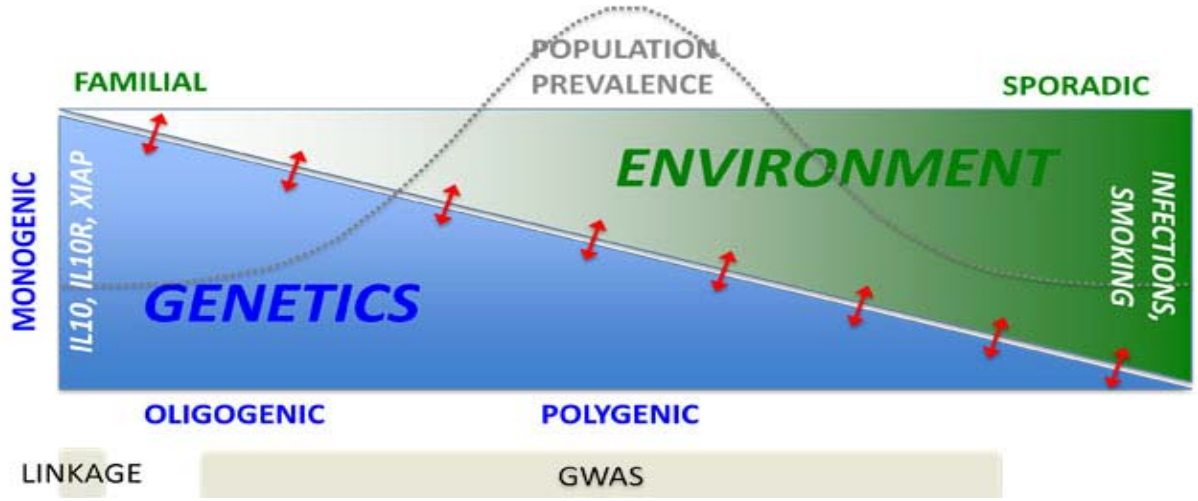


Figure 1.4: Inflammatory bowel disease (IBD) as an example of a complex disease. As indicated by the dashed line, IBD cases that arise from the change or alteration in a single gene are rare, and often cluster in families. Instead, most of the cases are associated with the effects of multiple genes in combination with environmental factors. Particularly, the gene environmental interactions (red arrows) play an important role in disease susceptibility, revealing the complexity of the disease mechanism. Figure adopted from http://www.genes-environment-inflammation.de/rtg/vision.

The methodologies employed to understand the role of genetic component in human diseases have evolved in recent years due to technological advances and accumulation of biological

knowledge. The general principle of these methods is to evaluate the correlation between genetic variants and the disease under study. Two of the major analysis methodologies are linkage study and association study, as will be described below.

2.1 Genetic linkage study

A genetic linkage study is a family-based method used for mapping a disease trait to a genomic location by demonstrating co-segregation of the disease with genetic markers at a known chromosomal location. It is based on the observation that genetic markers residing physically close on a chromosome tend to remain linked during meiosis. Linkage study is a powerful tool to detect the chromosomal location of disease genes, and has been employed to identify a number of genes involved in monogenic Mendelian diseases (Genin et al., 2008). For example, by genotyping family members affected by cystic fibrosis using a collection of genetic markers across the genome, and examining how those genetic markers segregate with the disease across multiple families, researchers have identified multiple mutations in the *CFTR* (Cystic fibrosis trans-membrane conductance regulator) gene as the cause of cystic fibrosis (Kerem et al., 1989). Linkage studies were also proven to be powerful in discovering some variants that contribute to familial forms of multifactorial diseases, from neurodegenerative diseases such as Alzheimer, Parkinson, to tumour syndromes such as neurofibromatosis type 1 and type 2 (Pulst, 1999). However, they are less suited for the study of multifactorial diseases as a whole (Khoury et al., 1998; Risch et al., 1996). The lack of success can be attributed to various factors, but mainly to its limited power in pinpointing genetic factors that have moderate or low effect (level of marker-trait correlation), and the complex mechanisms (gene-gene interactions, gene-environment interactions etc.) (Tabor et al., 2002). Moreover, the relatively high prevalence of these multifactorial diseases suggested that the risk alleles are common in the general population, raising the "Common Diseases-Common Variants" hypothesis that motivates researchers to conduct genetic analyses at genome-wide scale (Schork et al., 2009) (although this hypothesis has long been debated).

2.2 Genetic association study: from candidate gene study to genome-wide study

Genetic association study aims at finding genetic variants or genomic regions that are associated with disease susceptibility by means of testing their correlation with the disease status. For a binary trait (affected/unaffected or case/control), a significantly higher frequency

of a SNP allele in the disease-affected group can be interpreted as that the tested genetic variant is associated with the disease risk.

The first wave of association studies, applied in the 1990s and early 2000s, were focused on candidate genes. Most of the candidate genes were selected because they are either functional candidates (i.e., they encode a protein implicated by an etiological hypothesis), or positional candidates located in chromosomal regions implicated by previous linkage studies. Candidate gene association studies were relatively cheap and quick to perform at that time (Patnala et al., 2013). One limitation of candidate gene approaches is that they rely heavily on the basis of biological hypothesis or the location of candidate within a previously defined region of linkage. Therefore, results gained from these "hypothesis-driven" approaches depend on the ability to select plausible candidates from the genome.

Advantageously, genome-wide association studies (GWAS), such as the pioneering work conducted by Klein et al. (2005), allow a systematic, comprehensive survey of genetic variants (SNPs) in the entire genome, and in a hypothesis-free manner. The extension from candidate gene approach to genome-wide approach has become realistic thanks to the fast growing understanding of human genome, the advancement in micro-array and sequencing technologies, and the abundance of analysis tools. One crucial advance that enables efficient genome-wide studies is the characterization of LD patterns across the genome. LD has an important role in the selection of SNPs for performing GWAS. For a chromosome region with known LD pattern, a few tag SNPs can be chosen such that they capture most of the common variations within that region (Frazer et al., 2007; Hirschhorn et al., 2005). Consequently, the disease-association of a genotyped SNP is tested directly, while the association of a SNP that is not genotyped but in LD with the genotyped SNPs can be tested indirectly (as illustrated in Figure 1.5). The progress of international HapMap project (Frazer et al., 2007), the 1000 Genomes Project (Genomes Project Consortium, 2010), and recently the Haplotype Reference Consortium (HRC) (Haplotype Reference Consortium, 2016) have enabled to elucidating common human genetic variants and LD patterns across the genome in various populations. Today, comprehensive catalogs of SNPs are deposited in public databases and are available for use without much restriction. Individual genotyping was also made possible along with the availability of advanced chip-based microarray technology. Two primary platforms have been used in most GWAS—Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA).

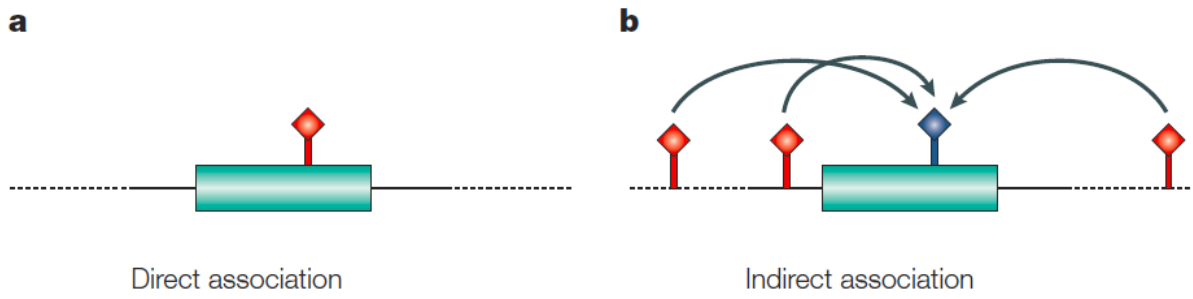


Figure 1.5: Testing SNPs for disease-association via direct or indirect association. (a) A case in which a genotyped SNP (red diamond) is tested for association with a disease trait directly. (b) A case in which an ungenotyped SNP (blue diamond) is tested for association with a disease trait indirectly, as it is in LD with the other three genotyped SNPs. Adopted from Hirschhorn and Daly (2005).

GWAS has experienced tremendous success since its first publication in 2005 on a study of age-related macular degeneration (Klein et al., 2005). Up to the time of March 2017, 2,518 human GWAS have been conducted. These studies examined more than 280 diseases or traits. Hundreds or thousands of individuals have been involved in these studies. More than 24,000 SNP-trait associations have been found (Figure 1.6). These results are collected in a GWAS catalog established by National Human Genome Research Institute (NHGRI) and European Bioinformatics Institute (EBI) that are available at https://www.ebi.ac.uk/gwas/ (MacArthur et al., 2017; Welter et al., 2014).

Figure 1.6: Published GWAS results for 17 trait categories. Figure retrieved from https://www.ebi.ac.uk/gwas/. Accessed in March 2017.

The GWAS approaches are primarily based on single-SNP analysis of hundreds of thousands to millions of SNPs. They survey each SNP one by one for their association with the disease trait under study. In the following, we describe several of the major components involved in the procedure of conducting a GWAS analysis.

Data pre-processing. Data pre-processing is an important step prior to perform association analysis. Essentially, quality control (QC) procedures should be conducted at the first stage. These include data filtering at both SNP-level and sample-level. SNP-level filtering aims at removing SNPs that have low variability, high genotyping error, or a large amount of missing data. Typically, SNPs with a call rate less than 95% (missing in more than 5% samples) are removed. SNPs having a minor allele frequency (MAF) less than 1%, which may result in inadequate statistical power or false positive results (if exact tests are not performed) in downstream association analysis are also excluded. The existence of genotyping error of a SNP is examined by testing for derivation from Hardy-Weinberg equilibrium (HWE) using a one degree-of-freedom Pearson goodness-of-fit test, often known as the χ^2 test (Reed et al., 2015). SNPs for which the HWE test have a *p*-value less than a certain threshold (for example

1×10^{-4}) are excluded for downstream analysis. Sample-level filtering aims at removing individuals due to sample contamination, missing data, correlation (for population-based investigations), and racial/ethnic or gender ambiguity or discordance. Criteria used for sample-level filtering were described in, for example, Anderson et al. (2010). The exact criteria can be study-dependent (Reed et al., 2015).

Next, the existence of population stratification (the presence of a systematic difference in allele frequencies between subpopulations possibly due to different ancestry) needs to be checked. This can be achieved by computing the principal components (PCs) of the genotype data using software such as EIGENSTRAT (Patterson et al., 2006). The computed PCs will be included as a covariate in the consequent association analysis to reduce spurious associations caused by systematic difference in allele frequencies in different populations.

Unmeasured SNPs (SNPs that are not on the chips, which often differ from one study to another) can be imputed based on reference haplotypes and their LD structure derived from extensive resources, such as the HapMap and 1000 Genomes data. Several imputation algorithms based on Markov Chain Monte Carlo (MCMC) technique have been proposed. Well-described packages for SNP imputation include BEAGLE, IMPUTE2, and MACH (Reed et al., 2015).

Choosing the genetic model. The genetic model describes the disease risk in subjects with different genotypes. Considering a genetic marker at a biallelic locus with two alleles a and A, where the risk allele a (or effect allele that may increase or decrease the risk) is often chosen as the allele that has the lower frequency among two alleles of a SNP, but can be also defined as the alternate allele compared to the reference sequence. The three possible genotypes of a subject at the locus are aa, aA, and AA. The disease penetrance associated with each genotype (denoted as γ_{aa}, γ_{aA} and γ_{AA} respectively) is the probability of getting the disease in subjects carrying that genotype. The relative risk of a genotype is the ratio of its penetrance to that of a reference genotype. To give an example, if AA is chosen as the reference genotype, the relative risks for individuals carrying aa or aA are defined as

$$RR_{aa} = \frac{\gamma_{aa}}{\gamma_{AA}} \text{ and } RR_{aA} = \frac{\gamma_{aA}}{\gamma_{AA}}, \text{ respectively.}$$

The models for disease penetrance include additive, recessive and dominant model. Their associated penetrance functions are defined as

- Additive model: $\gamma_{aa} = \gamma_{aA} \times RR = \gamma_{AA} \times RR^2$;
- Dominant model: $\gamma_{aa} = \gamma_{aA} = \gamma_{AA} \times RR$;
- Recessive model: $\gamma_{aa} = \gamma_{aA} \times RR = \gamma_{AA} \times RR$.

The additive model assumes the logarithm of the penetrance of a genotype is proportional to the number of risk alleles it has. The recessive or dominant model assume the penetrance is the same for homozygous (aa or AA) and heterozygous (aA) for a given allele (the a allele for dominant model, the A allele for recessive model). Generally, there is no accepted answer to the question of which model to use. One could choose the optimal model if the underlying mechanism is known, however, this is often not the case. A common practice is to examine the additive model, since it has reasonable power to detect both additive and dominant effects (Bush et al., 2012). Yet, an additive model may also be underpowered to detect some recessive effects (Lettre et al., 2007). Sophisticated approaches have considered performing analysis using all three models then combining their results using a weighing strategy, which could allow detecting both additive and strong non-additive effects (Balding, 2006). A general regression model that includes an additive effect and deviation from additive effect was also proposed and often used in animal and plant genetics (Wilson, 1980). The power of this general model was recently compared to that of the additive model through simulations (Dizier et al., 2017).

Statistical methods. Several statistical methods for single-marker analysis have been proposed (Balding, 2006). Quantitative traits (e.g., height, blood pressure and cholesterol level) are generally analyzed using linear models, such as linear regression and Analysis of Variance (ANOVA). These methods test the null hypothesis of no difference between the trait means in different genotype groups. A requirement for applying these methods is that the trait measurements are approximately normally distributed within each genotype group and share a common variance. Binary traits are analysed using contingency table or logistic regression approaches. The contingency table approach explores the association of a genotype with the trait via the construction of a frequency table that compares the counts of genotypes between case group and control group (Fisher, 1922). The logistic regression approach is extended

from the linear regression approach and has the goal to search for the dependence between a SNP and the probability of expressing a trait. An advantage of logistic regression is that it has the flexibility to incorporate covariates, such as age, sex, environmental exposures (exposure to the sun, tobacco etc.), and also principal components of genotype data to account for potential population stratification.

Result diagnosis. Given the penetrance model and the statistical analysis method, the association analysis can be performed conveniently using any of the available software, including PLINK (Purcell et al., 2007), STATA, and R/Bioconductor (Gentleman et al., 2004; R development core team, 2014). Prior to interpreting the outcomes, the existence of spurious associations, especially those resulting from population stratification, needs to be diagnosed. A Quantile-Quantile (Q-Q) plot that compares the observed SNP association statistics with their expected values under the null hypothesis of no association with the disease is routinely created. The observation that the majority of the SNP statistics follow the null distribution while only a handful of them deviate from it suggests there is no population structure unaccounted for when perform the analysis. This is revealed in the Q-Q plot that most of the data points fall on (or close to) the $y = x$ line. The degree of deviation from this line is measured by the genomic inflation factor λ, defined as $\lambda = Med_1 / Med_2$, where Med_1 is the median of the observed SNP statistics, and Med_2 is the median of the statistics under the null hypothesis (Devlin et al., 1999). A λ value close to 1 suggests the (potential) population substructure has been appropriately adjusted.

Multiple testing correction. There are generally hundreds of thousands to millions of statistical tests conducted simultaneously in a GWAS. With each test bearing its own false positive probability, the cumulative likelihood of finding one or more false positive associations can therefore be high. GWAS imposes a strict level of significance to reduce the number of false positives. This level is routinely determined based on Bonferroni correction, where the actual significance level α is specified as the nominal significant value of 0.05 divided by the number of SNPs that are tested (denoted as N), i.e., $\alpha = 0.05 / N$. This criterion controls strictly the family-wise error rate at 0.05 (FWER, defined as the probability of making at least one false discoveries (Thomas, 1989)). However, Bonferroni correction is known to be too conservative because the number of tests is huge and these tests are generally correlated as a consequence of LD among SNPs. This leads to over-correction and decreased

power. Alternate methods that aim at increasing power include the use of false discovery rate (FDR) criterion (e.g., $FDR \leq 0.05$) (Benjamini et al., 1995), or replacing the correction factor (N) by the number of effective (independent) tests (Li et al., 2005). Yet, because highly confident results are essential for downstream analysis and pharmaceutical operation, the significance level of 5×10^{-8} (equivalent to a nominal significant *p*-value of 0.05 after Bonferroni correction for testing one million SNPs) emerged as a standard for reporting significant associations (Jannot et al., 2015).

Replication analysis and meta-analysis. SNPs passing the significance threshold in a discovery study are urged to be replicated in one (or more) independent studies. The NHGRI working group outlined several criteria for establishing a positive replication (Chanock et al., 2007). These include using identical phenotype definition, collecting sufficient amount of replication samples, and conducting replication studies in independent datasets drawn from the same population as in the discovering study. For the purpose of increasing significance and refining effect size estimated from multiple studies, the results of multiple GWAS can be pooled together to perform a meta-analysis. Meta-analyses empower the synthesis of results from multiple studies without requiring the sharing of individual-level data—only summary statistics from a study need to be shared. Several software packages can be used to perform meta-analyses, including STATA, METAL and GWAMA (Mägi et al., 2010; Willer et al., 2010).

3 Further exploring the missing heritability of human diseases via multi-marker analysis of GWAS data

Up to now, genome-wide association studies are mainly based on single-marker analysis which requires stringent threshold (5×10^{-8}) to declare significance. Although such studies have successfully led to the identification of many genetic variants associated with complex traits, most of these variants confer relatively small increments in risk and explain only a part of the whole genetic component underlying diseases or traits, leading many to the question of how the remaining, missing heritability can be explained (Eichler et al., 2010; Manolio et al., 2009). This may potentially due to various factors (as discussed in detail in Manolio et al. (2009) and Eichler et al. (2010)), including the presence of larger number of variants of smaller effect yet to be discovered; the existence of rare variants (with MAF $\leq$ 0.01) not present on the genotyping chips or difficult to impute; the presence of complex mechanisms not taken into account (gene-gene and gene-environment interactions etc.). It can also be attributed to the joint effect of multiple SNPs, each having a low marginal effect but acting jointly on disease risk.

To address the limitations of single-SNP approaches commonly used in GWAS and to capture more of the complex genetic component underlying multifactorial diseases, many multi-marker analysis approaches have been proposed to aggregate the information of multiple SNPs into an integrative model and to study their joint effect on a disease. Multi-marker analysis provides various advantages over the single-SNP approach. First, by aggregating SNPs into sets and analyzing each set as a unit, it could reduce the number of tests thus relaxes the stringent threshold for reaching statistical significance. Secondly, by grouping SNPs properly, the power can be improved in settings where SNPs are individually only moderately significant. In particular, though any single SNP may serve as a poor surrogate of an ungenotyped SNP underlying disease susceptibility, by considering them together, it can better capture the true effect of the causal SNP. Thirdly, when there are multiple causal SNPs, conducting a joint analysis has the potential to inspect the cumulative effect of these SNPs on the disease as a whole. With these advantages, multi-marker analysis approaches are expected to discover more disease-associated variants and explain more of the missing heritability that has been missed by single-SNP analysis.

Multi-marker analysis methods have emerged over the last decade. They differ from one another according to their way of grouping SNPs, the type of data required for the analysis (individual genotype/phenotype data or GWAS summary statistics), the detailed analysis strategy they implement, etc. Based on the level of primary genetic entity that is studied, we classify multi-marker analysis into three categories: (1) SNP-based analysis, (2) gene-based analysis, and (3) pathway/network-based analysis (Figure 1.7). In the following, we first illustrate the main components involved in each of the analysis categories.

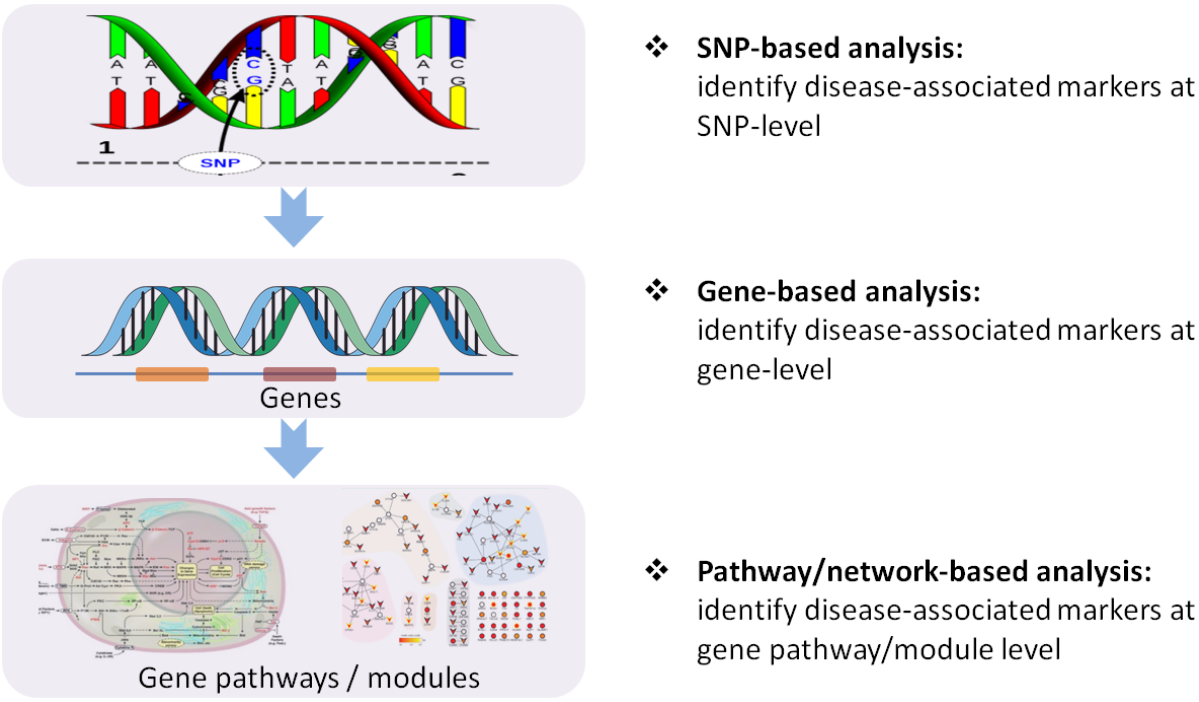


Figure 1.7: Three level of multi-marker analysis (SNP-based, gene-based, and pathway/network-based).

3.1 SNP-based multi-marker analysis

SNP-based multi-marker analysis surveys the genetic component of a disease at SNP-level. SNPs can be grouped if they are in the same gene, pathway, or a specific genomic region. In the following, we introduce several SNP-based analysis approaches that fall in the linear regression framework and machine learning framework.

Linear regression approaches. Linear regression models have the ability to search for linear combinations of SNPs that can best explain the trait. They are also flexible in incorporating

covariates (age, gender, environmental exposure etc.), in modeling the interaction between SNPs and covariates, and can provide statistical significance measure over each factor that is evaluated. Ordinary least squares (OLS) regression is a well-studied technique for modeling the relationship between a dependent variable and explanatory variables. In an OLS model for a quantitative trait, the trait Y is modeled as a linear combination of the SNPs $X_1,\dots X_p$ and possible covariates $X_{p+1},\dots X_m$ as $Y=\beta_0+\sum_{i=1}^{m}X_i\beta_i+\varepsilon$, where β_0 is the intercept and β_i $(i=1,...,m)$ are regression coefficients for the SNPs and covariates. ε is the error term. The coefficients $\boldsymbol{\beta}=(\beta_0,\beta_1,...,\beta_m)$ are routinely estimated by minimizing a loss function defined as $L_{\mathrm{OLS}}(\boldsymbol{\beta})=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{m}x_{ij}\beta_j\right)^2$, where n is the number of samples. However, OLS requires $n\geq m$, thus are generally inapplicable for GWAS data in which the number of SNPs is typically larger than the sample size. Additionally, in an OLS the estimator of regression coefficients can be highly unstable when the SNPs are correlated (De Vlaming et al., 2015).

Penalized least squares regressions, also called shrinkage methods, are more appropriate for multi-marker regression analyses. In a shrinkage method, the loss function $L(\boldsymbol{\beta})$ is usually defined as $L(\boldsymbol{\beta})=L_{\mathrm{OLS}}(\boldsymbol{\beta})+P(\lambda,\boldsymbol{\beta})$, in which $P(\lambda,\boldsymbol{\beta})$ is a penalty with a tuning parameter λ. It has the effect of shrinking the coefficients of SNPs that are less correlated with the phenotype towards zero. There are many types of penalties, including $P_{\mathrm{Ridge}}(\boldsymbol{\beta},\lambda)=\lambda\sum_{i=1}^{m}\beta_i^2$ for Ridge regression (Hoerl et al., 1970), $P_{\mathrm{Lasso}}(\boldsymbol{\beta},\lambda)=\lambda\sum_{i=1}^{m}|\beta_i|$ for Lasso regression (Tibshirani, 1996), and $P_{\mathrm{Enet}}(\boldsymbol{\beta},\lambda)=\lambda\sum_{i=1}^{m}|\beta_i|+(1-\lambda)\sum_{i=1}^{m}\beta_i^2$ for Elastic Net regression (Zou et al., 2005). There is also a shrinkage method called HyperLasso that is designed specifically for simultaneous analyzing a set of SNPs and covariates (Hoggart et al., 2008). This method implements a Bayesian-inspired penalized maximum likelihood approach with a Normal-Exponential-Gamma (NEG) prior over each regression coefficient. The NEG distribution has a sharp peak at zero, which imposes a strong penalty on the coefficients when they are close to zero, thus leads to sparse models.

Shrinkage methods have been applied to various GWAS studies for detecting the marginal and interactive effect of SNPs. For example, Sun et al. (2009) used Ridge regression to detect SNPs associated with rheumatoid arthritis. By incorporating information on multiple correlated genetic variants, they identified a SNP near the *HLA-B* gene that was not significant in their single-SNP analysis. Wu et al. (2009) applied Lasso regression to GWAS data of celiac disease and identified both marginal and interactive factors associated with this disease. Waldmann et al. (2013) used both Lasso and Elastic Net to identify SNPs affecting milk fat content. Barrett et al. (2015) applied HyperLasso to identify functional variants of melanoma from genetic loci that were pinpointed by earlier GWAS analyses.

As for the performance of different shrinkage methods, each of them has its own strengths and limitations. Ridge regression has a better performance for predicting phenotype labels given new genotype data. However, it does not achieve SNP selection. Lasso regression allows for automatic SNP selection by shrinking some of the coefficients to zero, but it tends to have problems when the SNPs are highly correlated (Waldmann et al., 2013). Elastic Net regression incorporates a combined penalty of Lasso and Ridge regression, thus holds the features of both methods. Yet, there is no conclusive evidence as for which method outperform others overall. For instance, using a simulation study, Ogutu et al. (2012) found Lasso outperformed Ridge regression, whereas other studies found that Ridge regression and Elastic Net outperformed Lasso (Bøvelstad et al., 2007; Waldmann et al., 2013).

Machine learning approaches. Machine learning approaches employ models and algorithms that have the ability to learn the SNP-trait association pattern. Random Forests (RF) is a machine learning method that has been successfully applied to genetic studies for the purpose of prioritizing SNPs, predicting disease status, and identifying SNP-SNP interactions (Li et al., 2016; Schwarz et al., 2007; Sun et al., 2007; Szymczak et al., 2016). A RF is an ensemble of decision trees, where each tree is grown using a bootstrap sample of the whole dataset (Breiman, 2001). A node in a tree is chosen as the SNP that can best reduce the trait impurity within child nodes. The effect of each SNP on the trait can be quantified by a variable importance score (Zhang et al., 2009). One prominent feature of RF is that it can capture the nonlinear interactions between SNPs, making it a desirable technique for unveiling the complex genetic architecture underlying multifactorial diseases.

RF was shown to perform well in simulations and real applications but these studies included no more than hundreds of SNPs (Lunetta et al., 2004; Schwarz et al., 2007). Technique advances such as the implementation of the Random Jungle (RJ) tool have made it possible to construct large RFs for genome-wide data (Schwarz et al., 2010). However, direct application of RF to genome-wide data still poses a computational challenge, and only a few studies were reported in the literature (Goldstein et al., 2010; Schwarz et al., 2007; Zou et al., 2012). For these reasons, two-stage RFs, which select candidate SNPs at the first stage and apply RF to the selected SNPs at the second stage, were exploited in more detail. For example, in a study of WTCCC coronary artery disease, Roshan et al. (2011) first performed a single-SNP analysis to assess the significance of association for each SNP. Then only SNPs ranked at the top of the whole list were selected for downstream RF analysis. Chung et al. (2012) proposed a similar two-stage RF to prioritize candidate SNPs in each pathway. At the first stage, a RF was built using all SNPs in a pathway. Then SNPs with a variable importance score greater than a threshold were selected to rebuild the RF. These two-stage approaches were shown to avoid overfitting and can generate more accurate models with a lower prediction error.

Support Vector Machine (SVM) is a supervised learning method that can be used for both classification and regression. In its simplest form, a SVM seeks to identify the optimal hyperplane that can separate the samples into two classes and achieve the largest margin between the classes. SVM was shown to have excellent power in detecting epistasis in both simulated and real genetic data (Chen et al., 2008; Listgarten et al., 2004). In Listgarten et al. (2004), a number of genetic variants associated with breast cancer risk were discovered using a SVM model. These variants are collectively better at predicting breast cancer patients than single variants. Chen et al. (2008) explored several SVM-based strategies that were able to uncover the interaction among SNPs. Nonetheless, due to the same computational and overfitting issue as for RF, two-stage SVMs were favored as compared to applying SVM directly to genome-wide data (Kim et al., 2013; Roshan et al., 2011).

3.2 Gene-based multi-marker analysis

Gene-based analyses are those studying the genetic component of a disease at gene-level. They have emerged as a major complement to GWAS (Neale et al., 2004). Several reasons are behind. First, genes are the basic physical and functional unit of heredity. Cellular processes are ultimately directed by genes and driven by their products (proteins). Secondly,

aggregating SNP-level information into genes can reduce the multiple testing burden. A large-scale GWAS usually involves testing of more than one million SNPs while these SNPs can be mapped to around 20,000 genes. Additionally, gene-based analysis is an essential intermediate step for performing integrative analysis of GWAS results with biological knowledge at gene pathway and network level, as will be described in Section 3.3.

Various gene-based analysis methods have been proposed. These methods share an initial step of mapping SNPs to genes. Typically, SNPs located between the 5' UTR (five prime untranslated region) and 3' UTR (three prime untranslated region) of a gene can be mapped to that gene. More sophisticated strategies such as those taking into account the LD structure or regulatory effect were also investigated (Pers et al., 2015; Taşan et al., 2015), and will be discussed in more detail in the Discussion section of this thesis. Apart from the SNP to gene mapping issue, gene-based methods differ from each other for various features. In the following, we describe them in terms of whether they are based on analyzing individual genotype/phenotype data or GWAS summary data.

3.2.1 Methods based on analyzing individual genotype/phenotype data

Among the methods that analyze individual genotype/phenotype data, the SKAT method (Sequence Kernel Association Test) allows borrowing information between different SNPs to improve the power to detect the effect of a gene (Wu et al., 2011). SKAT is based on a logistic kernel-machine model and has the flexibility to include covariates in the analysis. It estimates a matrix of genetic similarity between pairs of individuals at the level of all SNPs of the gene using kernel functions. The significance of a gene is evaluated using a variance-component score test under a mixed model, whose test statistic follows a mixture of chi-square distributions. Several extensions of SKAT were also proposed, which make it feasible for conducting analysis of familial data (Chen et al., 2013; Oualkacha et al., 2013; Svishcheva et al., 2014), and analysis including both rare variants and common variants (Ionita-Laza et al., 2013).

Another useful gene-based analysis toolkit is MAGMA (Multi-marker Analysis of GenoMic Annotation) (Leeuw et al., 2015). MAGMA includes both the tool to analyze individual-level data and also the tool to analyze summary data. The tool that analyzes individual-level data characterizes the relationship between the phenotype and the SNPs of a gene via multiple

linear principal components regression. This model has a similar form as the OLS model as described previously. The major difference is that instead of modeling directly on the genotype data, it projects the SNP matrix of a gene onto its principal components (PC), then prunes away PCs with very small eigenvalues before using the remaining ones as predictors for the phenotype in the regression model. This improves power by removing redundant parameters and guarantees the model is identifiable in the presence of highly correlated SNPs. The association of a gene with the disease is assessed using an F-test.

Multifactor Dimensionality Reduction (MDR) is a powerful method in detecting gene-gene (also SNP-SNP) interaction using individual genotype/phenotype data (Hahn et al., 2003). The main idea of MDR is to reduce the dimensionality of multi-locus information by pooling multi-locus genotypes into high-risk and low-risk groups, thus reducing to a one-dimensional variable. Cross-validation (CV) and permutation test are used to select the interaction pattern that has the best ability to classify and predict disease status. Yet, it is can be difficult to perform high order gene-gene interaction analyses via MDR at genome-wide level because it requires exploring a huge search space and suffers from a computational burden due to high dimensionality (Oh et al., 2012). Many MDR extensions have been proposed, including Gene-based MDR that allows for perfroming fast and efficient high order gene-gene interaction analysis (Oh et al., 2012), and the model-based MDR (MB-MDR), which is a parametric extension of the MDR method that was shown to have increased power over MDR in identifying gene-gene interactions for most genetic models (Cattaert et al., 2011).

3.2.2 Methods based on analyzing GWAS summary data

Many gene-based methods are based on analyzing GWAS summary data. Such methods are becoming more and more prominent since summary data have become abundant after years of GWAS effort. These methods are typically conducted by combing the *p*-values of SNPs mapped to a gene into a gene *p*-value. One naive while popular approach is to take the most significant SNP *p*-value among all SNPs mapped to that gene. This strategy is easy to implement and is sensitive in capturing the best association signal. It has been utilized in various early studies (Askland et al., 2012; Wang et al., 2007). However, taking only the best SNP to represent the whole gene, other SNP signals present in the gene will be ignored. Thus the overall gene effect can be under-evaluated if a trait is highly polygenic. Another limitation of this approach is that long genes harboring many SNPs tend to have a lower *p*-value, even if

none of its SNPs is truly associated with the disease. This is the consequence of performing multiple tests simultaneously, for which testing κ hypotheses there is a chance of $1-(1-y)^{K}$ to get the smallest *p*-value lower than y (for any $0<y<1$). Thereby the resulted gene *p*-values are inflated by gene length.

Various strategies have been introduced to address these issues. They typically combine SNP *p*-values (or test statistics) into a representative statistic (denoted as T), then evaluate its significance of deviation from the background distribution under the null hypothesis of no gene-disease association. The schematic diagram of such approaches is depicted in Figure 1.8. Some examples of defining T are presented in Table 1.2.

Computing the probability density functions of these statistics requires the correlation information of the SNP *p*-values, which, however, is generally unknown. Two strategies are employed to account for the correlation between SNP *p*-values. One strategy approximates the correlation using SNP LD estimated from HapMap or 1000 Genome Project reference panels, or from a custom set of individual genotype data when they are available (Liu et al., 2010). The idea is intuitive—two SNPs tend to have high dependence in their *p*-values if they are in high LD, whereas they are likely to have independent *p*-values if they are not in LD. Given the estimated correlation structure among SNP *p*-values, the gene *p*-value based on T can be computed either by Monte Carlo approximation such as employed by the VEGAS method (Liu et al., 2010), or by analytical calculation as employed by the MAGMA (Leeuw et al., 2015) and PASCAL method (Lamparter et al., 2016).

Table 1.2: Commonly used gene representative statistics for gene-based analysis.

Gene representative statistics (P_i represent SNP *p*-values)	**Related method**
$T = -2\sum_{i=1}^{m} \ln P_i$	COMBASSOC (Curtis et al., 2008)
$T = -2\sum_{i=1}^{m} \ln(1-P_i)$	Pearson's method (Pearson, 1938)
$T = \sum_{i=1}^{m} X_i$; where $X_i = Q_{\chi_1^2}(P_i)$ is the upper quantile of the χ_1^2 distribution evaluated at P_i	VEGAS (Liu et al., 2010), VEGAS2 (Mishra et al., 2015), PASCAL (Lamparter et al., 2016), fastBAT (Bakshi et al., 2016), MAGMA (Leeuw et al., 2015)
$T = \max_{i \le m} X_i$, or equivalently, $T = \min_{i \le m} P_i$	VEGAS, VEGAS2, PASCAL, MAGMA
$T = \max_{i \le m} Z_i$; where $Z_i = Q_{N(0,1)}(P_i)$ is the upper quantile of the standard normal distribution evaluated at P_i	MAGENTA
$T = -2 \times Q_1\left(\ln P_1, \ln P_2, \cdots, \ln P_m\right)$; Q_1: the first quartile	TopQ (Lehne et al., 2011)
$T(k) = \prod_{i=1}^{k} P_{(i)}$; $1 \le k \le N$ is a truncation point chosen a priori by user	Rank Truncated Product (Dudbridge et al., 2003)
$T = \prod_{i=1}^{N} P_i^{I(P_i \le \tau)}$; τ is a truncating parameter, typically set as $\tau = 0.05$	Truncated Product (Zaykin et al., 2002)

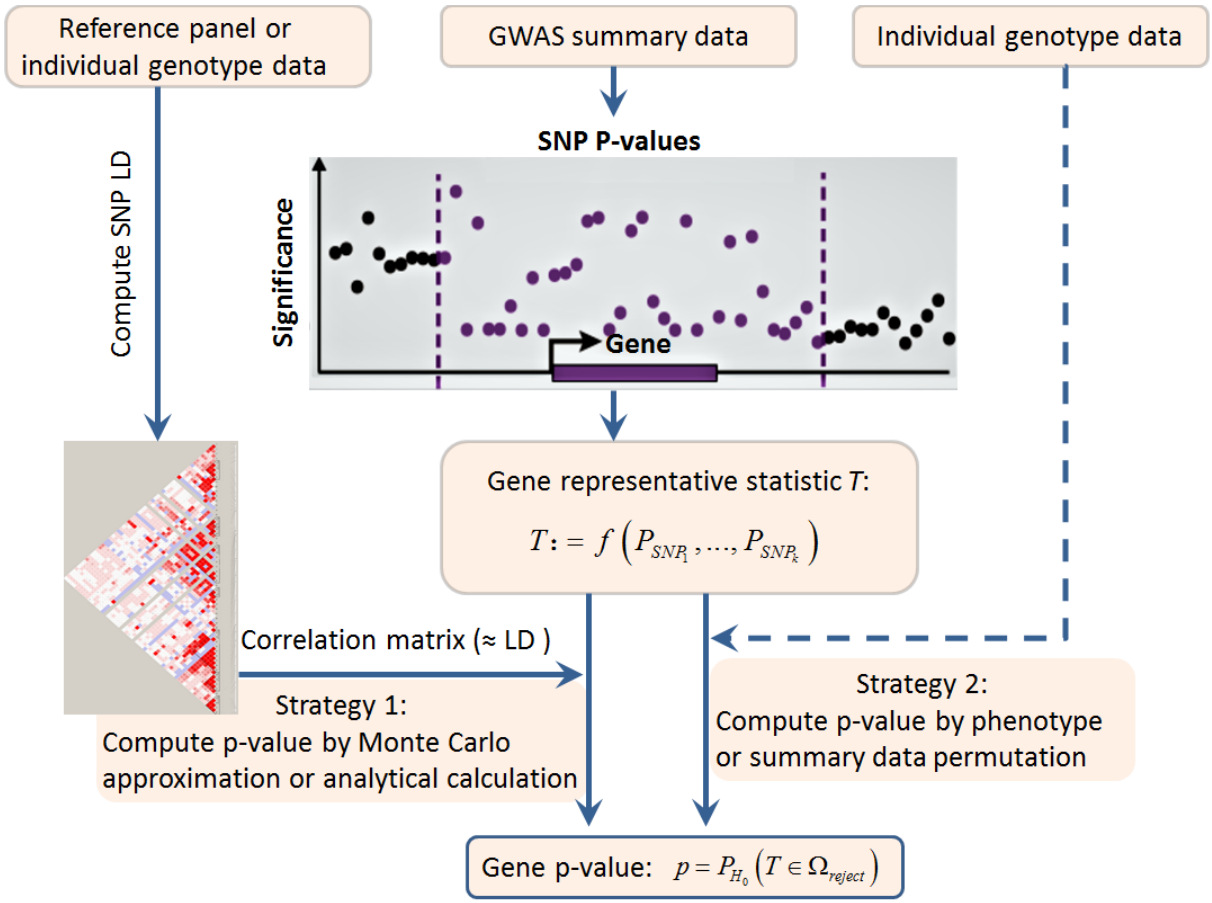


Figure 1.8: Diagram of computing gene *p*-values using GWAS summary data. A gene representative statistic T is first computed from the *p*-values of SNPs mapped to that gene. Then the gene *p*-value is computed by evaluating the significance of T for its deviation from the background distribution under the null hypothesis of no gene-disease association. To compute the significance of T, one strategy (Strategy 1) is based on Monte Carlo approximation or analytical calculation, where the SNP *p*-value correlation structure is approximated by the LD computed from a reference population or individual genotype data. Another strategy (Strategy 2) is based on phenotype or summary data permutation. For phenotype permutation, the individual genotype/phenotype data is required.

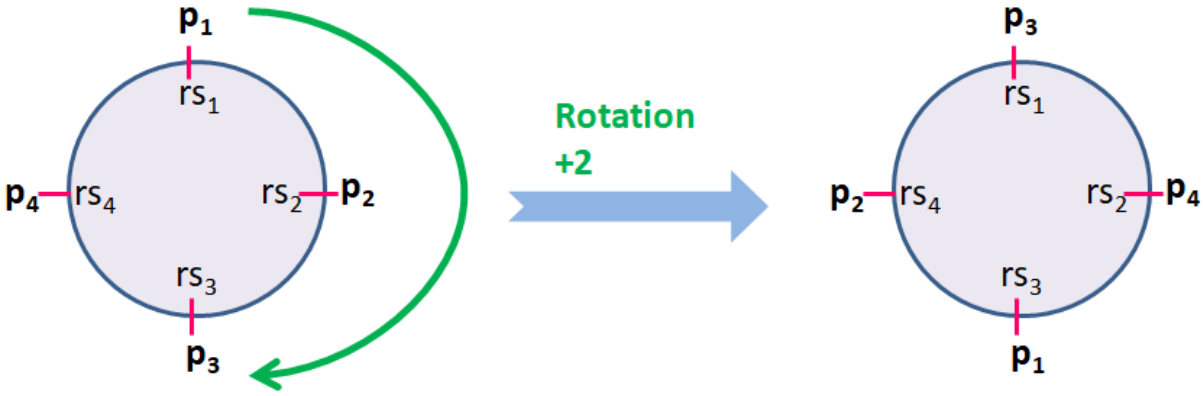


Figure 1.9: An illustrative example of the Circular Genomic Permutation strategy. CGP considers the genome as a circle. SNP *p*-values of a GWAS are ordered on the circle according to the position of the SNP. A CGP sample is generated by rotating the *p*-values for a randomly chosen position (=2 in this example) and reassigning these *p*-values to each SNP.

Another strategy to incorporate the correlation among SNP *p*-values is to use permutation techniques (Figure 1.8). Phenotype permutation is the gold standard of generating the null distribution, for which the LD structure and other possible confounding factors, such as gene size, are accounted for (Liu et al., 2010). Computing gene *p*-values via permutation is conceptually simple and is implemented as the "set-based test" in the PLINK software package (Purcell et al., 2007). Nevertheless, heavy computational demand has restricted its application at genome-wide scale. Moreover, there are several cases in which permutation-based method cannot be applied, including family-based GWAS, GWAS meta-analyses based on summary statistics, and studies in which the individual genotypic data are unavailable. Another permutation strategy, which applies directly to SNP summary data by randomly shuffling the SNP *p*-values, was proposed as an alternate to phenotype permutation. Though convenient and efficient, it is criticized for treating the SNP *p*-values as if they were independent (thus the correlation is not accounted for). Cabrera et al. (2012) proposed a permutation strategy called Circular Genomic Permutation (CGP) that is applied to summary data and can partly preserve the *p*-value correlation structure. As illustrated in Figure 1.9, CGP considers the genome as a circle, starting from chromosome 1 and ending at chromosome 22 then restarting from chromosome 1. SNP *p*-values of a GWAS are ordered on the circle according to the position of the SNP. A CGP sample is generated by rotating the *p*-values for a random position and reassigning them to each SNP at their new position. In this

manner, CGP keeps the relative position between SNP *p*-values unchanged during the permutation process, thus the correlation structure is partly preserved. This strategy has been applied to several studies and was shown to have similar performance compared to phenotype permutation when applied to a pathway-based analysis (Brossard, 2013; Chambers et al., 2013; Mott et al., 2014; Stainton et al., 2015). Thereby, we have taken advantage of the CGP strategy and developed an efficient method to compute gene *p*-values, as will be introduced in Chapter III.

3.3 Pathway-based and network-based multi-marker analysis

As mentioned previously, the single-SNP analysis approaches have methodological bottlenecks and have resulted in limited power. Gene-based analysis could partly overcome their limitations, but it is not flawless. Genes that are genuinely associated with disease status but do not reach the multiple testing significance threshold cannot be captured. The joint and interactive effects of multiple genes are also missed. This urges the investigators to develop alternate and complementary strategies. Integrative analysis approaches that combine knowledge of biological pathways and/or biological networks with GWAS results to identify functional gene modules associated with disease status have emerged as a prominent research direction (Figure 1.10). The rationale behind these methods is that biological organizations are fundamentally modular—instead of working in isolation, groups of genes, proteins or metabolites are known to work together through physical and/or functional interaction (Mitra et al., 2013).

Pathway and network-based approaches appear to be well suited for the analysis of massive GWAS data. They have a number of benefits relative to the analyses performed at individual SNP or gene level either from biological or statistical considerations. First, they aggregate molecular events across multiple genes in the same pathway or network subunit, thus reduce the number of hypotheses to be tested and can increase the likelihood that a test passes the statistical significance threshold. Secondly, a common disease is the result of the joint action of multiple genes within a pathway. Although each single gene may confer only a small disease risk, their collective action is likely to have a significant role in the development of a disease. Thirdly, locus heterogeneity, in which alleles at different loci cause disease in different populations, will increase the difficulty in replicating associations of a single-marker with a disease. Therefore, replication of association findings at the SNP or gene level can be

difficult if there are redundant genes with similar roles present (Sun, 2012). In comparison, pathway and network-based approaches that combine information from multiple loci in a functional unit could produce more stable and robust results than single-marker approaches do (Qiu et al., 2008). Furthermore, the ultimate goal of genetic studies of complex diseases is to decipher the link between genotype and phenotype. Despite the efforts made by extensive studies in search for genes causing complex diseases, the links between genetic variants and complex traits, which are essential for unraveling the pathogenesis of complex diseases, have remained elusive. In this sense, pathway and network-based approaches provide a complementary role to single-marker approaches for interpreting the molecular basis underlying human diseases.

In the following sections, I will first summarize the biological pathway and network resources that are hugely available for performing integrative analyses. Afterwards, I will introduce pathway-based analysis, and then network-based analysis that is the main focus of this thesis.

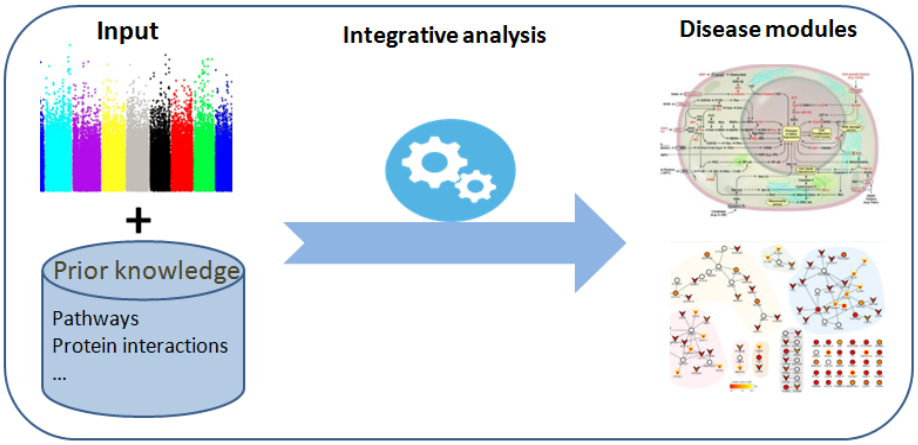


Figure 1.10: Diagram of integrative analysis of GWAS data.

3.3.1 Resources of biological pathways and networks

Thanks to the high-throughput "omics" (e.g., genomics, transcriptomics, proteomics, and metabolomics) technologies, our resources of biological data are increasing at exponential rate. According to a report published in 2013 in the journal *Nucleic Acid Research*, there are 1552 biological databases publically accessible online (at the time of writing that report) (Fernández-Suárez et al., 2013). These databases are developed for various purposes, curated at different knowledge levels and via diverse approaches. A comprehensive overview of these

biological databases was given in Zou et al. (2015). Among them, those describing biological pathways and networks are most widely used for integrative analysis of GWAS data and are the main focus of this thesis. They will be introduced below.

Biological pathways. Pathways are an important component in systems biology. A pathway is defined as a series of actions among molecules in a cell that leads to a certain product or a change in a cell. A pathway can trigger the assembly of new molecules, such as a protein or lipid, can turn genes on and off, or spur a cell to move. These actions are usually controlled and catalyzed by enzymes. Some of the most well-known pathways are involved in metabolism, the transmission of signals, and the regulation of gene expression. Perturbations in pathways are found to cause disorders.

Pathway resources are accumulating rapidly. Researchers have discovered many important pathways through laboratory studies of cultured cells, bacteria, fruit flies, mice and other organisms, many of which are similar to the counterparts in humans. For the purpose of effectively archiving and easily accessing to the ever-expanding knowledge of established pathways, an increasing number of databases have been established during the last decade. The Pathguide resource collects links to many databases of manually curated and computationally predicted pathways (http://www.pathguide.org). Some of the well-known collections are listed in Table 1.3. These include Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016), Reactome (Croft et al., 2013), and Gene Ontology (GO) (Ashburner et al., 2000). Both KEGG and Reactome contain manually curated (MC) pathways for different biological processes, whereas GO contains mostly electronic annotations (EA) for human genes and attempts to describe gene functions using three hierarchical categories: molecular functions, biological processes, and cellular components. Other commercial pathway providers, such as Pathway Studio (https://www.pathwaystudio.com/) and Ingenuity Pathway Analysis platform (https://www.qiagenbioinformatics.com/), also curate pathways from multiple sources of information, including literature reviews as well as experimental evidence. Still, many biological pathways remain to be discovered or explored in more detail. It will take years of effort to identify and understand the complex connections among all the molecules in all biological pathways, as well as to understand how these pathways work interactively.

Table 1.3: Some of the well-known pathway databases. MC: manually curated; EA: electronic annotations.

Pathway database	Curation method	Description	URL
KEGG	MC	KEGG pathway is a collection of manually drawn pathway maps representing the knowledge on the molecular interaction and reaction networks for metabolism, genetic information processing, cellular processes, etc.	http://www.genome.jp/kegg
Reactome	MC	Reactome is an open-source, open access, manually curated and peer-reviewed pathway database. Pathway annotations are authored by expert biologists, in collaboration with Reactome editorial staff and cross-referenced to many other bioinformatics databases	http://www.reactome.org
Gene Ontology	MC/EA	GO provides controlled vocabularies for the description of biological process, molecular function, and cellular component of gene products. The controlled vocabularies of terms are structured to allow annotation of gene products to GO terms at varying levels of detail	http://www.geneontology.org
WikiPathways	MC/EA	WikiPathways is an open space for biological pathway editing. Users can freely contribute and modify the content	http://wikipathways.org/index.php/WikiPathways
Ingenuity Pathway Analysis	MC/EA	IPA is a large curated database of biological pathways created from millions of individually modeled relationships between proteins, genes, complexes, cells, tissues, drugs, and diseases	http://www.ingenuity.com/products/ipa
Pathway Commons	EA	Pathway Commons aims to store and disseminate knowledge about biological pathways. Information is sourced from public pathway databases and is readily searched, visualized, and downloaded	http://www.pathwaycommons.org

Biological networks. Biological networks and pathways are similar concepts but with certain distinctions. Both comprise functionally related genes, proteins, and other molecular components that carry out biological processes. In comparison, a pathway describes the series of biochemical reactions and physical events among molecules (e.g., complex formation, phosphorylation events, and conformation changes), whereas a network characterizes the relationship among molecules and represents them by means of graph. Unlike pathways resources that are mostly acquired through laboratory studies or careful manual curation by domain experts, biological networks are constructed via a broader range of techniques and can capture various types of relationships among biological entities.

Many types of biological networks have been characterized to date. These include metabolic network, cell signaling network, gene regulation network, drug interaction network, protein-protein interaction network, and many others. Protein-protein interaction (PPI) contributes to the most of our current knowledge of biological networks and has been the major resource used for performing integrative analysis of GWAS data. In a narrow sense, PPI describes the highly specific physical contacts established between two or more protein molecules as a result of biochemical events. In a broad sense, the word "PPI" has been used for describing various types of relationship among proteins and their coding genes, including physical interaction, gene co-expression, and co-occurrence of them in literature. For this reason, a PPI is sometimes synonymous to a "functional protein network", "protein association network", or "gene network".

Interactions between proteins can be detected by many techniques. These techniques fall into three major categories according to where the analysis is performed: in vivo, in vitro, and in silico. For in vivo techniques, a given experiment is conducted in a whole living organism. In vivo PPI detection methods include yeast two-hybrid and synthetic lethality (Brückner et al., 2009; Nijman, 2011). In vitro studies are performed in a controlled environment outside their normal biological context. Several in vitro methods for detecting PPI are affinity chromatography, tandem affinity purification, protein arrays, protein fragment complementation, phage display, X-ray crystallography, co-immunoprecipitation, and NMR spectroscopy (Junker et al., 2011; Lehne et al., 2009; Rao et al., 2014). In silico methods refer to the analyses conducted via computational algorithms. Some in silico PPI detection/prediction methods are protein sequence-based approaches (Singh et al., 2010),

protein 3D structure-based approaches (Porollo et al., 2007), gene fusion, gene co-expression analyses and text-mining approaches (Papanikolaou et al., 2015). A limitation of in silico method is that the resulted interactions may not have experimental evidence, and can contain spurious information. Nonetheless, its ability to perform large-scale analysis will provide a more comprehensive and deeper coverage of the protein interaction map.

Numerous PPI databases have been established to collect published PPI data and to provide convenient access to the data. Table 1.4 lists some of the major PPI databases. They can be divided into three subgroups according to their source of origin: (1) primary databases, which include experimentally verified protein interactions collected from either small-scale or large-scale published studies that have been manually curated; (2) meta-databases, which contain only experimentally proven PPIs obtained by integrating multiple primary databases; (3) prediction databases, which include mainly PPIs predicted in silico. Many databases also provide friendly graphical user interfaces (GUI) for data accessing, where users input one or multiple identifiers such as protein names or accession number according to RefSeq, Universal Protein Resource (UniProt), Ensembl gene ID or Entrez gene ID. In return, users obtain interaction information about the input proteins (or their coding genes). This information usually contains the interactors of the proteins, the evidence/source of interactions, and the description of protein entities. Some databases also provide primary tools for customized network visualization and manipulation. It is of note that although all of them provide knowledge on protein interactions, each database has its own knowledge source, curation and storage protocols. It was observed that the overlap between these PPI databases is relatively small (Rao et al., 2014), thereby a combined investigation of multiple databases in a research work can be beneficial. In the following, we will introduce two PPI databases that are utilized in this thesis work.

Table 1.4: PPI databases. These databases are divided into three subgroups: (1) primary databases that include experimentally verified protein interactions collected from either small-scale or large-scale published studies that have been manually curated; (2) meta-databases that include experimentally proven PPIs obtained by integrating multiple primary databases; (3) prediction databases that include mainly predicted PPIs derived using different approaches, combined with experimentally proven PPIs. Data was accessed in March 2017.

Database name	**#proteins**	**#interactions**	**Species**	**URL**
Primary Databases				
HPRD	30,047	41,327	Human	http://www.hprd.org/
BioGRID	65,617	1,423,105	All	http://thebiogrid.org/
MINT	25,530	125464	All	http://mint.bio.uniroma2.it/
IntAct	98,289	720711	All	http://www.ebi.ac.uk/intact/
DIP	28877	81784	All	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
OPHID/I2D	unknown	1,279,157	Human	http://ophid.utoronto.ca/ophidv2.204/
Meta-Databases				
PINA	17,109	166,776	All	http://omics.bjcancer.org/pina/
APID	29,701	349,144	All	http://cicblade.dep.usal.es:8080/APID/init.action
InWeb_InBioMap	17,653	625,641	Human	https://www.intomics.com/inbio/map/#home
Prediction Databases				
STRING	9,643,763	1,380,838,440	All	https://string-db.org
PIPs	7750	79441	Human	http://www.compbio.dundee.ac.uk/www-pips/dbStats.jsp
UniHI	36023	573995	Human	http://www.unihi.org/

PINA. The Protein Interaction Network Analysis (PINA) platform is a comprehensive web resource for protein interaction network construction, filtering, visualization, and management (Cowley et al., 2011). PINA integrates PPI data from six public curated databases (IntAct, MINT, BioGRID, DIP, HPRD and MIPS/MPact), and builds a non-redundant protein interaction dataset for six model organisms *(Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster, Caenorhabditis elegans and Saccharomyces cerevisiae)*. PINA also provides a variety of built-in tools to filter and analyze the network for gaining insight into it, such as to retrieve protein interaction modules identified by clustering algorithms, and to identify topologically important proteins. PINA can be accessed via its website http://omics.bjcancer.org/pina/ or via the Cytoscape plugin PINA4MS http://apps.cytoscape.org/apps/pina4ms.

STRING. The Search Tool for the Retrieval of INteracting Genes/proteins (STRING) is currently the largest PPI database (Szklarczyk et al., 2017). Up to the time of March 2017, it contains 1,380,838,440 interactions among 9,643,763 proteins for a comprehensive coverage of diverse organisms. Each interaction represents a known or predicted relationship between genes or gene products (proteins). These include direct (physical) and indirect (functional) relationship derived from various sources, such as integration from primary PPI databases, systematic genome comparisons, high-throughput experiments, gene co-expression and text-mining analyses. Notably, all interactions in STRING are provided with a probabilistic confidence score, derived by separately benchmarking groups of interaction against the manually curated functional classification scheme of the KEGG database and generally correspond to the probability of finding the linked proteins within the same KEGG path (Kanehisa et al., 2009). A final "combined score" quantifying the overall interaction confidence between a pair of proteins is computed by combing all sub-scores via the formula $S_{Combined} = 1 - \prod_{sub} (1 - S_{sub})$. This combined score is often higher than the sub-scores, expressing increased confidence when the interaction is supported by multiple sources of evidence. Based on this score, the overall interaction between two proteins is classified as low confidence if $S_{Combined} < 0.4$, medium confidence if $0.4 < S_{Combined} < 0.7$, and high confidence if $S_{Combined} > 0.7$. The STRING database can be assessed conveniently through the website https://string-db.org, the stringApp Cytoscape

application http://apps.cytoscape.org/apps/stringapp, and the STRINGdb R/Bioconductor package.

3.3.2 Pathway-based analysis of GWAS data

Pathway-based analysis of GWAS data has the goal of identifying pathways with their genetic architecture significantly altered for a disease status. Over recent years, various pathway-based methods have been proposed (Chen et al., 2010; Guo et al., 2009; Wang et al., 2007; Zhang et al., 2010). Some of them overlap with the methods designed for gene-based analysis, such as those based on analyzing individual-level data or combining SNP *p*-values. This is because a pathway is similar to a gene in that it also consists of a fixed set of SNPs, thus a pathway can be viewed as a "giant gene".

The methods that are designed more specifically for pathway analysis include over-representation analysis (ORA) and functional class scoring (FCS) analysis (Khatri et al., 2012) (Figure 1.11). ORA, also known as functional enrichment analysis, has the goal to identify pathways over-represented by a list of susceptible genes selected on the basis of gene-level significance, for example, those having a significant *p*-value ($p \leq 0.05$) after multiple testing correction. Over-representation of a pathway is usually computed by hypergeometric test or binomial test. The related methodologies and analyzing tools will be presented in more detail in Section 3.3.3. ORA has been widely utilized for pathway-based analysis because it is easy to implement. Nonetheless, ORA has limitations. The definition of a list of susceptible genes is not straightforward. They are usually chosen based on a stringent significance threshold, which can be a salient issue when a GWAS is underpowered. Consequently, the majority of genes that do not reach the significance threshold are neglected in the analysis, including those bearing small to moderate marginal effects.

The FCS methods aim at pinpointing pathways enriched in overall high association signals. Unlike ORA that focuses on a set of selected susceptible genes and thus ignores the effect of remaining genes, a FCS takes into account the overall effect of all genes in the pathway (genes not involved in the GWAS are not included). In a FCS analysis, testing for the enrichment of signals of a pathway can be conducted either in a self-contained (association) or a competitive (enrichment) manner (Khatri et al., 2012). In a self-contained test, the null hypothesis is "a pathway is not associated with the disease under study". This test usually

includes two steps. At first, a pathway statistics (T) is formed either directly from the SNP statistics, or indirectly by first computing gene statistics (SNP statistics → (gene statistics) → pathway statistics). Then, the association significance of the pathway can be evaluated on the basis of T using the same methodologies as that of the gene-based analysis (i.e., Monte Carlo simulation, analytical computation, and permutation). Therefore the approaches introduced in Section 3.2.2 can be applied directly.

Alternately, in a competitive (enrichment) test, the null hypothesis is "the pathway genes are no more associated with the disease than genes outside the pathway". The first implementation of a competitive approach is the GSEA method introduced by Wang et al. (2007), based on an adaptation of an earlier method proposed by Subramanian et al. (2005) designed for the analysis of gene expression data. In their approach, genes are ranked in descending order according to their association with the trait (computed by gene-based methods). The statistics of a pathway, called enrichment score (ES) in their approach, is defined as a Kolmogorov-Smirnov running sum statistic. This statistic measures the difference in the rank of pathway genes relative to the rank of genes outside the pathway. A high ES value indicates a pathway includes genes of strong association evidence that are ranked at the top of the gene list. ES is tested for its derivation from the null distribution by permutation. Several variants of this approach, as well as other types of GSEA, have been proposed for the purpose of correcting biases and increasing power. These include GSA-SNP (Holden et al., 2008), SSEA (Weng et al., 2011), i-GSEA4GWAS (Zhang et al., 2010), and SeqGSEA (Wang et al., 2014). The relative strengths of these approaches have been evaluated in many genetic studies (Chen et al., 2010; Guo et al., 2009; Wang et al., 2007; Zhang et al., 2010).

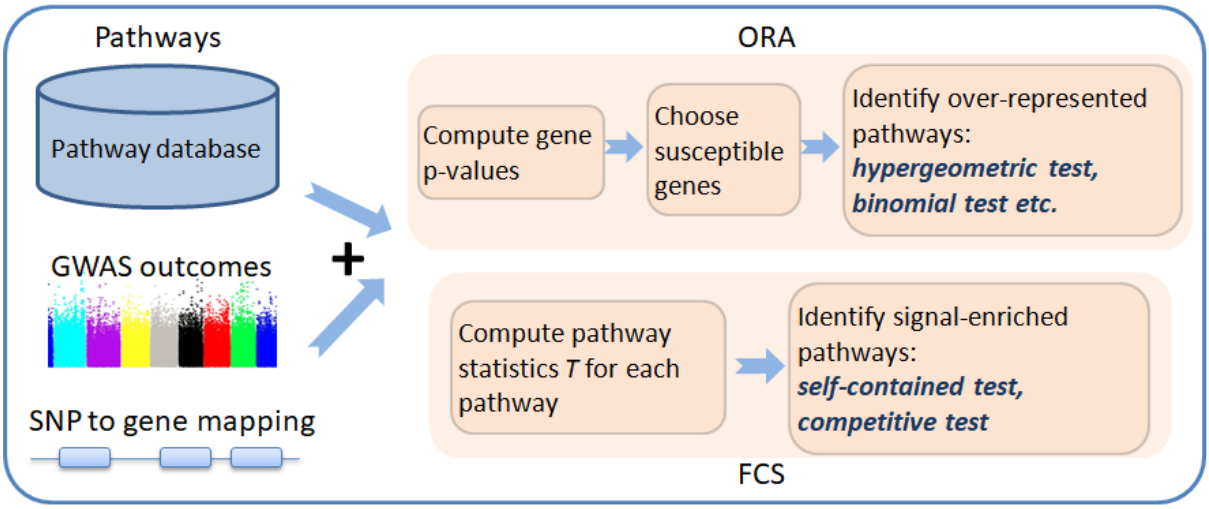


Figure 1.11: The principle of over-representation analysis (ORA) and functional class scoring (FCS) analysis. For ORA, susceptible genes are selected on the basis of gene-level significance computed using gene-based methods, then the pathways enriched in these genes are identified. For FCS, the statistic for each pathway is computed from GWAS results; then a pathway is tested for its significance of signal enrichment using either a self-contained test or a competitive test, or both.

3.3.3 Network-based analysis of GWAS data

Pathway-based analysis has been successfully applied to unveil the biological mechanism of many diseases. The identified pathways provide new insights that may be missed in a single-marker analysis. However, such approaches also have limitations: (1) although some prominent pathways are well studied, the knowledge on biological pathways remains fragmented and incomplete (Jin et al., 2014); (2) existing pathway annotations cover predefined pathways that may be too general in their delivery of disease-related biological functions (Ruano et al., 2010; Sun, 2012); (3) the connection information among genes is lacking within major annotation databases used for pathway analysis, such as the GO database; (4) most pathway-based method consider different pathways as independent sets and ignore their possible crosstalk. Specifically, the crosstalk between pathways refers to instances for which one or more components of one signal transduction pathway affects another (Figure 1.12). Two pathways are suspected to crosstalk with each other if there is a considerable interaction between their protein members (Li et al., 2008). Crosstalks were commonly observed between signaling pathways, for example, between Camp-dependent kinase and MAP kinase through a protein tyrosine phosphatase (Saxena et al., 1999).

Nonetheless, they are seldom considered by existing methods for pathway-based analysis of GWAS data.

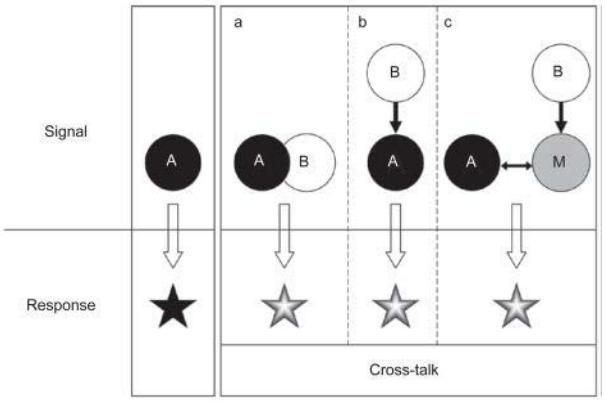


Figure 1.12: An example of pathway cross-talk. A cross-talk exists between two pathways A and B when both of the following criteria are met: functionally, the combinatorial signal from A and B must produce a different response than that triggered by A or B alone; mechanistically, A and B must be connected in at least one of the three depicted ways: (a) components of the two pathways physically interact; (b) components of one pathway are enzymatic or transcriptional targets of the other; and (c) one signal modulates or competes for a key modulator or mediator ("M") of the other. Figure adapted from Guo et al. (2009).

Compared to pathways that contain information of fixed sets of genes, biological networks include richer knowledge on the interaction or relatedness of genes and molecules. This knowledge allows generating novel gene sets that may represent novel biological functions. Thereby, integrating network information with GWAS data to perform network-based analysis may enable complementary discovery of disease-relevant markers and their interactive effects. In the following, we will go through several components that are involved in network-based analysis of GWAS data.

3.3.3.1 Network terminologies

A network is conventionally denoted as $G=(V,E)$, where V is a set of nodes (also called vertices), and $E=\{(p,q)\mid p,q\in V\}$ are edges connecting the nodes. In the case of protein-protein interaction network, the nodes represent proteins/genes and the edges represent their interactions. Each edge can have a weight (or score), typically ranging from 0 to 1, which

describes the strength/confidence of a connection. When a network has weights on its edges, it is called a weighted network.

A network can be either undirected or directed. In an undirected network, there is no direction on its edges, whereas in a directed network there is a direction on its edges, such that an edge $p \to q$ is different from an edge $q \to p$. In biological systems, many relationships are directed, for instance, gene A regulates gene B ($A \to B$), protein C phosphorylates protein D ($C \to D$), but the reverse relationship does not exist or represents a different process. Yet, due to the difficulties for precisely characterizing complex biological system, the directions are generally unknown.

Using mathematical representation, all the above network features can be fully described by a $N \times N$ matrix A, called adjacency matrix ($N = |V|$ is the number of nodes). For an unweighted network, A_{pq} takes a binary value that indicates whether there is an edge from p to q. A_{pq}=1 if there is an edge from p to q; A_{pq}=0 otherwise. For a weighted network, A_{pq} takes a real value representing the edge weight. For an undirected network, this matrix is symmetric (A_{pq}=A_{qp}).

Two nodes in the network are called neighbors if there is an edge connecting them. The degree of a node p is defined as the number of neighbors it has, and can be computed by $\deg(p) = \sum_q I\left(A_{pq} \neq 0\right)$, where $I(\cdot)$ is the indicator function. Nodes that have a degree greatly exceeds the average node degree of the network are called hub nodes. Hub nodes have an important role in maintaining the network structure. Particularly, the degree distributions of biological networks are known not to be at random. Several studies have revealed that the degree distributions of many biological networks have a scale-free property (Maslov et al., 2002; Nacher et al., 2009). In such networks, the probability that a randomly selected node has degree k approximately follows a power-law $P(k) \sim k^{-\gamma}$, where γ is a constant that typically ranges from 2 to 3. The biological networks possessing this scale-free property are more tolerant to random functional failures and errors, but are also more vulnerable to hub nodes perturbations (He et al., 2006).

The distance between two nodes p and q is the minimum number of edges that needs to be traversed to reach q from p. The path through the network which achieves this distance is called their shortest path. In a biological network, this distance can partially reveal the functional relevance between two molecules. For example, nodes with a distance of one (neighbors) usually have a direct functional relationship (regulation, phosphorylation, transport across membranes), while nodes with a larger distance can be involved at a different stage of the same biological process such as those involved in the same pathway.

An induced subnetwork of G, denoted by $G' = (V', E')$, is the network formed by a subset of the whole nodes ($V' \subseteq V$) and the edges among them. For example, the interaction network of proteins involved in the immune system is a subnetwork of the whole interaction network. Although the global network structure has been extensively explored to analyze the properties of biological networks, recently much attention has been paid to the subunits of the networks, called network modules, which represent a connected subnetwork that can carry out a specific biological function or event. Perturbations in modules are found to be the cause of many complex diseases. The searching of such functional modules associated with a disease is the major goal of this thesis, and will be discussed in more detail in the following sections.

3.3.3.2 Tools for network visualization and manipulation

Visualization concerns the representation of data in a pictorial or graphical format. A proper visualization can help answer existing questions and raise new hypotheses. Network visualization is particularly essential for understanding the global network conformation and highlighting important substructures. An increasing volume of research has benefited from network visualization and manipulation to gain insight into the complex systems under investigation.

Many network visualization and manipulation tools/packages have been developed in the last years and others are constantly created. Table 1.5 lists some of the well-known ones in the field of biology. Among many of these tools stands the Cytoscape open-source software for integration, visualization and computation modeling of molecular networks together with other systems-level data (Shannon et al., 2003). Cytoscape is versatile in that it can be conveniently extended through adding plugins (also called Apps) of various functionalities, thus enables scientists from multidisciplinary fields to contribute to the expansion of the

ecosystem. In the last five years, both the amount of available Cytoscape plugins and the number of installations have grown dramatically. Up to the time of March 2017, there are 310 plugins available with a total of 777 thousand downloads. These plugins cover a broad range of utilities, including online network data import, network visualization, manipulation, topological analysis, functional module detection, enrichment analysis, pathway annotation, and so on. A comprehensive travel guide to Cytoscape plugins has been previously provided by Saito et al. (2012), and can be found directly from the Cytoscape homepage (http://www.cytoscape.org/).

Table 1.5: Network visualization and analysis tools.

Tool	**Description**	**URL**
Gephi	An open-source and free software for graph visualization and manipulation	https://gephi.org/
Cytoscape	An open source software platform for visualizing, manipulating and annotating molecular interaction networks and biological pathways. Plugins of various functions can be installed through Cytoscape App Store	http://www.cytoscape.org/
Graphviz	An open source graph visualization software	http://www.graphviz.org/
Igraph	A programming package implemented in R, Python and C/C++ language for network visualization and analyzing	http://igraph.org/redirect.html
GraphWeb	A public web server for graph-based analysis of biological networks	http://biit.cs.ut.ee/graphweb/
NaviCom	A web application for visualization of multilevel omics data on top of biological network maps	https://navicom.curie.fr./bRidge.php

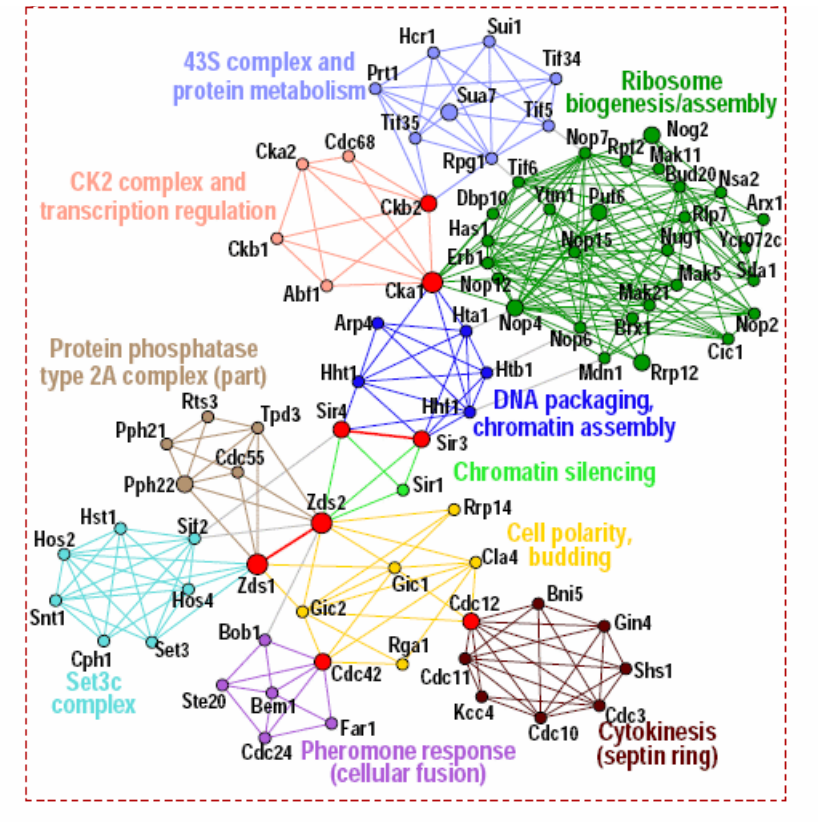


Figure 1.13: An illustrative example of the guilt-by-association principle. Proteins assigned to the same network community (represented in different colors) are found to be functionally closely related. Figure adapted from Palla et al. (2005).

3.3.3.3 Methods for conducting network-based analysis

Network-based analysis of GWAS data is based on the principle of "guilt-by-association", first raised by Oliver (2000). This principle states if two genes/proteins interact with one another in a network, they usually participate in the same, or related, cellular functions. It has been widely tested and supported in various studies (Li et al., 2016; Petsko, 2009). One example that well demonstrated this principle is given in Palla et al. (2005). The author identified protein communities (sets of strongly interconnected proteins) using a k-clique-percolation method. Their results indicated the proteins assigned to the same community have very similar function annotations (Figure 1.13). This "guilt-by-association" principle severs as the basis for many network-based applications with various objectives, including gene/protein function prediction (Piovesan et al., 2015; Tian et al., 2008), gene set over-representation analysis (Glaab et al., 2012), patients subgroup classification (Hofree et al., 2013), and here for network-based analysis of GWAS data.

Table 1.6: Tools for performing network-based analysis.

Tool	URL	Ref
jActiveModules	http://apps.cytoscape.org/apps/jactivemodules	Ideker et al. (2002)
dmGWAS	https://bioinfo.uth.edu/dmGWAS/	Jia et al. (2011)
Heinz	https://github.com/ls-cwi/heinz	Dittrich et al. (2008)
LEANR	https://cran.r-project.org/	Gwinner et al. (2016)
EW_dmGWAS	http://bioinfo.mc.vanderbilt.edu/dmGWAS	Wang et al. (2015)
STAMS	https://simtk.org/projects/stams	Hillenmeyer et al. (2016)
PINBPA	http://apps.cytoscape.org/apps/pinbpa	Wang et al. (2014)
DAPPLE	http://archive.broadinstitute.org/mpg/dapple/dapple.php	Rossin et al. (2011)
NETAM	http://www.sailing.cs.cmu.edu/	Lee et al. (2016)
ancGWAS	http://www.cbio.uct.ac.za/~emile/software.html	Chimusa et al. (2015)
PUPPI	https://sourceforge.net/projects/puppi/	Lin et al. (2016)
PANOGA	http://panoga.sabanciuniv.edu/	Bakir-Gungor et al. (2014)
COSINE	https://cran.r-project.org/	Ma et al. (2011)
cMonkey	http://djreiss.github.io/cMonkey/	Reiss et al. (2006)

Network-based analysis has been extremely popular in recent years. Many methods have been developed and many analysis tools have been implemented (Table 1.6). These methods can be broadly divided into two categories: the *active module search* category and the *seed gene oriented* category, as will be described below.

Category 1: active module search methods. Active module search methods are conducted by overlaying GWAS outcomes onto a PPI to identify disease-dependent "active modules"—subunits of the network showing significant enrichment of association signals. These subunits are also called as "functional module", "response modules", or "network hotspots" (Nibbe et al., 2010; Wang et al., 2015; Wu et al., 2009). To identify active modules, GWAS SNP *p*-values are first summarized into gene scores. This step involves the SNP to gene mapping and the combination of multiple SNP *p*-values into a single gene *p*-value. The related methodologies were described in Section 3.2.2 for gene-based analysis. The obtained gene *p*-values are transformed into gene scores, with a larger score indicating higher association significance. A common practice is to use the inverse normal transformation of *p*-value to z-

scores $z = \Phi^{-1}(1-p)$, such that z follows the standard normal distribution under the null hypothesis of no gene-disease association. This scoring function, however, usually results in about half of the genes have a positive score, which can lead to generating large modules even with random inputs (Rajagopalan et al., 2004). More sophisticated scoring methods that aim at addressing this issue have been developed. Rajagopalan and Agarwal (2004) corrected z by a linear factor such that only genes having a *p*-value less than a threshold (for example 0.05) will get a positive score. Dittrich et al. (2008) considered a signal-noise decomposition of the raw gene *p*-values, implemented as a mixture model that enables controlling the resultant subnetwork size by an adjustment parameter and achieving automated statistical significance over the resulted module.

The resulted gene scores are overlaid onto the PPI to build a disease-specific scored network. Given this network, the methods for searching "active modules" can vary substantially according to the different definitions of "active module" and the strategies to find them. Conventionally, the activeness of a module S is quantified by a module score defined as the summation of scores over the module genes $Z(S) := \sum_{g \in S} z_g$, or its normalized form that is adjusted for the background mean and variation (Dittrich et al., 2008; Ideker et al., 2002). Yet, the definition of a "module" can be more diversified. Various publications have defined a module as a connected subnetwork (i.e., its genes are connected to each other directly or indirectly within the subnetwork) (Dittrich et al., 2008; Ideker et al., 2002). Under such specification, the task of searching "active modules" is equivalent to solving the Maximum-Weight Connected Subnetwork (MWCS) problem as illustrated in Figure 1.14.

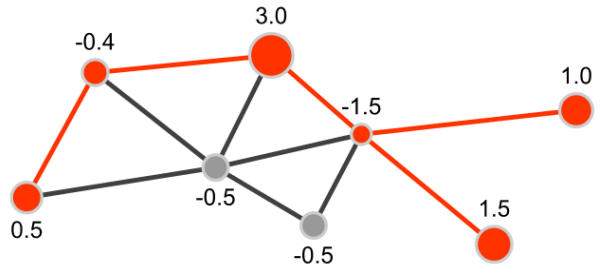


Figure 1.14: The Maximum-Weight Connected Subnetwork (MWCS). Given an undirected and connected network $G = (V, E)$, with each node associated with a weight z, its subnetwork $G_S = (V_S, E_S)$ is called a Maximum-Weight Connected Subnetwork if G_S is a connected network, and it has the maximum weight among all connected subnetworks, where the weight of a subnetwork is defined as $Z(S) := \sum_{v \in V_S} z_v$. In the example given in this figure, the subnetwork colored in red is a MWCS.

Finding the exact solution of the MWCS problem, however, is computationally challenging. The only known approach that allows for finding the exact MWCS was the HEINZ method proposed by Dittrich et al. (2008). In HEINZ, the MWCS problem was transformed into a prize-collecting Steiner tree problem (PCST) and solved by an integer-linear programming method. Several studies have considered the simpler variant of this problem as the constrained MWCS problem (Backes et al., 2012; Qiu et al., 2008). For example, Qiu et al. (2008) considered finding the MWCS that contains a root node and includes k nodes. Then active modules are found by starting from each node in the network as a root and setting k at all possible values to find the corresponding constrained MWCSs, though this approach can be computationally less efficient.

Table 1.7: The principles of three heuristic optimization strategies.

Simulated annealing. Simulated annealing is a probabilistic optimization technique that mimics the physical process of heating a material and then slowly decreasing its temperature. It was the first heuristic approach applied to the active module search problem (Ideker et al., 2002). To begin, a connected subgraph is chosen at random. At each iteration, nodes are added or removed from this subgraph. These changes are retained if they result in a connected subgraph with a higher score. The changes may also be kept with a probability that scales with the "annealing temperature" if they result in a subgraph with a lower score. After each iteration, the temperature decreases such that the accepted changes are increasingly likely to be beneficial. The final high-scoring subgraph is returned as the "active module".

Genetic algorithms. Genetic algorithms mimic natural selection among individuals of a population that drives biological evolution. The evolution usually starts from a population of randomly generated individuals. At each iteration, more fit individuals are stochastically selected to produce the next generation population. Over successive generations, the population evolves toward an optimal solution.

Greedy algorithms. Greedy algorithms are heuristic optimization algorithms that make the locally optimal choice at each stage. A greedy strategy does not, in general, produce an optimal solution, but may yield locally optimal solutions that approximate a global optimal solution in a reasonable time.

Beside the approaches that aim at finding exactly a MWCS, heuristic search strategies, such as those based on simulated annealing (Ideker et al., 2002), genetic algorithms (Klammer et al., 2010; Ma et al., 2011), or greedy algorithms (Jia et al., 2011), have been exhaustively explored to identify active modules. The principles of these heuristic strategies are summarized in Table 1.7. The rationality behind them is that although the MWCS is the optimal module from a mathematical point of view, other high-scoring subnetworks are also of biological interest regardless of whether their scores are strictly maximal.

The first heuristic method designed for identifying active modules is jActiveModules (Ideker et al., 2002). It searches modules in a way of simulated annealing. Several variants of jActiveModules were also developed and were applied to various genetic studies (Nacu et al., 2007; Rajagopalan & Agarwal, 2004; Ulitsky et al., 2009). Another widely applied heuristic method is the Dense Module Search algorithm implemented in the dmGWAS R package (Jia et al., 2011). The schematic diagram of DMS is shown in Figure 1.15. Briefly, DMS defines the score of a module of k genes as $Z_S = \sum z_i / \sqrt{k}$. It grows a module from a seed gene and iteratively adds the neighboring gene that can lead to the maximum increment of the module score. Module growth terminates if adding neighboring genes does not yield an increment of module score by at least $Z_S \times r$ ($r = 0.1$ by default). Each gene in the gene network is set as a seed once to generate a module. The modules with their scores ranked at the top x% of all modules (determined by the user) are selected as the final result.

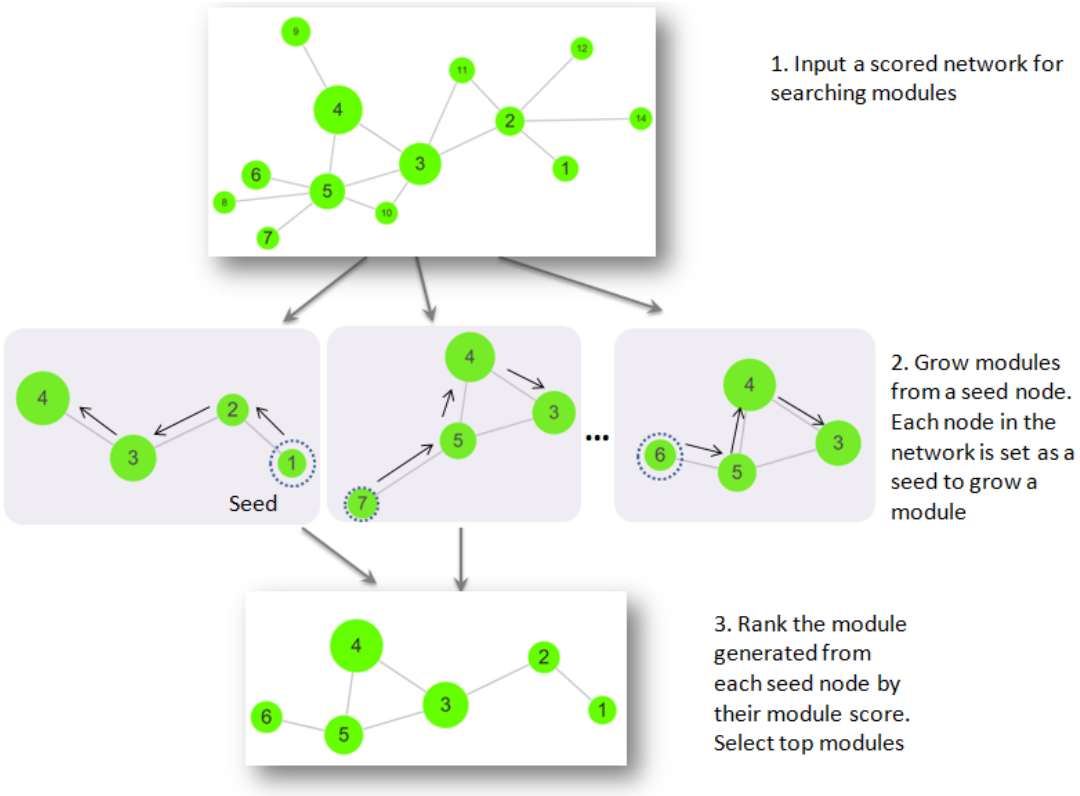


Figure 1.15: The DMS active module search strategy. DMS defines the score of a module of k genes as $Z_S = \sum z_i / \sqrt{k}$. It grows a module from a seed gene and adds iteratively the neighboring gene that can lead to the maximum increment of the module score. Module growth (along the arrows) terminates if adding neighboring genes does not yield an increment of module score by at least $Z_S \times 0.1$. Each gene in the gene network is set as a seed once to generate a module. Modules with their scores ranked at the top x% (determined by the user) are selected for downstream analysis.

It is worth noting that the module defined as a connected subnetwork does not utilize information on edge weights (when they exist), and has no emphasis on the module interconnection strength (denoted as ρ, usually defined as the ratio of the number of edges to the number of possible edges in the module). Several methods have considered searching for modules that are both enriched in high signals and have strong interconnection. For these methods, the module quality is quantified by $Q_1(S) = Z(S) + \lambda \times \rho(S)$, where λ is a tuning parameter which keeps a balance between the module score and module connectivity. Approaches for finding modules that maximize $Q_1(S)$ include the genetic algorithm implemented in the COSINE tool (Ma et al., 2011), the greedy algorithm implemented in EW_dmGWAS (Wang et al., 2015) and STAMS (Hillenmeyer et al., 2016), and a convex

optimization approach which approximates the discrete combinatorial problem by a continuous optimization problem (Wang et al., 2008).

There is also another method, named SConES (Azencott et al., 2013), designed for searching active modules. Unlike the approaches that find the MWCS or the approaches that maximize $Q_1(S)$, SConES identifies modules by maximizing a module quality defined as $Q_2(S) = Z(S) - \lambda \times \zeta(S) - \eta |S|$. In this function, the second term $-\lambda \times \zeta(S)$ is a function that has the effect to encourage connected nodes to be selected together. The third term $-\eta |S|$ is a sparsity regularizer that controls the number of nodes to be selected thus leads to sparse results. An exact algorithm based on graph-cut theory was provided to solve the maximization problem exactly.

Inspired by the formulation of SConES, we developed an active module search method, named SigMod, in order to identify modules that are both enriched in high signals and have strong interconnection. The detail of SigMod was described in Liu et at. (2017) and will be presented in Chapter IV of this thesis.

Category 2: seed gene oriented methods. Other than overlaying all GWAS association information onto a network to perform a global search for active modules, seed gene oriented methods focus on the topological property among or around a set of "seed genes". Seed genes are typically chosen as those having a strong disease-association evidence summarized from GWAS outcomes, for example, genes harbouring SNPs reaching genome-wide significance ($p \leq 5\times10^{-8}$) (Rossin et al., 2011), or genes having a significant *p*-value from gene-based method after multiple comparison correction, although a less stringent threshold can be used when highly significant signals are lacking. Disease genes reported from previous studies can also be added to the seed gene pool to increase the volume of prior knowledge.

Providing these genes that are of biological interest, the study objectives are mainly to screen genuine causal/functional genes within the seed gene list and to prioritize novel candidate genes beyond this list. The network approach for identifying disease susceptibility genes is motivated by the observation that genes contributing to the same trait often share functional relationships (King et al., 2003). Therefore, one may increase the power to detect disease genes by pinpointing genes located closest on the network or connected to other causal genes. One pioneering study implementing this idea was conducted by Taşan et al. (2015). In their

work, all genes that overlap with the genome-wide significant loci were collected as candidate disease genes (seed genes). A prix fixe-constrained optimization procedure was conducted to prioritize genes that form a functionally coherent network. This procedure finds the optimal combination of genes, such that only one gene is picked from each locus while those picked genes have the strongest interconnection. This approach was shown to have increased power than the approach that chooses the gene closest to the best SNP in terms of elucidating the mechanism of complex diseases.

For the purpose of prioritizing novel candidate genes, genes outside the candidate list are ranked according to their topological distances to genes inside the list. Early methods have used definitions of distance mainly as the shortest path (Schwikowski et al., 2000). However, many biological networks have a heavy-tailed scale-free degree distribution and the average shortest path length between nodes is small, making the shortest path a less desirable measure. More refined distance measures that take into account the overall network topology are proposed. Probably the most widely used technique for defining gene proximity is based on the property of network propagation, or network flow, as described in Qi et al. (2008) and illustrated in Figure 1.16. Briefly, fluid is pumped into each of the seed genes at a constant rate. It diffuses from the seed gene to other genes in the network via the edges connecting them. At the meantime of receiving fluid, each gene also pumps the received fluid to its neighboring genes. This process continues until the fluid system becomes stable. The resulting fluid represents the influence of seed genes over other genes. Genes retaining more fluid are likely to be functionally closer to seed genes and thus of higher possibility to be involved in disease susceptibility. Several approaches implementing this idea are GeneWanderer (Köhler et al., 2008), HotNet (Vandin et al., 2012), TieDIE (Paull et al., 2013) etc. A study conducted by Lee et al. (2011) has exploited six network propagation methods to assess their capability in boosting the statistical power to prioritize disease candidate genes. It is of note that network propagation approaches greatly depend on the quality of reference network. Both the accuracy and completeness of the network information play an essential role in the study performance. This is a key issue for network-based analysis as a whole, which will be discussed later in more detail in this thesis.

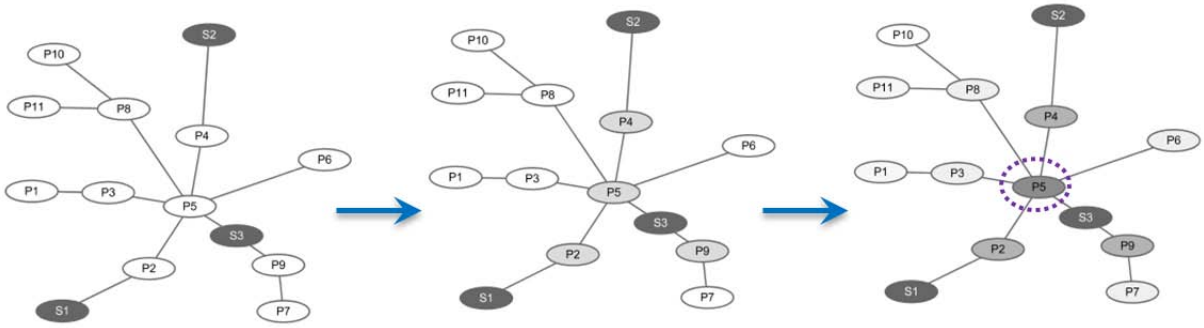


Figure 1.16: An illustrative example of network propagation. Nodes represent genes and lines represent their interactions. S1, S2, S3 represent seed genes. Fluid is pumped into each of the seed node at a constant rate. At every iteration of the network propagation process, nodes pump flow to their neighbors (thus also receive flow from their neighbors). Node greyscale represents the amount of flow they receive at each iteration. After multiple iterations, the amount of flow of each node converges. P5 stands out as the best candidate gene since it receives the largest amount of flow.

3.3.3.4 Methods for evaluating network-based analysis outcomes

It is well known that analyses performed at genome-wide scale tend to have high rates of false positives (Brzyski et al., 2017; Shen et al., 2013). To improve the reliability of findings, strict criteria have been established for reporting GWAS results, such as the use of a stringent threshold to declare significance, and to replicate findings in a replicating dataset, as introduced in Section 2.2. However, similar criteria are less well defined for network-based analysis. The performance of a network analysis is known to be influenced by noise from both the GWAS and the network data. Besides, many network algorithms have a heuristic nature and cannot find the optimal solution to their established problem. To reduce the amount of false positives caused by either the noise from input data or by the analyzing method, the outcomes need to be evaluated for their genuine relevance with the disease.

Nonetheless, currently there is no gold standard for network-based analysis result evaluation. Ideally, experimental verification of the role of identified genes on the disease can be conducted. Several studies have used experimental techniques to verify the co-expression, transcription regulation or other interaction relationship among genes in an identified module (Li et al., 2015). However, due to the complexity of disease mechanisms, the expense of experimental materials, and the demand of laboratory work, this approach is only applicable

for small modules that consist of a few genes, whereas it is nearly impossible to be conducted for assessing the role of large modules.

As an alternate option, in silico approaches are more feasible for assessing module-disease relevance. In silico approaches address whether the identified result is statistically relevant to the disease. Two types of test have been widely formulated and used. The competitive test evaluates whether the association signals enriched in a module are significantly higher than the signals outside the module. The self-contained test evaluates whether the module is significantly associated with the disease. These tests are conceptually the same as those performed for pathway analysis, which was introduced in Section 3.3.2. They are usually conducted in a combinatory manner to increase the overall confidence (Jia et al., 2012; Jia et al., 2011).

Using multiple datasets to perform discovery and replication analysis is another way to improve result reliability. In a study to identify gene modules associated with schizophrenia, the authors used two schizophrenia GWAS data to identify replicable results (Jia et al., 2012). They generated raw modules independently in each dataset and evaluated each module in the other dataset. Only those showed consistent association signals were selected to build the final module. Furthermore, the final module was assessed again using a third independent dataset to ensure its genuine-association with schizophrenia. This strategy provides robust results with multistage evaluation and replication, hence are more likely to identify genes underlying the disease susceptibility.

3.3.3.5 Methods for interpreting network-based analysis outcomes

As opposed to single-marker analysis that leads to the identification of genetic variants having highest significance (the amount of which is generally small), network-based analysis usually results in a collection of moderately significant genes that jointly influence the disease status. Additional to inspecting the function of each gene individually, it is a routine step to conduct annotation analysis to investigate the biological process in which they jointly involve. Such analysis helps inspect the genes in a view of systems biology and can shed light on the understanding of the biological mechanisms of the disease. It also justifies that these genes are indeed functionally related—as assumed by the "guilt-by-association" principle placed at the beginning of network-based analysis.

To annotate a gene list, one common approach is to perform over-representation analysis (ORA), as has been mentioned in Section 3.3.2 for pathway-based analysis. In general, ORA accepts a list of querying genes that will be compared with a given background set to test whether some functional categories, e.g., GO terms or KEGG pathways, are significantly enriched in the querying genes. Many tools for ORA have been developed. Table 1.8 lists some of them. These tools differ mainly by the annotation database they use. For example, BINGO only uses GO terms, while DAVID includes multiple pathway databases. Several tools, such as g:Profiler, GOrilla, and GSEAPreranked, also allow annotation of a ranked gene list, where the rank can be specified based on the disease-association score of each gene.

It has been argued that the agnostic gene overlap approach which assumes the equal weight of each gene in a functional process, or each gene has equivalent chance to be assigned to a biological term, is not optimal (Dong et al., 2016; Glaab et al., 2012). On the one hand, genes are not independent. They are rather linked with each other in the pathway map. On the other hand, genes in a pathway or a gene set are not of equal importance. For example, a gene may act as the regulator of a number of genes in the same pathway. The perturbation of this gene may have a larger impact on the pathway than the perturbation of its target genes. To address this issue, there is a trend to take into account the functional relationship among genes when performing ORA. For example, the IF method measures the contribution of a gene to a pathway based on the type of interaction (e.g., induction or repression) it has with upstream genes and its position in the pathway (Draghici et al., 2007). GANPA assigns a gene in a pathway with higher weight if it has more connections with other pathway members (Fang et al., 2012). A comprehensive review of these methods was given by Mitrea et al. (2013).

Table 1.8: Gene function annotation tools.

Annotation tool	**Annotation databases**	**URL**	**Ref**
PANTHER	GO, Reactome, PANTHER pathways	http://pantherdb.org/	Mi et al. (2013)
g:Profiler	GO, KEGG, Reactome	http://biit.cs.ut.ee/gprofiler/	Reimand et al. (2016)
clusterProfiler	GO, KEGG	https://bioconductor.org/	Yu et al. (2012)
EnrichNet	GO, KEGG, Reactome, Wiki Pathways, BioCarta and others	http://www.enrichnet.org/	Glaab et al. (2012)
DAVID	Over 60 functional categories	https://david.ncifcrf.gov	Huang et al. (2009)
PathwAX	KEGG	http://pathwax.sbc.su.se/	Ogris et al. (2016)
LEGO	GO	http://tianlab.cn/Research/softwares/	Dong et al. (2016)
BINGO	GO	http://apps.cytoscape.org/apps/bingo	Maere et al. (2005)
EnrichmentMap	GO	http://apps.cytoscape.org/apps/enrichmentmap	Merico et al. (2010)
ClueGO	GO, KEGG and BioCarta	http://apps.cytoscape.org/apps/cluego	Bindea et al. (2009)
GOlorize	GO	http://apps.cytoscape.org/apps/golorize	Garcia et al. (2006)
GSEAPreranked	MSigDB or customized gene sets	http://software.broadinstitute.org	Subramanian et al. (2005)
GOrilla	GO	http://cbl-gorilla.cs.technion.ac.il/	Eden et al. (2009)

Beside exploiting the module genes for the pathways they over-represent, clustering them into functionally similar groups is another useful annotation strategy that enables to summarize the major functions these genes have. The biological mechanisms are known to be extremely complex and reveal a "many-genes-to-many-terms" mapping profile. In one direction, individual genes can be associated with multiple biological terms, while in the opposite direction, individual biological terms can involve multiple genes. One powerful tool that allows to overcome this nested complexity and to present a summarized overview of the annotation structure is DAVID (Huang et al., 2009). DAVID considers two genes as functionally similar if their annotation profiles are similar. For example, if two genes encode similar sodium transporters, they are expected to have major functional annotations in common. Providing a list of query genes, DAVID constructs a gene-term annotation matrix using thousands of annotation terms from 14 categories integrated into the DAVID database (including GO terms, KEGG Pathways, Swiss-Prot Keywords, SMART Domains, UniProt Sequence Features etc.). Genes are clustered into the same functional group if their chance-corrected measure of co-occurrence is above a certain threshold. This analysis can be particularly useful if the gene list is large and contains genes sharing highly similar function, such as they represent gene families or protein complexes.

4 Asthma

The major objective of this thesis is to develop network-based analysis methods, and apply them to asthma GWAS data to identify biological processes and prioritize new candidate genes related to asthma. In the following, I will give an introduction to the definition, the epidemiology and pathogenesis of asthma. I will also outline the environmental and genetic components of asthma.

4.1 Definition of asthma

Asthma is a common chronic inflammatory disease of the airways of lungs. It is characterized by variable and recurring symptoms, reversible airflow obstruction, and bronchospasm. During asthma attacks, patients suffer from symptoms including episodes of wheezing, coughing, chest tightness, and shortness of breath, that may occur a few times a day mostly in the early morning or in the night. Depending on the person, the symptoms can become worse at night or with exercise. Recurrent asthma symptoms frequently cause sleeplessness, daytime tiredness, reduced activity levels and school and work absenteeism (WHO media center, 2017). Though there is no clear consensus on how to precisely define asthma, most cases are mild and can be diagnosed and treated by family doctors.

It is recognized that asthma is not a single disease but that the syndrome encompasses consistent groups of various characteristics (Wenzel, 2012), including age of asthma onset (childhood-onset, adult-onset asthma), the severity of disease (mild, moderate and severe asthma), occupational exposures and the varying response to treatment. Other subtypes may be defined by the character of the inflammatory infiltrate (eosinophilic or neutrophilic).

4.2 Epidemiology of asthma

Asthma is one of the most common chronic diseases in the world. It is currently estimated that 334 million people suffer from asthma worldwide and approximately 250,000 annual deaths are attributed to asthma (Martinez et al., 2013). Asthma prevalence varies from country to country, with the lowest value of less than 1% and highest value of up to 20% (Figure 1.17). Asthma has been found to have a higher prevalence in developed countries than in developing countries. Asthma is the most frequent chronic disease in children with the highest prevalence (>20%) observed in Anglo-Saxon countries (UK, New Zealand, Australia) and the

lowest prevalence in Greece (4-7%), India (6%) or China (4%). The prevalence of asthma in adults is lower, varying on average between 4% and 9% but with large variation according to geographical location. Asthma prevalence also varies according to ethnicity: in the US, asthma is more prevalent in Latino-Americans (14.2%) and in people of African-ancestry (9.5 %) than in people of European-ancestry (7.8%) (Moorman et al., 2011). Though asthmatic patients have benefited from the regular use of inhaled glucocorticoids and the number of deaths caused by asthma has been decreasing, the overall impact of asthma remains high and the prevalence of asthma has been increasing since the 1960s.

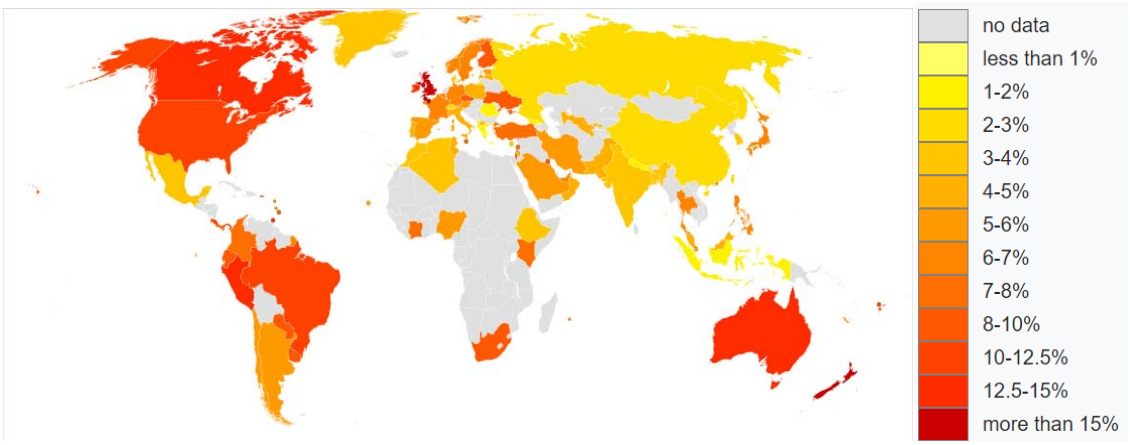


Figure 1.17: Asthma prevalence in different countries of the world as of 2004. From https://en.wikipedia.org/wiki/Asthma.

4.3 Pathogenesis of asthma

Asthma consists of a dynamic process involving immune mechanisms, chronic inflammation and airway epithelium remodeling which occur concomitantly or successively. Inflammation plays a central role in the pathophysiology of asthma. Airway inflammation involves an interaction of many cell types and mediators within the airways. It gives rise to the main features of the disease: bronchial inflammation and airflow limitation that cause recurrent episodes of cough, wheeze, and shortness of breath. The detailed processes by which these interactive events occur and lead to clinical asthma are still not fully known. However, despite the existence of distinct asthma subtypes (e.g., intermittent, mild persistent, moderate persistent, or severe persistent), airway inflammation remains a ubiquitous mechanism.

4.4 The environmental component of asthma

The increase of asthma prevalence in recent decades is probably due in large part to a change in lifestyle and exposure to environmental factors, including the many factors that have been associated with asthma occurrence and exacerbation, such as smoking (active and passive), air pollution, exposure to allergens, viral infections, unhealthy working conditions, diet. Epidemiological studies have also established that the timing of exposure to environmental factors in the life cycle is a crucial variable in determining risk (Figure 1.18), and the risks differ for childhood-onset and adult-onset asthma (Ober et al., 2011).

The prenatal environment is the first exposure of life and can establish lifelong risks for asthma. It has been reported that maternal asthma is among the most significant and consistent risk factors for childhood asthma (Abdulrazzaq et al., 1994; Holberg et al., 1998; Litonjua et al., 1998), thereby suggesting the prenatal environment differs between asthmatic and non-asthmatic mothers and contributes to subsequent risk of asthma in the fetus. Other studies have also shown that exposure to smoking during pregnancy is associated with a greater risk of asthma-like symptoms for the child (Gergen et al., 1998; Gilliland et al., 2001; Gold, 2000). These observations demonstrate the influential effect of prenatal environment on asthma.

Early life exposure to smoking or viral respiratory infections, such as respiratory syncytial virus and rhinovirus, can increase the risk of developing asthma in early childhood (National Asthma Education Prevention Program, 2007). Other factors can act as protective factors for developing asthma. For example, the exposure to farm animals, attending daycare, having a dog in the home, or drinking unprocessed cow's milk, during the first years of life, have been associated with protection against asthma in childhood, and this provides an explanation for the increasing prevalence of asthma risk in westernized countries (Ober & Vercelli, 2011). This is in line with the so-called “hygiene hypothesis” which proposes that exposure to infectious agents, symbiotic microorganisms (such as the gut flora or probiotics), and parasites, triggers protective responses during the development of the immune system (Okada et al., 2010).

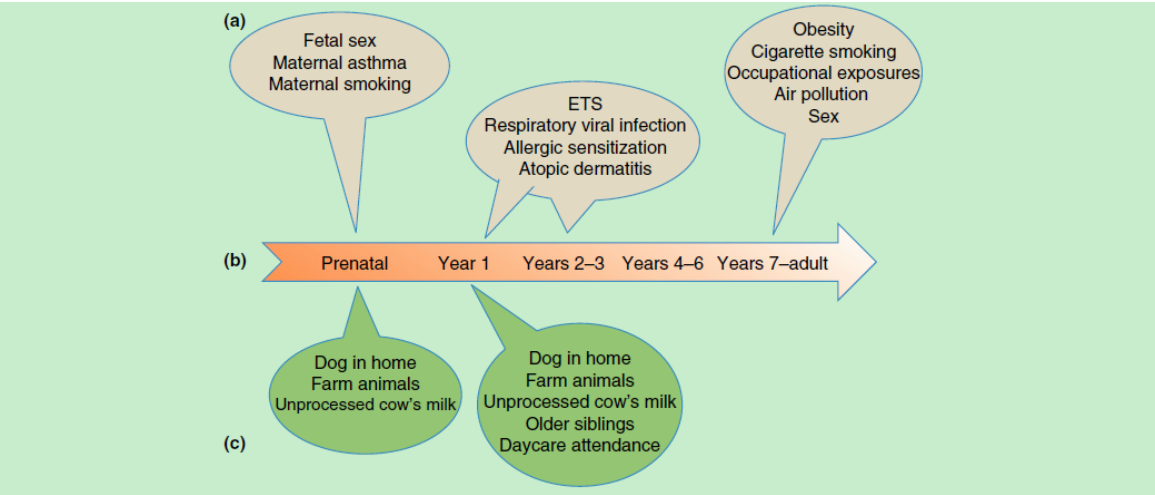


Figure 1.18: Risk and protective factors that influence asthma risk throughout the lifecycle. (a): Environmental exposures that have been associated with increased risk for asthma; (b): Stages of the lifecycle that are "sensitive" to the epidemiologic risk and protective factors; (c): Environmental exposures associated with protection from asthma. From Ober and Vercelli (2011).

Risk factors for asthma in adults include obesity or a high body mass index (BMI) (Chen et al., 2002; Gilliland et al., 2003), occupational exposures (e.g. to latex gloves, substances like ammonia, chemicals such as adhesives) (Bardana, 2008) and air pollution (Schwartz, 2004). Indoor allergens such as dust mites, animal dander, and mold are also important factors for triggering asthma both in children and adults.

4.5 The genetic component of asthma

Early studies have indicated that genetics plays an important role in the development of asthma and allergy (Willemsen et al., 2008). Studies of familial aggregation of asthma showed that the relative risk of developing asthma for siblings of asthmatic subjects with respect to subjects from the general population varies between 2.5 and 3.0. Twin studies consistently reported higher concordance rates for asthma in monozygotic twins (between 0.43 and 0.75) than in dizygotic twins (between 0.21 and 0.45). Analyses of familial transmission of asthma suggested a polygenic pattern of inheritance and did not reveal the effect of a major gene in a consistent manner (Los et al., 1999).

Given the substantial role of genetic factors in asthma and asthma-related phenotypes (including skin tests to allergens, IgE levels, lung function phenotypes), many genetic studies

have been conducted to identify these factors, including genetic linkage studies, candidate gene studies and, more recently, genome-wide association studies. Genome-wide linkage screens, mostly based on the affected sib-pair method and related methods, have revealed more than 70 regions linked to asthma or asthma-related phenotypes, although some of these regions were not consistently replicated. These regions included candidate genes but also novel genes identified by positional cloning (i.e., association analysis within linkage regions), but the functional role of these genes is still not well known (Bouzigon et al., 2010; March et al., 2013). To date, more than 1,000 association studies with candidate genes have been published with more than 200 loci reported associated with asthma and related phenotypes. However, only about 30 genes were found consistently associated with asthma in at least five independent studies (Table 1.9). These genes can be classified into four main categories: (1) genes involved in innate immunity and immune-regulation (e.g., *CD14*, *TLR2*, *TLR4*); (2) genes involved in Th2 immune response (e.g., *IL4, IL13*, *IL4RA*, *FCER1B*); (3) genes involved in airway epithelium biology and mucosal immunity (*CCL5*, *CCL11*, *SPINK5*); genes involved in lung function and airway epithelium remodelling (*ADRB2*, *TNF*, *NOS1*, *ADAM33*) (Halapi et al., 2004; Kabesch, 2005; Levy et al., 2005)

The availability of high-density genotyping arrays has led to the genome-wide association study era of asthma. GWAS are advantageous over candidate gene studies for their hypothesis-free nature and ability to systematically screen the genome, allowing for discovery of novel asthma-associated loci. Up to now, 23 loci have been associated with asthma *per se* by 15 GWAS (Table 1.10). Ten of these GWAS were conducted in populations of European ancestry (at least at the discovery stage), two in Japanese, one in Latino and one in ethnically-diverse populations (including populations of European-ancestry, African-Ancestry, and Latino). Of these 15 GWAS, seven studies included only childhood asthma and two studies included only adult asthma. These studies led to the identification of 23 loci, among which five (2q12, 5q22.1, 6p21.32, 9p24.1, 17q12-q21) were reported by at least two independent studies at the genome-wide significance level. The most strongly associated locus is the 17q12-q21 locus (*ORMDL3*, *GSDMB* and *GSDMA* genes) that was reported by the first asthma GWAS (Moffatt et al., 2007) and was confirmed by several GWAS in populations of various ethnic origin. This locus was reported to confer stronger risk in childhood-onset asthma than in adult-onset asthma and to interact with environmental exposure to smoking in early life (Bouzigon et al., 2008). Beside the 17q12-q21 locus, the 2q12 (*IL1RL1 / IL18R1*)

and 9p24 (*IL33*) loci were reported by two large meta-analyses of asthma GWAS, one conducted in the European GABRIEL consortium (Moffatt et al., 2010) and the other one in the American EVE consortium that included multi-ancestry populations (Torgerson et al., 2011). The broad HLA region at the 6p21.32 locus was reported by GWAS performed in European-ancestry populations and in Japanese, and included association signals for both childhood asthma and adult asthma. The 5q22.1 locus (*TSLP*) was identified in ethnically-diverse populations and in Japanese. Altogether, these asthma GWAS uncovered a smaller number of loci than similarly sized studies of other multifactorial diseases, which may be partly due to the significant role of environmental exposures on disease risk and the phenotypic heterogeneity that is a hallmark of this disease. Because the genetic loci identified to date explain only part of the genetic risk, complementary approaches to GWAS, such as the network-based analysis proposed in this work, may prove useful in revealing novel genes underlying asthma and bringing further insight into the genetic component of this complex disease.

Table 1.9: Genetic loci associated with asthma and asthma-related phenotypes (bronchial hyper-responsiveness, IgE levels) through candidate-gene studies and replicated in at least five independent studies. Adapted from March et al. (2013).

Gene	**Chromosomal locus**	**Function**
GSTM1	1p13.3	Detoxification, removal of products of oxidative stress
FLG	1q21.3	Epithelial integrity and barrier function
IL10	1q31-q32	Cytokine — immune regulation
CTLA4	2q33	Control/inhibition of T cell responses/immune regulation
IL13	5q31	Induces TH2 effector functions
IL4	5q31.1	TH2 differentiation
CD14	5q31.1	Microbe detection — recognizes pathogen associated molecular patterns
ADRB2	5q31-q32	Smooth muscle relaxation
SPINK5	5q32	Epithelial serine protease inhibitor
HAVCR1	5q33.2	T cell responses — hepatitis A virus receptor
LTC4S	5q35	Leukotriene synthase — inflammatory mediator
LTA	6p21.3	Inflammatory mediator
TNF	6p21.3	Inflammatory mediator
HLA-DRB1	6p21	Major histocompatibility complex class II — antigen presentation
GPRA	7p14.3	Regulation of metalloprotease expression, neuronal effects
NAT2	8p22	Detoxification

Gene	Chromosomal locus	Function
GSTP1	11q13	Detoxification, removal of products of oxidative stress
FCER1B	11q13	Receptor for IgE — atopy
IL18	11q22.2-q22.3	Inflammation
CC16	11q12.3-q13.1	Potential immunoregulatory function — epithelial expression
STAT6	12q13	IL-4 and IL-13 signaling
NOS1	12q24.2-q24.31	Nitric oxide synthase — cellular communication
CMA1	14q11.2	Chymase — mast cell expressed serine protease
IL4R	16p12.1-p12.2	Alpha chain of receptors for IL-4 and IL-13
CCL11	17q21.1-q21.2	Eoxtaxin-1 — eosinophil chemoattractant
CCL5	17q11.2-q12	RANTES — chemoattractant for T cells, eosinophils, basophils
ACE	17q23.3	Regulation of inflammation
TBXA2R	19p13.3	Platelet aggregation
TGFB1	19q13.1	Influences cell growth, differentiation, proliferation, apoptosis
ADAM33	20p13	Cell—cell and cell—matrix interactions
GSTT1	22q11.23	Detoxification, removal of products of oxidative stress

Genes are ordered according to chromosomal location. Abbreviations: IgE, immunoglobulin E; IL, interleukin; RANTES, regulated and normal T cell expressed and secreted; T_H, Thelper.

Table 1.10: Genetic loci associated with asthma *per se* by genome-wide association studies.

Discovery population	**Replication population**	**Phenotype**	**Region**	**Significance of the best signal**	**Reported genes**	**Ref**
European ancestry	European ancestry	Asthma	1q21.3	2.3×10^{-8}	*IL6R*	Ferreira et al. (2011)
African ancestry	African ancestry	Asthma	1q23.1	4.0×10^{-9}	*PYHIN1*	Torgerson et al. (2011)
European ancestry	European & African ancestries	Childhood asthma (3-12 years of age)	1q31.3	1.6×10^{-13}	*DENND1B*	Sleiman et al. (2010)
Multi-ancestry	Multi-ancestry	Asthma	2q12	2.0×10^{-15}	*IL1RL1*	Torgerson et al. (2011)
European ancestry	European ancestry	Adult asthma		1.1×10^{-9}	*IL1RL1, IL18R1*	Ramasamy et al. (2012)
European ancestry	European ancestry	Asthma		3.4×10^{-9}	*IL18R1*	Moffatt et al. (2010)
European ancestry	Multi-ancestry	Childhood asthma (5-12 years of age)	4q12	2.0×10^{-8}	*SRIP1 MIR548AG1*	Ding et al. (2013)
Japanese	Japanese	Adult asthma	4q31	1.9×10^{-12}	*USP38, GAB1*	Hirota et al. (2011)
European ancestry	Multi-ancestry	Childhood asthma (4-12 years of age)	5q12.1	3.0×10^{-8}	*PDE4D*	Himes et al. (2009)
Multi-ancestry	Multi-ancestry	Asthma	5q22.1	1.0×10^{-14}	*TSLP*	Torgerson et al. (2011)

Discovery population	Replication population	Phenotype	Region	Significance of the best signal	Reported genes	Ref
Japanese	Japanese	Adult asthma		1.2×10^{-16}	*TSLP, WDR36*	Hirota et al. (2011)
European ancestry	European ancestry	Asthma	5q31	1.4×10^{-8}	*IL13*	Moffatt et al. (2010)
Latino Americans	Latino Americans	Childhood asthma	6p21.33	$p < 5\times10^{-6}$ (admixture mapping; *p*-value based on permutation)	*MUC22*	Galanter et al. (2014)
Japanese	Japanese	Adult asthma	6p21.32	4.1×10^{-23}	*NOTCH4*	Hirota et al. (2011)
Japanese	Japanese	Childhood asthma		2.3×10^{-10}	*HLA-DPA1, HLA-DPB1*	Noguchi et al. (2011)
European ancestry	Mixed	Adult asthma		2.0×10^{-8}	*HLA-DQA1*	Lasky-Su et al. (2012)
European ancestry	European ancestry	Adult asthma		1.1×10^{-8}	*BTNL2,HLA-DRA*	Ramasamy et al. (2012)
European ancestry	European ancestry	Asthma		7.0×10^{-14}	*HLA-DQ*	Moffatt et al. (2010)

Discovery population	Replication population	Phenotype	Region	Significance of the best signal	Reported genes	Ref
European ancestry	European ancestry	Severe childhood asthma (2-6 years of age)	7q22.3	3.0×10^{-14} (but significant heterogeneity; $p_{random}=2.7\times10^{-7}$)	*CDHR3*	Bønnelykke et al. (2014)
Japanese	Japanese, Koreans	Childhood asthma	8q24.11	5.0×10^{-13}	*SLC30A8*	Noguchi et al. (2011)
European ancestry	Multi-ancestry	Childhood asthma (1-18 years of age)	9p23	8.0×10^{-9}	*JKAMPP1 TYRP1*	Ding et al. (2013)
Multi-ancestry	Multi-ancestry	Asthma	9p24.1	2.0×10^{-12}	*IL33*	Torgerson et al. (2011)
European ancestry	European ancestry	Asthma		9.2×10^{-10}	*IL33*	Moffatt et al. (2010)
Japanese	Japanese	Adult asthma	10p14	1.8×10^{-15}	*LOC338591*	Hirota et al. (2011)
European ancestry	Multi-ancestry	Childhood asthma (1-18 years of age)	10q24.2	5.0×10^{-8}	*HPSE2*	Ding et al. (2013)
European ancestry	European ancestry	Asthma	11q13.5	2.0×10^{-8}	*LRRC32*	Ferreira et al. (2011)
Japanese	Japanese	Adult asthma	12q13.2	2.3×10^{-13}	*IKZF4*	Hirota et al. (2011)
European ancestry	European ancestry	Asthma	15q22.2	2.4×10^{-9}	*RORA*	Moffatt et al. (2010) & Ramasamy et al. (2012)

Discovery population	Replication population	Phenotype	Region	Significance of the best signal	Reported genes	Ref
European ancestry	European ancestry	Asthma	15q22.33	3.9×10^{-9}	*SMAD3*	Moffatt et al. (2010)
European ancestry	European ancestry	Severe childhood asthma (2-6 years of age)	17q12-q21	6.4×10^{-23}	*GSDMB*	Bønnelykke et al. (2014)
European ancestry	European ancestry	Asthma		6.4×10^{-23} (childhood-onset asthma)	*GSDMB*	Moffatt et al. (2007)
Multi-ancestry	Multi-ancestry	Asthma		2.0×10^{-16}	*GSDMB*	Torgerson et al. (2011)
European ancestry	European ancestry	Severe asthma		1.0×10^{-8}	*ORMDL3*	Wan et al. (2012)
Latino Americans	Latino Americans	Childhood asthma		5.7×10^{-13}	*IKZF3*	Galanter et al. (2014)
European ancestry	European ancestry	Severe childhood asthma (2-6 years of age)		3.0×10^{-21}	*GSDMA*	Bønnelykke et al. (2014)
European ancestry	European ancestry	Asthma		3.0×10^{-17} (childhood-onset asthma)	*GSDMA*	Moffatt et al. (2010)
European ancestry	European ancestry	Asthma	22q12.3	1.0×10^{-8}	*IL2RB*	Moffatt et al. (2010)

5 Outline of the thesis work

Genetic association studies of asthma have been successful in identifying novel asthma-associated loci. As for many other complex diseases, the genetic variants at these loci account for a relatively small part of the whole asthma genetic susceptibility. Up to now, genome-wide association studies are mainly based on single-marker analysis which requires stringent threshold (5×10^{-8}) to declare significance. This may miss genetic variants having a small marginal effect and/or interacting with other variants. To complement the single-marker approaches, more sophisticated strategies, such as those integrating pathways and/or protein-protein interaction (PPI) networks with GWAS data to identify disease-associated functional gene modules, have become prominent. The main objectives of this thesis, thereby, were to develop network-based analysis methods and apply them to asthma GWAS data to identify biological processes and prioritize new candidate genes for asthma.

The asthma genetic data used in this thesis was acquired from the European GABRIEL Consortium that is hosted in UMR-946 laboratory (http://genestat.inserm.fr/fr/), which will be introduced in Chapter II.

The first study of this thesis, presented in Chapter III, aimed at extending an existing network-based analysis method and applying it to childhood-onset asthma GWAS data. This study included the development of a novel gene-based method, named fastCGP, to compute gene-level *p*-values from SNP *p*-values, and a bi-directional module searching strategy that extended the dmGWAS (Jia et al., 2011) approach to identify gene modules consistently associated with asthma. The results of this network analysis and their biological interpretation are presented in an article published in *Scientific Reports*.

The second study of this thesis, presented in Chapter IV, was to develop a novel active module search method, named SigMod, which has the potential to boost the performance of network-based analysis in general. SigMod aims at selecting a set of genes that are enriched in high association signals and tend to have strong interconnections. Compared to previous network analysis methods, SigMod has several advantages, including the high interpretability of the result, the robustness to background noise, and the ability to incorporate edge weights. SigMod was applied to both simulated data and childhood-onset asthma GWAS data using a biological network from the STRING database (Szklarczyk et al., 2017). The methodological

developments of SigMod together with the results of simulations and analysis of asthma GWAS data are presented in an article published in *Bioinformatics*.

The discussion, perspectives, and conclusion of this thesis are presented in Chapter V.

CHAPTER II. GABRIEL ASTHMA DATA

1 Description of the GABRIEL asthma genetic consortium

The asthma GWAS data used in this thesis comes from the GABRIEL consortium—a multidisciplinary study to identify the genetic and environmental factors of asthma in the European Community, that was funded by the European Community and the Wellcome Trust. This consortium consisted of over 150 scientists from 14 European countries.

The GABRIEL GWAS data includes a total of 10,365 case subjects and 16,110 controls recruited from 23 studies, all of which are European-ancestry. Details of these studies were given in Moffatt et al. (2010). Data on case subjects and population-matched controls were obtained from clinics and from cohort and cross-sectional population surveys in Europe. The study also included a few family studies, case subjects and controls of European descent from Canadian, Australian, and U.S. surveys. Asthma was considered to be present if it had been diagnosed by a physician. Childhood-onset asthma was defined as the presence of the disease in a person younger than 16 years of age and later-onset asthma as the disease that developed at 16 years of age or older. Some surveys contributed samples to both childhood-onset and later-onset groups. Other subgroups consisted of subjects with asthma that were developed at an unknown age, subjects with occupational asthma, and subjects with severe asthma. All participants or their parents provided written informed consent for their participation in the study, in accordance with the rules of local ethics committees.

2 Genotyping, quality control (QC) and genotype imputation of GABRIEL studies

All GABRIEL consortium datasets, except for the MRCA and MAGICS datasets, were genotyped at Centre National de Génotypage (CNG, Evry, France) using the Illumina Human610-Quad array. The MRCA and MAGICS datasets were genotyped using Illumina Sentrix Human-1 and Sentrix HumanHap300 BeadChips, as part of the first asthma GWAS (Moffatt et al., 2007). QC of individuals and SNPs genotyped at CNG was done in each dataset following the same protocol. Briefly, individuals were removed from analysis if they were not of European descent (based on principal component analysis of each GABRIEL dataset with all HapMap populations), had a low genotyping call rate (<95%) or were discrepant or ambiguous for genetic sex. Single Nucleotide Polymorphisms (SNPs) with call rate lower than 95% or minor allele frequency (MAF) lower than 0.01, or with Hardy-Weinberg equilibrium *p*-value < 10^{-4} were removed. QC for MRCA and MAGICS is detailed in Moffatt et al. (2007).

In each dataset, genome-wide imputations were performed using MACH 1.0 software and HapMap Phase 2 (Release 21) as reference panel. SNPs with imputation quality score (rsq) ≥ 0.5 and MAF ≥ 0.01 were kept for analysis.

3 Statistical analysis of childhood-onset asthma GWAS

In each study, association analysis between asthma and individual SNPs was performed using a logistic regression model that included allele dosage for each SNP and principal components to account for population structure. In family data, a robust sandwich estimation of the variance and a family cluster were used to take into account familial dependencies. All analyses were performed using Stata® version 10 (distributed by Stata Corporation, College Station, Texas, USA). The GWAS summary statistics of all GABRIEL studies are hosted by the Inserm unit UMR-946 (http://genestat.inserm.fr/fr/), where I am conducting my work.

For this thesis, we focused on childhood-onset asthma (COA), because it represents a more homogeneous entity. We randomly divided the 18 GABRIEL studies with childhood-onset asthma into two groups of nine studies each while keeping a total sample size similar in each group. The studies contained in each group are shown in Table 2.1. A total of 3,031 cases/2,893 controls were in the first group and 2,679 cases/3,364 controls were in the second group. The splitting of asthma COA GWAS into two groups was motivated by the need to use two sets of GWAS outcomes as input for the network analyses performed in this thesis. This allowed finding consistent results and cross-validating the results, as will be detailed in the following sections.

Meta-analyses of the study-specific asthma GWAS summary statistics were performed in each of the two groups and also in the whole set of 18 studies. We used both inverse variance fixed-effects model and random-effects model. Under the random-effect model, the estimate of the between-study variance was based on the DerSimonian and Laird method (Higgins et al., 2002). All meta-analyses were done with Stata® version 14.1 (STATA Corp., College Station, Texas, USA). Tests of significance of the meta-analyzed SNP effect sizes were performed using the Wald test. A conventional threshold of 5×10^{-8} was used to declare genome-wide significance. Tests of heterogeneity across studies in each group were based on the Cochran's Q statistic. To minimize the false-positive findings, we only examined SNPs for which at least two-thirds of the studies contributed to a meta-analysis, as done before in Moffatt et al. (2010). The outcomes of the meta-analyses (single-SNP *p*-values) of the two groups were named META1 and META2 respectively.

Table 2.1: The list of 18 GABRIEL COA GWAS surveys.

Study	Country	#Cases	#Controls	Total
META 1: 9 studies				
ALSPAC	United Kingdom	607	609	1 216
BAMSE	Sweden	239	246	485
CAPPS*	Canada	266	156	422
ECRHS	Pan European	279	620	899
MAGICS	Germany	630	572	1 202
MAS (pooled with MAGICS)	Germany	171	0	171
SLSJ*	Canada	373	390	763
TOMSK*	Russia	197	91	288
UFA	Russia	269	209	478
Total (META1)		**3,031**	**2,893**	**5,924**
META 2: 9 studies				
B58C	United Kingdom	213	200	413
BUSSELTON	Australia	188	390	578
EGEA*	France	482	598	1,080
GABRIELA	Germany	841	851	1,692
KSMU	Russia	112	116	228
MRCA-UKC*	United Kingdom	177	399	576
PIAMA	Netherlands	172	187	359
SAGE*	Canada	257	267	524
SAPALDIA	Switzerland	237	356	593
Total (META2)		**2,679**	**3,364**	**6,043**
Total (META1 + META2)		**5,710**	**6,257**	**11,967**

* Family study

4 Results

There was a total of 2,370,689 SNPs that passed all QC filters in the meta-analysis of META1 and META2. Estimates of the genomic control parameter (λ) showed little inflation of the test statistics under the fixed-effects model (lambda = 1.039 in META1 and 1.017 in META2); the lambda estimates were less than one (lambda = 0.873 in META1 and 0.870 in META2) under the random-effects model. Because there was evidence for heterogeneity across studies at a few loci, we used the *p*-values obtained under a random-effects model (p_{random}) as input for our network analyses, although this may be somehow a conservative choice. The Q-Q plots of META1 and META2 under a random-effects model are shown in Figure 2.1.

The SNP p_{random}-values of each meta-analysis is shown by a double Manhattan plot (called Miami plot) in Figure 2.2.

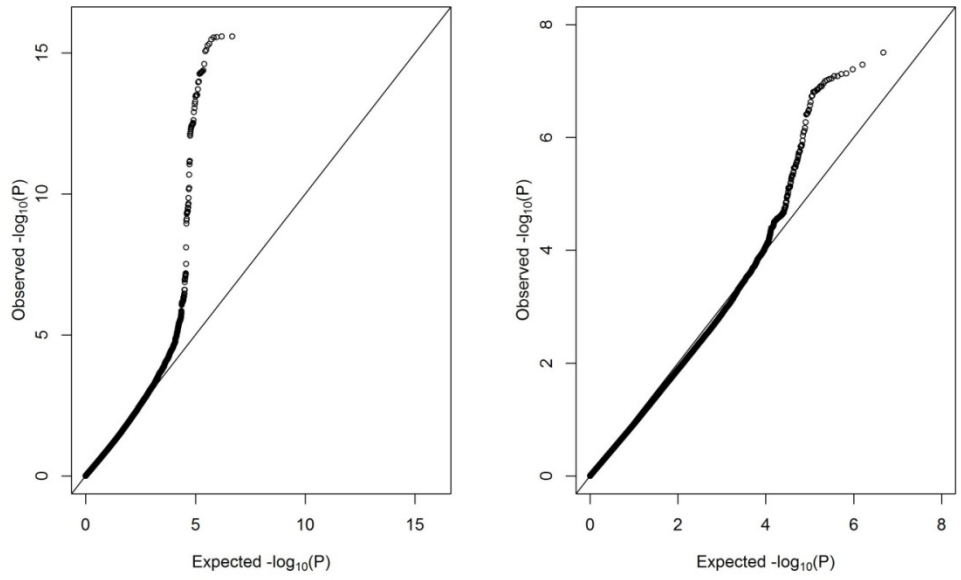


Figure 2.1: Log quantile-quantile (Q-Q) *p*-value plot. Left: Q-Q plot of META1; right: Q-Q plot of META2. The observed GWAS p_{random}-values for SNPs are plotted against the expected *p*-values. The genomic control parameter λ is 0.873 for META1 and 0.870 for META2.

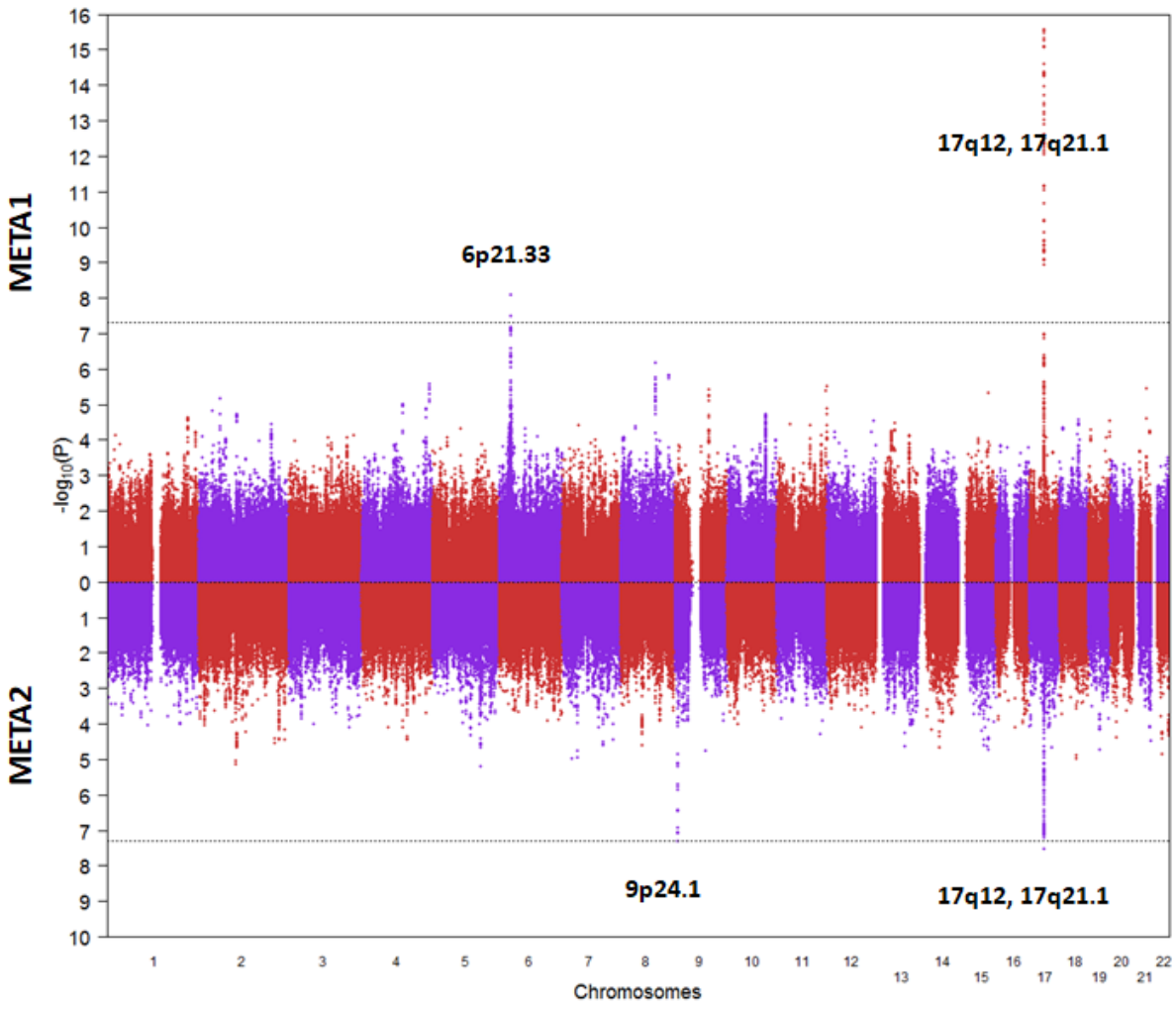


Figure 2.2: Double Manhattan plot of the results of two meta-analyses. The horizontal dashed lines indicate the genome-wide significance threshold ($p < 5\times10^{-8}$).

The distribution of p_{random} values for different thresholds is shown in Table 2.2. When using the genome-wide significance level of 5×10^{-8}, we observed 64 SNPs associated with COA in META1. All these SNPs are located at 2 loci, the 6p21.33 and 17q12-q21 loci (Figure 2.2 and Table 2.3). The strongest association on chromosome 17 was with rs9303281 within the *GSDMB* gene ($p_{random}=2.6\times10^{-16}$; odds ratio (OR)=1.36, 95% CI= (1.27-1.47)). The strongest association on chromosome 6 was with the SNP rs2596560 ($p_{random}=7.9\times10^{-9}$; OR=1.30, 95% CI=(1.19-1.43)), which is approximately 5kb downstream of *HLA-B* gene. In META2, only one SNP on 17q12-q21 reached the genome-wide threshold, rs4794820 ($p_{random}=3.1\times10^{-8}$, OR=0.79, 95% CI=(0.73-0.86)). This SNP is located 15kb apart from the top SNP in META1 and is distal to *ORMDL3* gene.

When we used a less stringent threshold of 5×10^{-7} that is suggestive of association, no additional locus was identified in META1 (five additional SNPs meeting that level were in the 6p21.33 region). However, in META2, one additional locus, 9p24.1, was detected; the top SNP (rs7848215) was close to the genome-wide level; $p_{random}=5.1\times10^{-8}$) and is proximal to *IL33* gene.

These results were compared to those of the meta-analysis of all 18 studies. In the latter analysis, two loci were genome-wide significant, 2q12 and 17q12-q21, and two other loci, 6p21.33 and 9p24.1, showed suggestive association ($p_{random} \leq 5\times10^{-7}$). We can note that the 2q12 locus was not detected in META1 or META2, the SNPs with the strongest association having $p_{random} = 1.9\times10^{-5}$ and 2.2×10^{-5} in META1 and META2 respectively. Finally, among the four loci evidenced in these analyses, three of them, 2q12, 9p24.1 and 17q12-q21 were previously reported by the GABRIEL meta-analysis of asthma GWAS based on genotyped SNPs (Moffatt et al., 2010), while the 6p21.33 locus is a new locus in European-ancestry populations, as it was only reported in an admixture mapping analysis in Latino populations.

Table 2.2: This table lists the number of SNPs that have a P_{random} value falling into each interval in META1, META2, and in the meta-analysis of the 18 GABRIEL studies.

Dataset / **Interval of P_{random}**	**META1**	**META2**	**Meta-analysis of all 18 studies**
$(-\infty, 5\times10^{-8}]$	64	1	138
$(5\times10^{-8}, 1\times10^{-7}]$	5	9	3
$(1\times10^{-7}, 1\times10^{-6}]$	32	24	30
$(1\times10^{-6}, 1\times10^{-5}]$	67	40	89
$(1\times10^{-5}, 1\times10^{-4}]$	394	184	266

Table 2.3. Loci significantly associated with asthma in META1, META2 and in the meta-analysis of all 18 GABRIEL studies

Loci with genome-wide significance - $p_{random} \leq 5\times10^{-8}$									
	META1 (9 studies)			**META2 (studies)**			**Meta-analysis of all 18 studies**		
Region	**SNP (position)†**	**Nearest genes‡**	**p_{random}**	**SNP (position)†**	**Nearest genes‡**	**p_{random}**	**SNP (position) †**	**Nearest genes‡**	**P_{random}**
2q12	rs3771166 (102,986,222)	*IL18R1, IL1RL1, IL18RAP*	1.9×10^{-5}	rs10206753 (102,968,362)	*IL1RL1, IL1RL2, IL18R1*	2.2×10^{-5}	**rs3771166 (102,986,222)**	***IL18R1, IL1RL1, IL18RAP***	**1.8×10^{-9}**
6p21.33	**rs2596560 (**31,355,318)	***HLA-B, MICA***	**7.9×10^{-9}**	rs2596560 (31,355,318)	HLA-B, MICA	0.34	rs2596464 (31,412,961)	*MICA, HCP5*	2.5×10^{-7}
17q12-q21	**rs9303281 (38,074,046)**	***GSDMB, ZPBP2, ORMDL3***	**2.6×10^{-16}**	**rs4794820 (38,089,344)**	***LRRC3, ORMDL3, GSDMA***	**3.1×10^{-8}**	**rs9303277 (37,946,469)**	***IKZF3, GRB7, ZPBP2***	**1.7×10^{-20}**
Additional loci - $p_{random} \leq 5\times10^{-7}$									
9p24.1	rs7848215	*RANBP6, IL33*	1.6×10^{-2}	rs7848215	*RANBP6, IL33*	5.1×10^{-8}	rs1342326 (6,190,076)	*RANBP6, IL33*	5.1×10^{-7}

†The SNP position is according to build 37;

‡The gene where eventually the SNP lies is first indicated followed by the previous gene and next gene;

SNPs that reached the genome-wide significance level ($p_{random} \leq 5\times10^{-8}$) are in bold;

Results are indicated for the top SNP at a locus in each meta-analysis if p_{random} was $\leq10^{-4}$ otherwise we used the SNP showing the strongest association across the three meta-analyses.

CHAPTER III. NETWORK-BASED ANALYSIS OF CHILDHOOD ASTHMA GWAS DATA

1 Summary

Asthma is a multifactorial disease arising from many genetic and environmental factors. Genome-wide association studies (GWAS) of asthma have been successful in identifying novel asthma-associated loci, but as for other complex diseases, genetic variants at these loci account for only a part of asthma genetic susceptibility. One limitation of GWAS is that they rest on single-marker analysis and is underpowered to detect genetic variants with small marginal effects and interacting with other genetic variants.

To complement the typical single-marker analysis in GWAS, more sophisticated analysis strategies, such as network-based analysis that integrates biological networks with GWAS outcomes, have been proposed to allow detecting sets of functionally related genes that jointly affect disease risk. In this study, we performed a network-based analysis of asthma. We used two GWAS outcomes (named META1 and META2, respectively) that are the results of meta-analyses of nine childhood-onset asthma (COA) GWAS each (5,924 subjects for the first dataset, 6,043 subjects for the second dataset). These GWAS were part of the European GABRIEL asthma consortium data as described in Chapter II. The PPI data we used was retrieved from PINA database (http://omics.bjcancer.org/pina/), which contains protein interaction information integrated from multiple primary databases and has been introduced in Section 3.3.1 of Chapter I.

To conduct network-based analysis, we first proposed an exact and efficient gene-based method, named fastCGP, to compute gene *p*-values from SNP *p*-values. fastCGP takes advantage of the CGP technique (as introduced in Section 3.2.2 of Chapter I) to correct the best-SNP *p*-value for gene length while takes into account the LD among SNPs. We have implemented fastCGP in an analytical manner so that to achieve computational efficiency. The gene *p*-values computed by fastCGP were transferred into scores and overlaid to the PPI to build a scored-network. We proposed a bi-directional module searching method that extended the dmGWAS (Jia et al., 2011) approach in order to identify consistent gene modules from the scored-network.

Application of these strategies to the asthma GWAS outcomes (META1 and META2) detected a gene module of 91 genes significantly associated with COA ($p \leq 10^{-5}$). This module consists of a core network and five peripheral subnetworks including known and high-confidence candidates for asthma. Out of the 91 genes that belonged to the selected module, 19 genes had nominally significant *p*-value in both META1 and META2 datasets. They included 13 genes at 4 loci found significantly associated with asthma in previous GWAS (2q12, 5q31, 9p24.1, 17q12-q21), and six genes at six distinct loci that are novel: *CRMP1* (4p16.1), *ZNF192* (6p22.1), *RAET1E* (6q24.3), *CTSL1* (9p21.33), *C12orf43* (12q24.31) and *JAK3* (19p13-p12). Additionally, we found the core genes of the module were connected to *APP* (encoding amyloid beta precursor protein), a major player in Alzheimer's disease that is known to have immune and inflammatory components. This link between *APP* and asthma-associated genes indicates that asthma and Alzheimer's disease may share common underlying mechanisms. It can open new routes for elucidating the functional role and relationships of these genes in asthma, and also potentially, in Alzheimer's disease. Functional analysis of the module genes using the DAVID tool (Huang et al., 2007) revealed four functionally related gene clusters involved in innate and adaptive immunity, chemotaxis, cell-adhesion and transcription regulation, which are biologically meaningful processes underlying asthma risk. Altogether, this study prioritized new candidate genes and brought deeper insight into their functional relationships with asthma.

2 Article published in *Scientific Reports* (doi:10.1038/s41598-017-01058-y)

SCIENTIFIC REPORTS

Received: 26 August 2016
Accepted: 21 March 2017
Published online: 17 April 2017

OPEN

Network-assisted analysis of GWAS data identifies a functionally-relevant gene module for childhood-onset asthma

Y. Liu[1,2], M. Brossard[1,2], C. Sarnowski[1,2], A. Vaysse[1,2], M. Moffatt[3], P. Margaritte-Jeannin[1,2], F. Llinares-López[4], M. H. Dizier[1,2], M. Lathrop[5], W. Cookson[3], E. Bouzigon[1,2] & F. Demenais[1,2]

The number of genetic factors associated with asthma remains limited. To identify new genes with an undetected individual effect but collectively influencing asthma risk, we conducted a network-assisted analysis that integrates outcomes of genome-wide association studies (GWAS) and protein-protein interaction networks. We used two GWAS datasets, each consisting of the results of a meta-analysis of nine childhood-onset asthma GWASs (5,924 and 6,043 subjects, respectively). We developed a novel method to compute gene-level *P*-values (fastCGP), and proposed a parallel dense-module search and cross-selection strategy to identify an asthma-associated gene module. We identified a module of 91 genes with a significant joint effect on childhood-onset asthma ($P < 10^{-5}$). This module contained a core subnetwork including genes at known asthma loci and five peripheral subnetworks including relevant candidates. Notably, the core genes were connected to *APP* (encoding amyloid beta precursor protein), a major player in Alzheimer's disease that is known to have immune and inflammatory components. Functional analysis of the module genes revealed four gene clusters involved in innate and adaptive immunity, chemotaxis, cell-adhesion and transcription regulation, which are biologically meaningful processes that may underlie asthma risk. Our findings provide important clues for future research into asthma aetiology.

Asthma is a common chronic inflammatory disease of the airways, characterized by varying age at onset and clinical presentation[1]. It is currently estimated that 334 million people suffer from asthma worldwide and 14% of the world's children experience asthma symptoms[1]. Although environmental factors play an important role in asthma, estimates of heritability of asthma range from 35% to 75%[2], which suggests significant genetic contribution. There have been considerable efforts to characterize the genetic factors underlying asthma, including candidate gene studies, positional cloning studies and more recently genome-wide association studies (GWAS)[3,4]. Although these studies have been successful in identifying novel loci, the genetic factors identified to date explain only a small part of asthma risk. Moreover, heterogeneity is a hallmark of asthma. Genetic heterogeneity according to age of onset of asthma has been evidenced, with genetic factors appearing to play a more important role in childhood-onset asthma[5,6].

Typically, GWAS focus on testing association of disease with individual SNPs over the genome and only top-ranked SNPs with strongest statistical evidence for association are reported. GWAS are therefore underpowered to detect genetic variants which have small marginal effect but rather act jointly or interact with each other in disease or trait variability. To complement the typical single-marker analysis, more sophisticated analyses of GWAS data, which integrate biological knowledge with GWAS outcomes, have been proposed to allow

[1]INSERM, Genetic Variation and Human Diseases Unit, UMR-946, Paris, France. [2]Université Paris Diderot, Université Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Paris, France. [3]Genomic Medicine Section, National Heart Lung Institute, Imperial College London, London, UK. [4]Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland. [5]McGill University and Genome Québec Innovation Centre, Montréal, Québec, Canada. Y. Liu and M. Brossard contributed equally to this work. Correspondence and requests for materials should be addressed to Y.L. (email: yuanlong.liu@inserm.fr) or F.D. (email: florence.demenais@inserm.fr)

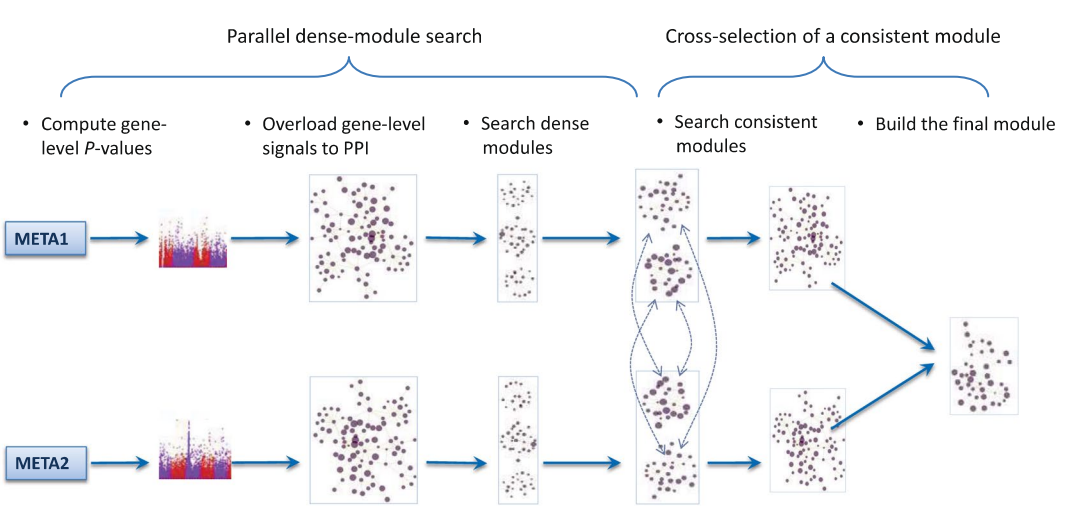


Figure 1. Workflow of the parallel dense-module search and cross-selection strategy. Individual SNP-level *P*-values from two independent childhood-onset asthma GWAS datasets (META1 and META2) were used as input for our network analysis. Gene-level *P*-values, computed from SNP-level *P*-values using fastCGP were converted to z-scores and overloaded to the PPI. The Dense Module Search algorithm was applied to each scored-PPI in parallel to search for dense modules. Modules with highest consistency between the two datasets were selected to build the final module.

detecting sets of functionally related genes that jointly affect disease risk. Among many of these approaches stands the network-assisted analysis that integrates GWAS results with protein-protein interaction (PPI) network to identify gene modules (subnetworks) enriched in association signals. The rationale behind it is the principle of "guilt-by-association", which states connected genes (or gene products) are usually participating in the same, or related, cellular functions[7,8]. Therefore, network-assisted analysis is a promising approach to discover functionally related genes that have a small marginal effect but rather act jointly in disease susceptibility.

In spite of such advantages, network-assisted analysis is also facing challenges. In a classical GWAS, association tests are typically performed at the SNP level, yet the basic entity of PPI network is gene products (proteins). An essential question is how to aggregate signals at SNP-level into gene-level. A popular approach is to take the best SNP from all SNPs mapped to a gene as gene-level *P*-value. However, longer genes represented by more SNPs are more likely to have small *P*-values by chance[9,10]. Another challenge of network-assisted analysis is that most algorithms for searching gene modules are sensitive to the input data: the PPI network and GWAS results. It is well known that GWAS vary in their results for many reasons such as study design, genetic background of the populations, disease heterogeneity, and influence of environmental exposures or simply because of random variation. Therefore, appropriate network-analysis strategies need to be implemented to identify reliable disease-associated gene modules.

In the present study, we conducted a network-assisted analysis by integrating childhood-onset asthma (COA) GWAS results with experimentally verified human PPI network information to identify a set of interconnected genes that significantly contributes to COA risk. We used two large GWAS datasets which consisted of the results of meta-analyses of nine COA GWAS each (5,924 subjects for the first dataset, 6,043 subjects for the second dataset), that were part of the European Gabriel asthma consortium. To address the challenges mentioned above, we first developed an efficient method, named fastCGP, to compute gene-level *P*-values from GWAS SNP-level *P*-values. Then, we used a parallel dense-module search and cross-selection strategy to search for a consistent gene module between the two datasets. We identified a module of 91 genes significantly associated with childhood-onset asthma, including both known and novel candidate genes. Inspection of the interconnected components of this module together with functional enrichment analysis revealed biologically meaningful processes that underlie the risk of childhood-onset asthma.

Results

The different steps of the proposed parallel dense-module search and cross-selection strategy are summarized in Fig. 1.

Identification of a module enriched with childhood-onset asthma-associated genes. We used individual SNP-level *P*-values from two independent COA GWAS datasets as input data to perform network-assisted analysis. These two datasets were the results of meta-analyses of nine COA GWAS each (named META1 and META2, respectively) and included 2,370,689 unique SNPs. The two datasets that were meta-analysed were obtained by randomly splitting the total set of 18 Gabriel Consortium COA GWAS into two sets of similar size (3,031 cases and 2,893 controls in the first set; 2,679 cases and 3,364 controls in the second set).

We first combined SNP-level *P*-values into gene-level *P*-values using a novel gene-based method named fastCGP. fastCGP starts by mapping SNPs to genes (between the start site and 3′-untranslated region of each gene) using dbSNP Build 132 and human Genome Build 37.1, making a total of 24,120 genes with at least one SNP mapped. Then, gene-level *P*-values were taken as the best SNP *P*-values among all SNPs mapped to the

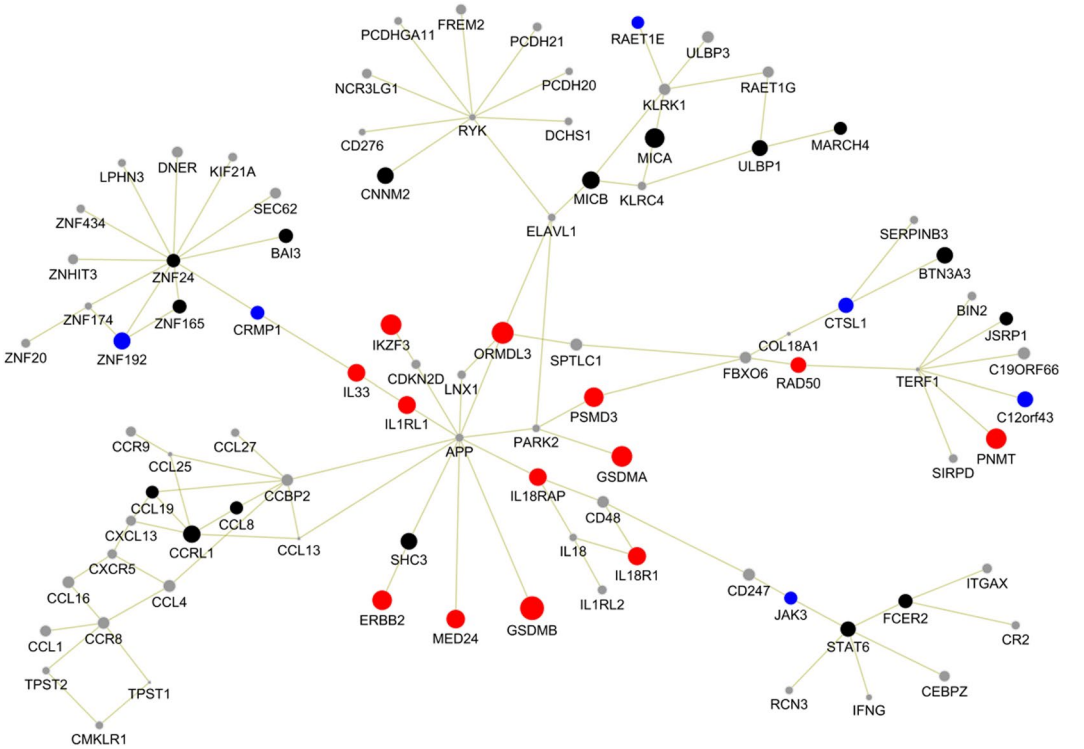


Figure 2. The gene module identified for childhood-onset asthma. The red coloured nodes represent genes at known asthma associated loci and nominally significant in both META1 and META2 datasets; the blue coloured nodes represent new module genes that are nominally significant in both META1 and META2 datasets; the black coloured nodes represent new module genes that are nominally significant in either dataset; the grey coloured nodes are not significant. The node size indicated the strength of the association (the maximum z-score of its corresponding gene in the two datasets).

gene and were corrected for the gene length bias using circular genomic permutation of SNP *P*-values, which allows taking into account the linkage disequilibrium (LD) between SNPs (see Methods and Supplementary Information for details). The resulting gene-level *P*-values in META1 and META2 are shown in Supplementary Fig. S1. Gene-level *P*-values were transformed to z-scores and overloaded to PPI, resulting in a scored-PPI for META1 and a scored-PPI for META2. The scored-PPIs consisted of 12,709 proteins and 123,608 interactions.

The Dense Module Search (DMS) algorithm[11] was applied to each scored-PPI in parallel to identify modules enriched in association signals (quantified by a module score, which is the mean of z-scores of module genes). A total of 10,439 modules were generated in META1 and 10,429 modules in META2. As expected, because of the nature of DMS algorithm, there is considerable gene overlap between the modules generated in each dataset (Supplementary Fig. S2). We reduced such redundancy by hierarchically merging the raw modules within dataset until all pairwise module similarities were smaller than 0.5 (a similarity of r implies two modules share ~100r% of their genes), which resulted in 1,127 modules in META1 and 952 modules in META2. The overlap between the merged modules was largely reduced (Supplementary Fig. S2).

To identify consistent modules between the two datasets, we computed all pairwise module similarities across datasets. The similarity between two modules was defined by the proportion of genes they share (see Methods). Among a total of 1,072,904 module pairs, 95% had low similarity of less than 0.20 but 77 pairs had similarity over 0.40, showing notable consistency of the involved modules (Supplementary Fig. S3). In order to obtain the most consistent results, we selected the top 10 module pairs with highest pairwise similarities. These module pairs had similarity ranging from 0.42 to 0.63. The modules belonging to these 10 module pairs were further merged within each dataset, resulting in a single connected subnetwork of 171 genes with 243 interactions in META1 and 201 genes with 289 interactions in META2 (Supplementary Fig. S4). Finally, we took the intersection of the two subnetworks and obtained a final module of 91 genes with 106 interactions (Fig. 2 and Supplementary Table S1).

Module assessment. The identified gene module showed significant association with childhood-onset asthma, with $P_{assoc} < 10^{-5}$ in both META1 and META2 (using 100,000 circular genomic permutations). This module had a significantly higher score than expected by chance, with $P_{zm} < 10^{-5}$ and $P_{z_m}^{mhrw} < 7.9 \times 10^{-5}$ evaluated in both META1 and META2, where P_{zm} was computed by comparing the identified module score with the scores of 100,000 topology-free random modules, and $P_{z_m}^{mhrw}$ with the scores of 12,709 random modules generated by the modified Metropolis-Hasting Random Walk (MHRW) algorithm that takes into account the connection among genes (see Methods for details). The gene module was also enriched in genes that are nominally significant in at least one dataset ($P_{sig}^{hyper} = 2.3 \times 10^{-12}$ and $P_{sig}^{mhrw} < 7.9 \times 10^{-5}$). Moreover, the number of pairwise connections

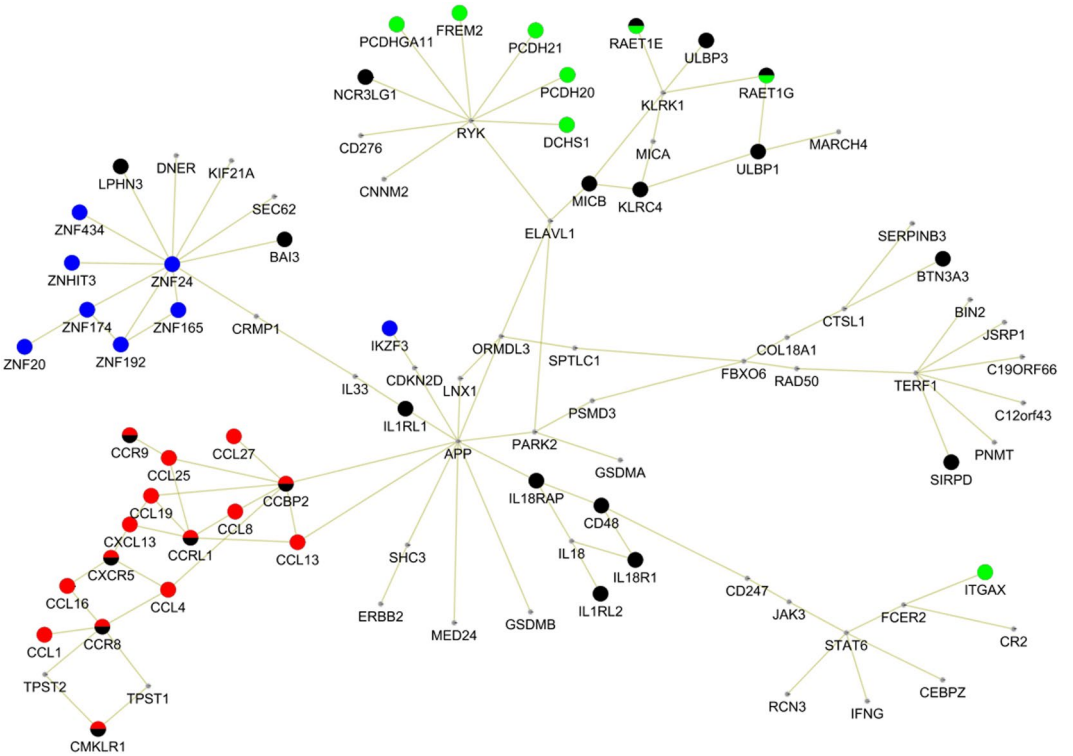


Figure 3. Clusters of functionally-related genes in the COA module. Four genes cluster including a total of 44 out of 91 module genes were identified using DAVID[13,41]. The genes coloured in black belong to the Immune Response cluster; the genes coloured in red belong to the Chemokines/Chemotaxis cluster; the genes coloured in green belong to the Cadherins/Cell-adhesion cluster and the genes coloured in blue belong to the Zinc finger proteins/Transcription regulation cluster. Genes belonging to multiple clusters are marked by mixed colours.

among nominally significant genes was significantly higher for genes in the selected module than for genes outside the selected module ($P_{con} = 3.0 \times 10^{-5}$), implying more functional relatedness among the selected genes.

Out of the 91 genes that belonged to the selected gene module, 19 genes had nominally significant *P*-value in both META1 and META2 datasets. These 19 genes included 13 genes at 4 loci found significantly associated with asthma in previous GWAS (2q12, 5q31, 9p24.1, 17q12-q21)[4,6], and six genes at six distinct loci that are novel: *CRMP1* (4p16.1), *ZNF192* (6p22.1), *RAET1E* (6q24.3), *CTSL1* (9p21.33), *C12orf43* (12q24.31) and *JAK3* (19p13-p12). Among the other 72 genes, 16 genes were nominally significant in one dataset (META1 or META2) while the remaining genes were connected to the nominally significant genes (Supplementary Table S1). The overall module contained a core subnetwork and five peripheral subnetworks connected to the core. The core subnetwork included genes at the known 2q12, 9p24.1 and 17q12-21 loci that were all connected through the *APP* gene (amyloid beta precursor protein) which occupies a central position in this subnetwork. The five peripheral subnetworks, each harbouring multiple nominally significant genes, were brought together through these core genes (Fig. 2). It is of note that the *APP* gene, which encodes the amyloid beta precursor protein, predisposes to dominant forms of Alzheimer's disease (AD) but also harbours rare variants with a protective effect on AD[12]. This protein is cleaved by secretases to form a number of peptides, some of which contribute to amyloid plaques in the brains of patients with AD while others have bactericidal and antifungal activities. We noticed that *APP* is a hub gene in the scored-PPI network, as it has the second highest number of interactors (1,727). Nonetheless, two elements show that it was identified in the final module not only for its "hubness", but also for its interactions with genes strongly associated with COA. First, *APP* was present in 97% of the raw modules generated by DMS, while two other hub genes with comparable number of interactors, *NRF1* (2,174 interactors) and *SUMO2* (1,098 interactors), were included in less than 5% of the raw modules. Second, setting the score of all direct interactors of *APP* in the identified module to zero led to a dramatic decrease from 97% to 4% of the raw modules containing *APP*. This demonstrates that *APP* was identified in the final module mainly for its interactions with strongly COA-associated genes. This link between *APP* and asthma-associated genes suggests potential relationship between AD and asthma that will be further discussed.

Functional clustering and annotations of the identified module genes. The functional and biological relatedness of the module genes were explored using the gene functional classification tool of DAVID[13]. This tool clusters genes into functionally related groups according to gene-to-gene annotation similarities using over 75,000 terms from 14 annotation sources (including KEGG, Gene ontology etc.), allowing a much more comprehensive analysis than enrichment analysis based solely on Gene ontology categories or KEGG pathways. We identified four functional gene clusters which altogether included 48% of the module genes (Fig. 3 and Table 1).

Gene cluster	Functional annotation	Number of genes	List of genes in a cluster
1	Immune response	22	*BAI3, BTN3A3, CCBP2, CCR8, CCR9, CCRL1, CD48, CMKLR1, CXCR5, IL18R1, IL18RAP, IL1RL1, IL1RL2, KLRC4, LPHN3, MICB, NCR3LG1, RAET1E, RAET1G, SIRPD, ULBP1, ULBP3*
2	Chemokines/Chemotaxis	15	*CCBP2, CCL1, CCL13, CCL16, CCL19, CCL25, CCL27, CCL4, CCL8, CCR8, CCR9, CCRL1, CMKLR1, CXCL13, CXCR5*
3	Cadherins/Cell-adhesion	8	*DCHS1, FREM2, ITGAX, PCDH20, PCDH21, PCDHGA11, RAET1E, RAET1G*
4	Zinc finger proteins/Transcription regulation	8	*IKZF3, ZNF165, ZNF174, ZNF192, ZNF20, ZNF24, ZNF434, ZNHIT3*

Table 1. Clusters of functionally-related genes characterised in the childhood-onset asthma module using DAVID[13,41].

The largest cluster (cluster 1) encompassed 22 genes scattered across the module while the three other functional clusters showed almost complete overlap with the peripheral subnetworks. These clusters were annotated by the most representative terms that were "immune response", "chemokines/chemotaxis", "cadherins/cell-adhesion" and "zinc finger proteins/transcription regulation", respectively (Table 1).

Discussion

Network-assisted analysis provides a powerful approach to explore the joint effects of multiple genetic factors on disease and to discover new candidate genes that are missed by single-marker analysis. This is of particular interest for asthma where the number of loci reported by GWAS is relatively small as compared to other common diseases. By integrating the results of two large-scale meta-analyses of childhood-onset asthma with a comprehensive protein-protein interaction network, we identified a gene module of 91 genes that significantly influences COA. This module includes known genes and novel promising candidates. Functional annotation of this module revealed biologically meaningful processes underlying childhood asthma.

The core of the identified module included 11 genes at the three loci that reached genome-wide significance in the meta-analysis of all 18 Gabriel Consortium childhood-onset GWAS, and were also replicated by many other studies[4], which demonstrates the validity of our strategy. The connection of these genes with *APP* in the core subnetwork is of great interest and is supported by a number of studies indicating that asthma and Alzheimer's disease (AD) may share common underlying mechanisms. Epidemiological studies have reported an increased risk of AD and dementia in patients with asthma or other allergic diseases[14,15]. Genetic factors involved in immune-related and inflammatory processes, which are key in asthma, are associated with AD[16]. Epigenetic signatures for both neuronal and immune-response genes were found in a mouse model of AD and in orthologous regions in humans[17]. It has also been recently suggested, in mouse and worm models of AD, that amyloid-β peptide may play a protective role in innate immunity[18]. Finally, in a mouse model of AD, an asthma drug was found to have a potential beneficial impact in AD by decreasing the levels of amyloid-β peptides[19]. The link between *APP* and asthma genes, as highlighted in our module, can open new routes of research for elucidating the functional role and relationships of these genes in asthma, and also potentially, in AD.

The identified module highlighted six novel genes that were nominally significant in both META1 and META2 datasets. The functions of at least four of these genes make them strong candidates for asthma. *RAET1E* (retinoic acid early transcript 1E) belongs to the major histocompatibility complex (MHC) class I-related genes of the *RAETA* family which encode ligands for NKG2D receptor, known to be involved in innate and adaptive immune responses. In the identified module, *RAET1E* is connected to the NKG2D encoding gene *KLRK1*, and through *KLRK1*, to several genes of *MIC* and *RAET/ULBP* families which all encode NKG2D ligands that appear on the surface of stressed cells, such as virus-infected cells[20]. Some of these genes were also nominally significant. This clearly illustrates the usefulness of network analysis in pointing out a set of functionally-related genes that may collectively influence COA, while, individually, they only show nominal association or even no association. Another candidate is *CTSL1*, which encodes a proteinase that acts on the alpha-1 protease inhibitor, a major controlling element of neutrophil elastase activity associated with allergic airway inflammation and severe asthma[21]. *JAK3* (19p13-p12) encodes Janus kinase 3, a member of the Janus kinase family of tyrosine kinases that is predominantly expressed in immune cells and involved in cytokine receptor-mediated intracellular signal transduction. *CRMP1* (collapsin response mediator protein 1) encodes a member of a family of cytosolic phosphoproteins that are expressed in the nervous system but is also an interactor of IL33, a cytokine with a prominent role in asthma[6]. The other two potential candidates, *C12orf43* (chromosome 12 open reading frame 43) and *ZNF192* (encoding a zinc finger protein) have less well known functions.

Our network-assisted analysis is based on the assumption of "guilt-by-association" which states connected genes are usually participating in the same or related cellular functions. We certified the validity of this assumption through gene function clustering analysis. We characterized four gene clusters involving nearly half of the module genes. Three of these clusters, annotated as "chemokines", "cadherins" and "zinc finger proteins", are topologically overlapping with three peripheral subnetworks of the module and are related to the core subnetwork in various ways. The "chemokines" cluster is made of chemokines and their receptors, which are all interconnected in the PPI and are involved in several biological processes that may contribute to asthma pathogenesis, such as recruitment and activation of immune and inflammatory cells, collagen deposition and airway wall remodeling[22]. One component of this cluster, CCBP2 (chemokine binding protein 2), shows direct interaction with the core

protein APP and two nominally significant chemokines, CCL8 and CCL19. It is also part of the broad functional immune response cluster. Furthermore, CCBP2 was found associated with CCL2 chemokine levels in the cerebrospinal fluid of Alzheimer patients[23]. The "cadherins" cluster includes mainly protocadherins that are part of the cadherin superfamily involved in cell adhesion[24]. While protocadherins may contribute to the defect in epithelial barrier function observed in asthma, as suggested for some of them[25], a role of other protocadherins in asthma is still unknown. These proteins, which are interacting in the network, are also functionally clustering together with RAET1E, which is a strong asthma candidate (as described above) and is part of the broad immune response cluster. The "zinc finger proteins" cluster includes proteins that show widespread binding to regulatory regions across the genome[26] but their role in regulating expression of cytokines and other inflammatory proteins, as other transcription factors known to be implicated in asthma[27,28], remains to be established. The zinc finger proteins cluster is linked to the core subnetwork in two ways: through direct interaction of CRMP1 protein, encoded by a nominally significant candidate, with IL33, known to be associated with asthma and part of the core network, and through functional relationship with IKZF3, a zinc finger transcription factor regulating B lymphocyte differentiation, encoded by a gene at the known 17q12-q21 asthma locus[5,6] and part of the core network. All these results show that the integration of association signals with protein-protein interaction network plus functional clustering analysis bring together interrelated biologically meaningful processes that may underlie the risk for asthma.

The analysis strategy proposed in this study included a novel, exact and efficient algorithm to compute gene-level *P*-values from SNP-level *P*-values. Although other existing methods also allow such computation, including VEGAS2[29], MAGMA[30] and PASCAL[31], these methods use raw genotype data, or an external reference SNP panel (e.g., Hapamap2 or 1000 Genomes panel) when the original genotype data are unavailable, to compute the correlation among SNP statistics. Use of an external reference SNP panel has two limitations. First, some SNPs from a GWAS may not be part of the reference panel, thus will be discarded from the analysis and therefore the results of corresponding genes will be affected. Second, the LD structure estimated from an external reference SNP panel may not always reflect the true correlations among SNP *P*-values, especially in datasets from large genetic consortiums which are usually composed of different populations. Advantageously, fastCGP keeps all SNPs for analysis and utilizes the LD pattern existing in the input data. Though fastCGP is permutation-based in nature, the exact implementation we proposed does not require generating any CGP sample, and provides the best obtainable *P*-value without relying on a limited number of samples as required by typical permutation test procedures. It also avoids variation of the outcomes compared to simulation-based methods, such as VEGAS2. Our analytical implementation of fastCGP reduces considerably the computational time as compared to simulation-based approaches (see Supplementary Information). A potential limitation of fastCGP is that the circular genomic permutation strategy it implements corrects each gene-level *P*-value for the average LD across the genome. Thus, genes with higher LD level than average will be undercorrected while genes with lower LD will be overcorrected. However, this issue of within-gene LD variation may not be so critical for fastCPG as it only uses the best SNP *P*-value instead of all SNP *P*-values to compute the gene *P*-value. Moreover, comparison of fastCGP with VEGAS2 and MAGMA showed strong correlation between the results of fastCGP and the other two methods (use *-bestsnp* sub-model for VEGAS2 and *-snp-wise = top,1* sub-model for MAGMA. See Supplementary Information for details).

Besides correlations between SNPs within a gene, linkage disequilibrium may extend over a broad genomic region and create correlations between gene-based *P*-values of nearby genes. This issue of gene clusters has been addressed in pathway-based analysis[32,33], where sensitivity analysis is usually done by including and excluding such genomic regions (e.g., HLA region). For network analysis, such sensitivity analysis is not easy to implement because of the dynamic nature of the module search algorithm and the risk of dismantling the network by removing a few genes. Novel strategies that enable to address this linkage disequilibrium issue deserve further investigation.

It is well known that analyses performed at a genome-wide scale are prone to high rates of false positives. To ensure the reliability of findings, strict criteria have been established for reporting GWAS results, such as the use of a stringent threshold to declare significance and replication of results. However, such criteria are less well defined in pathway and network analyses. It is also worth noting that most module searching methods, including DMS, are based on heuristic or greedy algorithms which do not guarantee finding the module with highest score but may include irrelevant genes by chance[34]. To reduce the amount of false positive findings caused by either the noise from input data or by the module search method, previous network-based studies have used two or more GWAS datasets and implemented cross-evaluation strategies to identify modules showing consistent association signals across datasets[35,36]. In the current study, we used two large GWAS datasets, resulting from a meta-analysis of nine GWAS each, and designed a parallel dense-module search and consistency-based cross-selection strategy to increase the reliability of results. The consistency, defined in terms of similar gene composition between modules obtained from two independent datasets, has the ability to take into account the module topology, and hence to select modules that share genes with association signals and genes closely connected to these disease-associated genes, both of which may play a role in COA susceptibility. We also compared our parallel strategy to a non-parallel strategy by repeating network analysis using a single GWAS dataset made of the meta-analysis results of all 18 childhood-onset asthma GWAS. We selected the same number of genes (91) using the approach proposed by dmGWAS[11]. We found the non-parallel strategy selected less genes that were replicated across datasets (it selected 13 genes nominally significant in both META1 and META2 while the parallel strategy selected 19 such genes), and were less functionally related based on DAVID analysis (the non-parallel module contains two gene clusters that includes 9% of the module genes while the parallel module contains four clusters that includes 48% module genes). This indicates the advantage of using a parallel strategy at least for these asthma data but further studies applied to various datasets are needed to confirm these findings.

As for many other module selection strategies[11,36], our strategy involves the choice of a cut-off defined as the number of consistent modules across datasets to be selected for downstream analysis. We chose the 10 module

pairs (out of 1,072,904 pairs) having highest pairwise module similarity between the two datasets. Although selection of more module pairs may allow including additional candidate genes, it may also increase the complexity of downstream analysis. When we repeated the analysis by loosening the cut-off of module pairs selection, i.e. by selecting 15 or 20 module pairs instead of 10 pairs, no additional relevant information was obtained. Indeed, only one gene out of a maximum of 34 additional genes selected in the final module was nominally significant in the two datasets and belonged to a well-known asthma-associated region on chromosome 17q12-q21. In addition, clustering analysis using DAVID did not identify any additional functional cluster. This shows that our choice of the 10 most consistent module pairs is reasonable. Nonetheless, modules that contribute to asthma susceptibility but did not rank in the top modules may have been missed. More sophisticated methods that allow finding an optimal similarity cut-off deserve further consideration.

In summary, we have derived a comprehensive network-assisted analysis strategy and identified a module of 91 genes significantly contributing to asthma risk. As part of this strategy, we developed an exact and efficient gene-based method (fastCGP) to compute gene-level *P*-values, and a parallel dense-module search and cross-selection strategy to identify an asthma-associated module, both of which are key elements of our network analysis. This study was able to confirm known asthma genes and to pinpoint novel relevant candidates. It also highlighted many links between subnetworks of the identified module and functional relationships both within and across subnetworks, thus providing new clues for future research in both the genetics and pathogenesis of asthma.

Methods

Childhood-onset asthma (COA) GWAS datasets. We used two COA GWAS datasets that consisted of the outcomes of meta-analysis of nine COA GWAS each. These GWAS were part of the European GABRIEL Consortium and have been described in detail elsewhere[6]. Briefly, in these studies, asthma was considered to be present if it had been diagnosed by a physician and childhood-onset asthma was defined as the presence of the disease in a person younger than 16 years. The genotyping of all datasets was performed using the Illumina Human610-Quad Beadchip. Imputations were done using MACH 1.0 and Hapmap2 reference panel (release 21). After quality control filtering (imputation quality score $\geq$0.50 and minor allele frequency $\geq$1%), a total of 2,370,689 autosomal SNPs were kept in the analysis. A total of 18 GABRIEL childhood-onset asthma GWAS, all of European-ancestry, were randomly split into two datasets of 9 GWAS each of similar size (3,031 cases/2,893 controls in the first dataset and 2,679 cases/3,364 controls in the second dataset). A random-effect meta-analysis was performed in each dataset using Stata™ V10.0 (distributed by Stata Corporation, College Station, Texas, USA). The outcomes of these meta-analyses (single-SNP *P*-values) were named META1 and META2 respectively.

Computing gene-level *P*-values by fastCGP. To perform network analysis, gene-level *P*-values representing the significance of each gene for association with COA were computed. We developed an exact and efficient algorithm, named fastCGP, to calculate gene-level *P*-values from SNP-level *P*-values. To start, SNPs were mapped to genes (between the start site and 3′-untranslated region of each gene) using dbSNP Build 132 and human Genome Build 37.1. Each gene-level *P*-value P_g is taken as the best SNP *P*-value among all SNPs mapped to the gene. These *P*-values are biased by gene length (number of SNPs being mapped) since genes with more SNPs mapped tend to have a lower best SNP *P*-value by chance. We corrected for such bias using permutation-based approach. To keep similar patterns of LD among SNPs in the permutated data as in the original data, we implemented the Circular Genomic Permutation (CGP) strategy[37]. Briefly, CGP considers the genome as a circle, starting from chromosome 1 and ending at chromosome 22. SNP-level *P*-values of a GWAS are ordered on the circle according to the position of the SNPs. A CGP sample can be generated by rotating the *P*-values from a randomly chosen position and reassigning these *P*-values to each SNP. As in a typical permutation test, we defined the corrected gene-level *P*-value as $P_{\text{corrected}} = 1 - l/(L+1)$, where L is the total number of CGP samples, and l is the number of samples with $P_{\pi,g} > P_g$ ($P_{\pi,g}$ is the best SNP *P*-value of gene g based on a permutation sample). Particularly, in contrast to general permutation tests classically relying on a limited number of samples, we considered all non-repeating CGP samples in order to obtain the best obtainable *P*-value within this permutation-based framework. In such a case, L becomes the total amount of SNPs placed on the circle (hence the number of SNPs in a GWAS), while l can be calculated analytically and efficiently without generating any CGP sample. The detail of this analytical approach along with an illustrative example is given in Supplementary Information. We implemented fastCGP in R and made it publically available at https://github.com/YuanlongLiu/fastCGP.

We applied fastCGP separately to META1 and META2 to compute gene-level *P*-values. A total of 24,120 genes were analysed for each dataset.

Overloading gene-level signals to protein-protein interaction network. We converted gene-level *P*-values to z-scores by $z_i = \Phi^{-1}(1 - p_i)$, where Φ is the cumulative distribution function of the standard normal distribution. We downloaded the human protein-protein interaction network (PPI) from the Protein Interaction Network Analysis platform[38] (release of May 21, 2014). It integrates annotated protein-protein interaction data from six public curated databases (IntAct, BioGRID, MINT, DIP, HPRD and MIPS/MPact). To reduce the uncertainty of network data, we kept only the interactions having experimental evidence. We overloaded gene scores to the PPI to build a scored-PPI for each of META1 and META2.

Identification of a module enriched with childhood-onset asthma-associated genes. We applied the dense module search (DMS) algorithm implemented in dmGWAS R package[11] within each scored-PPI to search modules that consist of high score genes. Briefly, DMS defines the score of a module of k genes as $Z_m = \sum z_i/\sqrt{k}$. It grows a module from a seed gene and adds the neighbouring gene that can lead to the maxi-

mum increment of the module score. Module growth terminates if adding neighbouring genes does not yield an increment of module score by at least $Z_m \times 0.1$.

Due to the nature of DMS algorithm that uses every gene in the scored-PPI as a seed to grow module, thousands of modules might be generated and there is extensive overlap among them. To reduce such redundancy, we hierarchically merged the raw modules within each dataset until all pairwise Dice similarity were less than 0.50, where the Dice similarity between two modules A and B is defined as $s(A, B) = 2|A \cap B|/(|A| + |B|)$[39]. Here $|\bullet|$ represents the number of genes in a module; $A \cap B$ represents the genes shared by module A and module B.

The original dmGWAS paper suggested selecting 1% of the modules with highest normalized module score. In our study, we were rather interested in modules generated independently from separate datasets but having similar compositions across datasets, thus to improve the reliability of results. Module consistency was assessed by computing all pairwise module similarities, with one module from META1 and another module from META2. We selected the 10 module pairs with highest pairwise similarities and the selected modules were then merged within each dataset. The final module was constructed by taking the shared genes between the two merged modules.

Module assessment. We performed various types of statistical tests to assess different features of the final gene module. First, we assessed whether the module is significantly associated with COA, using META1 and META2 respectively. The null distribution of the module score was estimated by permuting the SNP-level P-values $n = 100{,}000$ times through CGP that takes into account the genomic structure. For each CGP permutation, gene-level P-values were recalculated using fastCGP and module scores were computed. The P-value of association of the module with COA was defined as $P_{assoc} = (\#\{Z_{m(s)} \geq Z_m\} + 1)/(n + 1)$.

Second, we evaluated whether the module selected by our strategy has a higher score than by chance. Two sets of random modules were generated as background. The first consists of $n = 100{,}000$ modules sampled from the scored-PPI without considering their connections (topology-free). Each module has the same number of genes as that of the module under test. The corresponding P-value is $P_{z_m} = (\#\{Z_{m(s)} \geq Z_m\} + 1)/(n + 1)$. The second set of random modules was generated by taking the connection among genes into account. Specifically, we constrained the genes to be connected with each other (directly or indirectly) within each random module, so that they are more biologically related and is more comparable to the module under test. We inherited the Metropolis-Hasting Random Walk (MHRW) algorithm[40] to generate random modules (see Supplementary Information). A total of $N = 12{,}709$ modules were generated. The P-value was defined as $P^{mhrw}_{z_m} = (\#\{Z_{m(s)} \geq Z_m\} + 1)/(N + 1)$. We computed P_{zm} and $P^{mhrw}_{z_m}$ in META1 and META2 respectively.

Third, we assessed whether the selected module is enriched in genes nominally significant in at least one dataset. These genes have high probability of association with asthma hence are of high interest. We used a hypergeometric test to assess whether the selected module contains a higher proportion of such genes than the background. The P-value is defined by $P^{hyper}_{sig} = 1 - F_h(k; K, n, N)$, which is the tail probability of a hypergeometric distribution that a module of K genes contains at least k nominally significant genes, while the whole scored-PPI of N genes contains n nominally significant genes. We also evaluated the significance by comparing with the MHRW random modules. The P-value was defined as $P^{mhrw}_{sig} = (\#\{k(s) \geq k\} + 1)/(N + 1)$, where $k(s)$ is the number of nominally significant genes in a random module.

Finally, we evaluated whether the nominally significant genes in the selected module are more interconnected than those outside the module. We sampled $n = 100{,}000$ times the same number of genes from the unselected nominally significant genes and computed the amount of the connections between them. The P-value is $P_{con} = (\#\{e(s) \geq e\} + 1)/(n + 1)$, where e is the number of connections between nominally significant genes in the identified module, while $e(s)$ is the corresponding number in a sample.

Functional clustering and annotation identified genes. To explore the functional relatedness of genes belonging to the selected module, we conducted the gene functional classification analysis provided by DAVID Bioinformatics Resource[41]. This tool generates a gene-to-gene similarity matrix based on shared functional annotation profiles using over 75,000 terms from 14 annotation sources of different types (ontology, protein domain/family, pathways, functional categories, or disease association) and use a heuristic fuzzy multiple-linkage partitioning to identify functionally related gene clusters. The gene functional classification analysis was run for the list of genes in the final module. We set the genes mapped to the PPI as background and used the default parameters of DAVID in our analysis.

References

1. Martinez, F. D. & Vercelli, D. Asthma. *The Lancet* **382**, 1360–1372 (2013).
2. Los, H., Koppelman, G. H. & Postma, D. S. The importance of genetic influences in asthma. *Eur. Respir. J.* **14**, 1210–1227 (1999).
3. Vercelli, D. Discovering susceptibility genes for asthma and allergy. *Nat. Rev. Immunol.* **8**, 169–182 (2008).
4. Welter, D. *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–1006 (2014).
5. Bouzigon, E. *et al*. Effect of 17q21 Variants and Smoking Exposure in Early-Onset Asthma. *N. Engl. J. Med.* **359**, 1985–1994 (2008).
6. Moffatt, M. F. *et al*. A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363**, 1211–1221 (2010).
7. Oliver, S. Proteomics: Guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
8. Li, Z.-C. *et al*. Identification of drug-target interaction from interactome network with 'guilt-by-association' principle and topology features. *Bioinforma. Oxf. Engl.* **32**, 1057–1064 (2016).
9. Askland, K., Read, C., O'Connell, C. & Moore, J. H. Ion channels and schizophrenia: a gene set-based analytic approach to GWAS data for biological hypothesis testing. *Hum. Genet.* **131**, 373–391 (2012).
10. Jia, P. *et al*. A bias-reducing pathway enrichment analysis of genome-wide association data confirmed association of the MHC region with schizophrenia. *J. Med. Genet.* **49**, 96–103 (2012).

11. Jia, P., Zheng, S., Long, J., Zheng, W. & Zhao, Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinforma. Oxf. Engl.* **27**, 95–102 (2011).
12. Jonsson, T. *et al*. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96–99 (2012).
13. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
14. Eriksson, U. K., Bennet, A. M., Gatz, M. & Dickman, P. W. & Pedersen, N. L. Non-Stroke Cardiovascular Disease and Risk of Alzheimer's Disease and Dementia. *Alzheimer Dis. Assoc. Disord.* **24**, 213–219 (2010).
15. Chen, C.-W. *et al*. Increased risk of dementia in people with previous exposure to general anesthesia: a nationwide population-based case-control study. *Alzheimers Dement. J. Alzheimers Assoc.* **10**, 196–204 (2014).
16. Heneka, M. T., Golenbock, D. T. & Latz, E. Innate immunity in Alzheimer's disease. *Nat. Immunol.* **16**, 229–236 (2015).
17. Gjoneska, E. *et al*. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).
18. Kumar, D. K. V. *et al*. Amyloid-β peptide protects against microbial infection in mouse and worm models of Alzheimer's disease. *Sci. Transl. Med.* **8**, 340ra72 (2016).
19. Hori, Y. *et al*. A Food and Drug Administration-approved asthma therapeutic agent impacts amyloid β in the brain in a transgenic model of Alzheimer disease. *J. Biol. Chem.* **290**, 1966–1978 (2015).
20. Carapito, R. & Bahram, S. Genetics, genomics, and evolutionary biology of NKG2D ligands. *Immunol. Rev.* **267**, 88–116 (2015).
21. Koga, H. *et al*. Inhibition of neutrophil elastase attenuates airway hyperresponsiveness and inflammation in a mouse model of secondary allergen challenge: neutrophil elastase inhibition attenuates allergic airway responses. *Respir. Res.* **14**, 8 (2013).
22. Zlotnik, A. & Yoshie, O. The chemokine superfamily revisited. *Immunity* **36**, 705–716 (2012).
23. Kauwe, J. S. K. *et al*. Genome-Wide Association Study of CSF Levels of 59 Alzheimer's Disease Candidate Proteins: Significant Associations with Proteins Involved in Amyloid Processing and Inflammation. *PLOS Genet* **10**, e1004758 (2014).
24. Yagi, T. & Takeichi, M. Cadherin superfamily genes: functions, genomic organization, and neurologic diversity. *Genes Dev.* **14**, 1169–1180 (2000).
25. Tellez, G. F., Nawijn, M. C. & Koppelman, G. H. Protocadherin-1: epithelial barrier dysfunction in asthma and eczema. *Eur. Respir. J.* **43**, 671–674 (2014).
26. Najafabadi, H. S. *et al*. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.* **33**, 555–562 (2015).
27. Holgate, S. T. Pathogenesis of asthma. *Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol.* **38**, 872–897 (2008).
28. Barnes, P. J. In *European Respiratory Monograph* **8**, 84–113 (2003).
29. Mishra, A. & Macgregor, S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res. Hum. Genet.* **18**, 86–91 (2015).
30. Leeuw, C. A., de Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
31. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLOS Comput Biol* **12**, e1004714 (2016).
32. Hong, M.-G., Pawitan, Y., Magnusson, P. K. E. & Prince, J. A. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.* **126**, 289–301 (2009).
33. Holmans, P. *et al*. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.* **85**, 13–24 (2009).
34. Liu, Y. *et al*. SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. *Bioinformatics*, doi:10.1093/bioinformatics/btx004.
35. Jia, P. *et al*. Network-Assisted Investigation of Combined Causal Signals from Genome-Wide Association Studies in Schizophrenia. *PLOS Comput. Biol.* **8**, e1002587 (2012).
36. Han, S. *et al*. Integrating GWASs and human protein interaction networks identifies a gene subnetwork underlying alcohol dependence. *Am. J. Hum. Genet.* **93**, 1027–1034 (2013).
37. Cabrera, C. P. *et al*. Uncovering networks from genome-wide association studies via circular genomic permutation. *G3 Bethesda Md* **2**, 1067–1075 (2012).
38. Wu, J. *et al*. Integrated network analysis platform for protein-protein interactions. *Nat. Methods* **6**, 75–77 (2009).
39. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**, 297–302 (1945).
40. Gjoka, M., Kurant, M., Butts, C. T. & Markopoulou, A. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *2010 Proceedings IEEE INFOCOM* 1–9, doi:10.1109/INFCOM.2010.5462078 (2010).
41. Huang, D. W. *et al*. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, R183 (2007).

Acknowledgements

This work was funded by the Marie Curie Initial Training Network MLPM2012 No. 316861. This work was also supported by the French National Agency for Research grants (ANR-USPC-2013-EDAGWAS, ANR-11-BSV1-027-GWIS-AM, ANR-15-EPIG-0004-05). Genotyping of the Gabriel data was supported by grants from the European Commission (No. LSHB-CT-2006-018996-GABRIEL) and the Wellcome Trust (WT084703MA). The authors also thank all groups of the GABRIEL asthma consortium (http://www.cng.fr/gabriel/results.html).

Author Contributions

Y.L., M.B., and F.D. conceived and designed the study. Y.L. performed methodological developments and network analysis. M.B. carried out the functional analysis and contributed to biological interpretation of results. F.D. supervised the research. C.S., A.V., P.M.J., M.H.D. and E.B. contributed materials and/or analysis tools. F.L. contributed methodology. M.M., W.C. and M.L. obtained financial support for the Gabriel consortium. Y.L., M.B., F.D. wrote the manuscript and all authors reviewed the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-01058-y

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SUPPLEMENTARY INFORMATION

Network-assisted analysis of GWAS data identifies a functionally-relevant gene module for childhood-onset asthma

Y. Liu[1,2*+], M. Brossard[1,2+], C. Sarnowski[1,2], A. Vaysse[1,2], M. Moffatt[3], P. Margaritte-Jeannin[1,2], F. Llinares-López[4], M.H. Dizier[1,2], M. Lathrop[5], W. Cookson[3], E. Bouzigon[1,2], F. Demenais[1,2*]

[+]These two authors contributed equally to the work

[1]INSERM, Genetic Variation and Human Diseases Unit, UMR-946, Paris, France

[2]Université Paris Diderot, Université Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Paris, France

[3]Genomic Medicine Section, National Heart Lung Institute, Imperial College London, London, UK

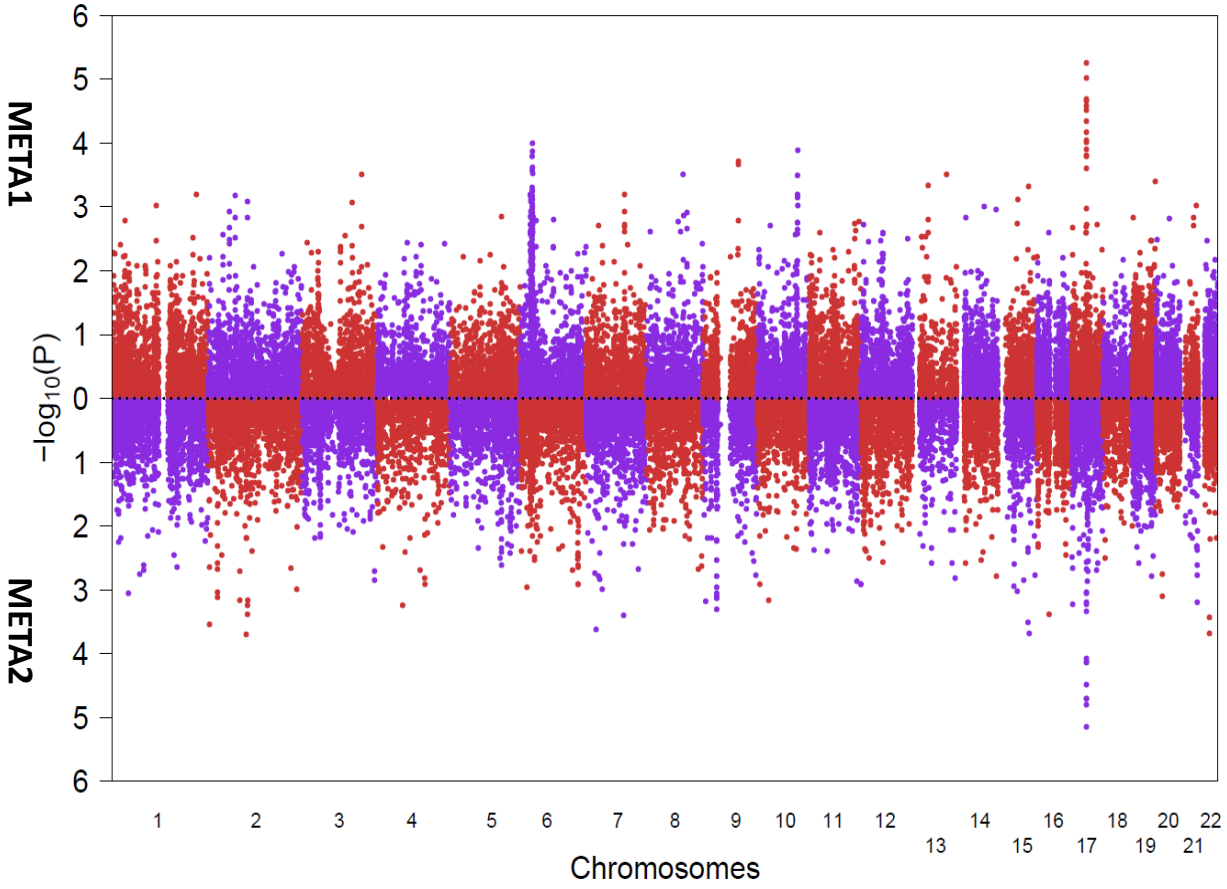
[4]Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[5]McGill University and Genome Québec Innovation Centre, Montréal, Québec, Canada.

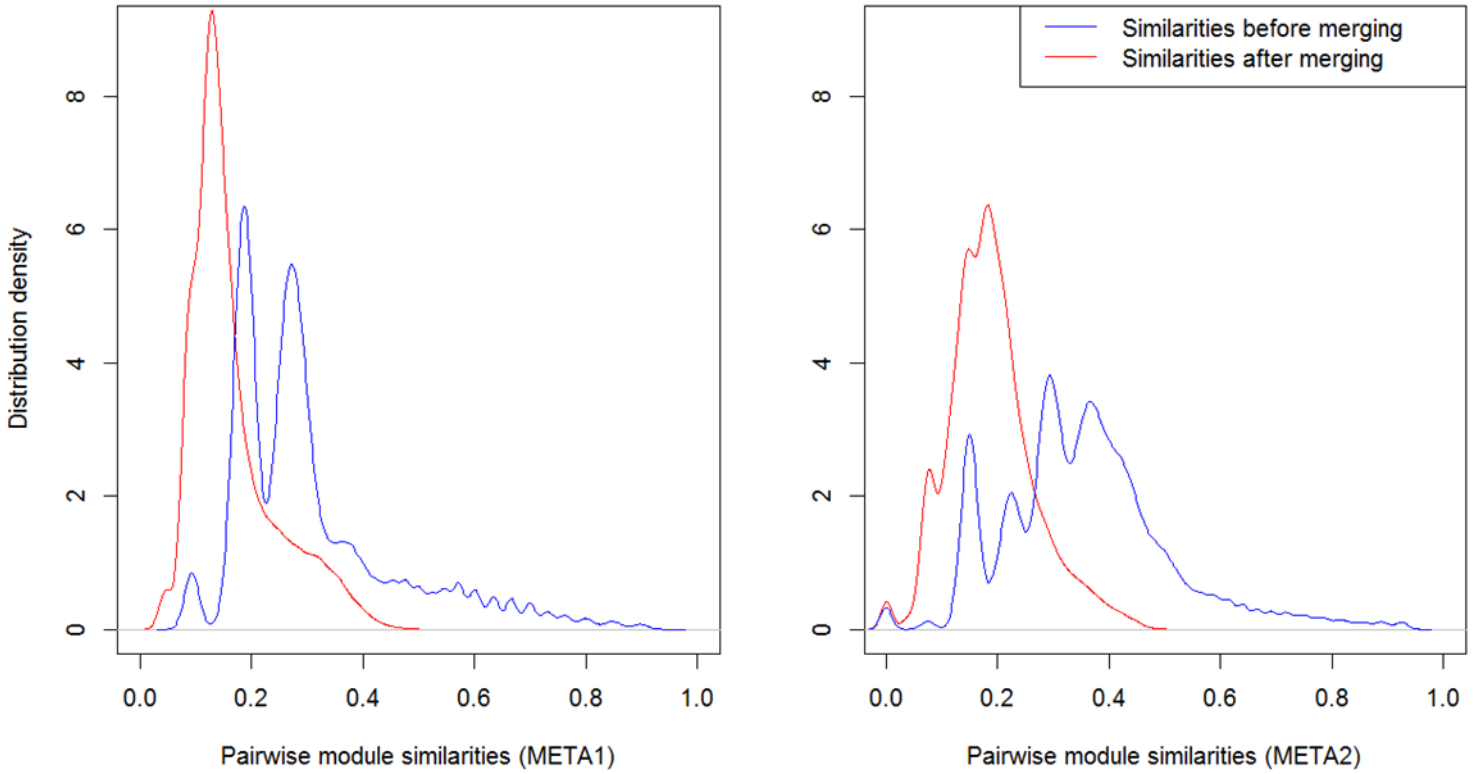
Correspondence should be addressed to Y. Liu (email: yuanlong.liu@inserm.fr) or F. Demenais (email: florence.demenais@inserm.fr)

Table of content

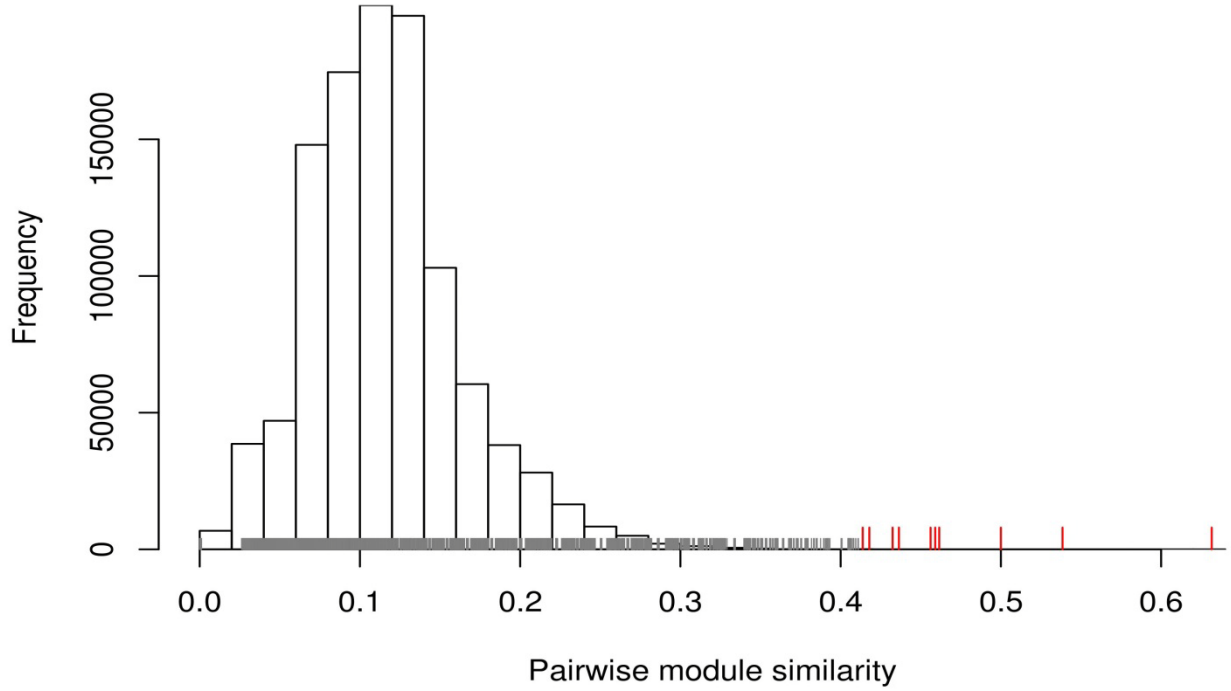
Supplementary Figure S1. Double Manhattan plot of gene-level *P*-values in META1 and META2. Gene-level *P*-values were computed from SNP-level *P*-values using fastCGP. META1 and META2 correspond to the results of meta-analysis of 9 COA GWAS each. The GWAS are part of the GABRIEL asthma consortium (ALSPAC, BAMSE, ECRHS, MAS/MAGICS, SLSJ, TOMSK, UFA, CAPPS studies for META1; B58C, BUSSELTON, EGEA, GABRIEL Advanced Surveys, KSMU, MRCA-UKC, PIAMA, SAPALDIA, SAGE studies for META2; see Moffatt et al[1] for details on these studies)



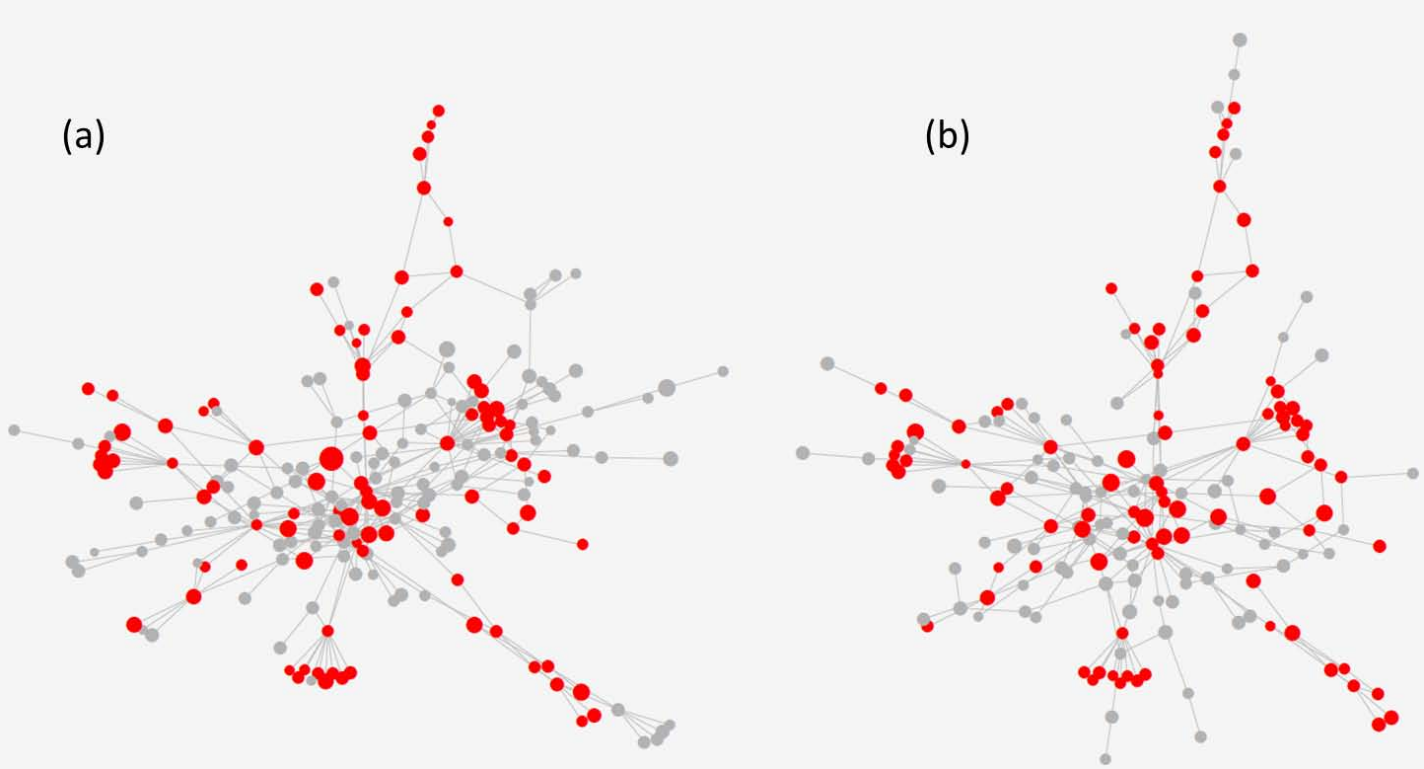
Supplementary Figure S2. Distribution of pairwise module similarities before and after merging the raw modules generated by the Dense Module Search algorithm. The pairwise module similarities (indicating overlaps between modules) were remarkably reduced in both META1 (left panel) and META2 (right panel) after hierarchically merging similar raw modules



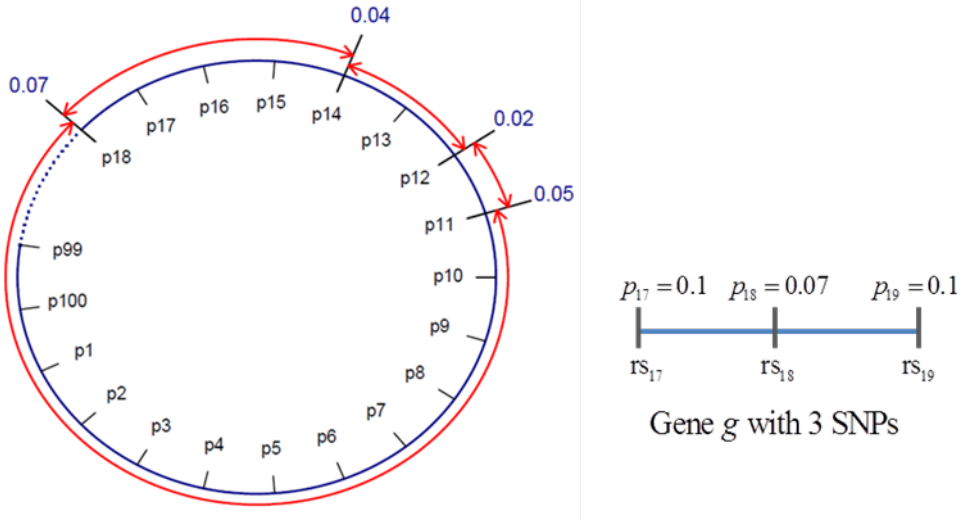
Supplementary Figure S3. Distribution of pairwise module similarities between modules in META1 and modules in META2. A total of 1,072,904 module pairs were constructed. The bins represent histograms of pairwise module similarities. The ticks represent rug plot of the similarities. Each tick represents one pairwise module similarity. The red ticks highlight the 10 highest pairwise module similarities



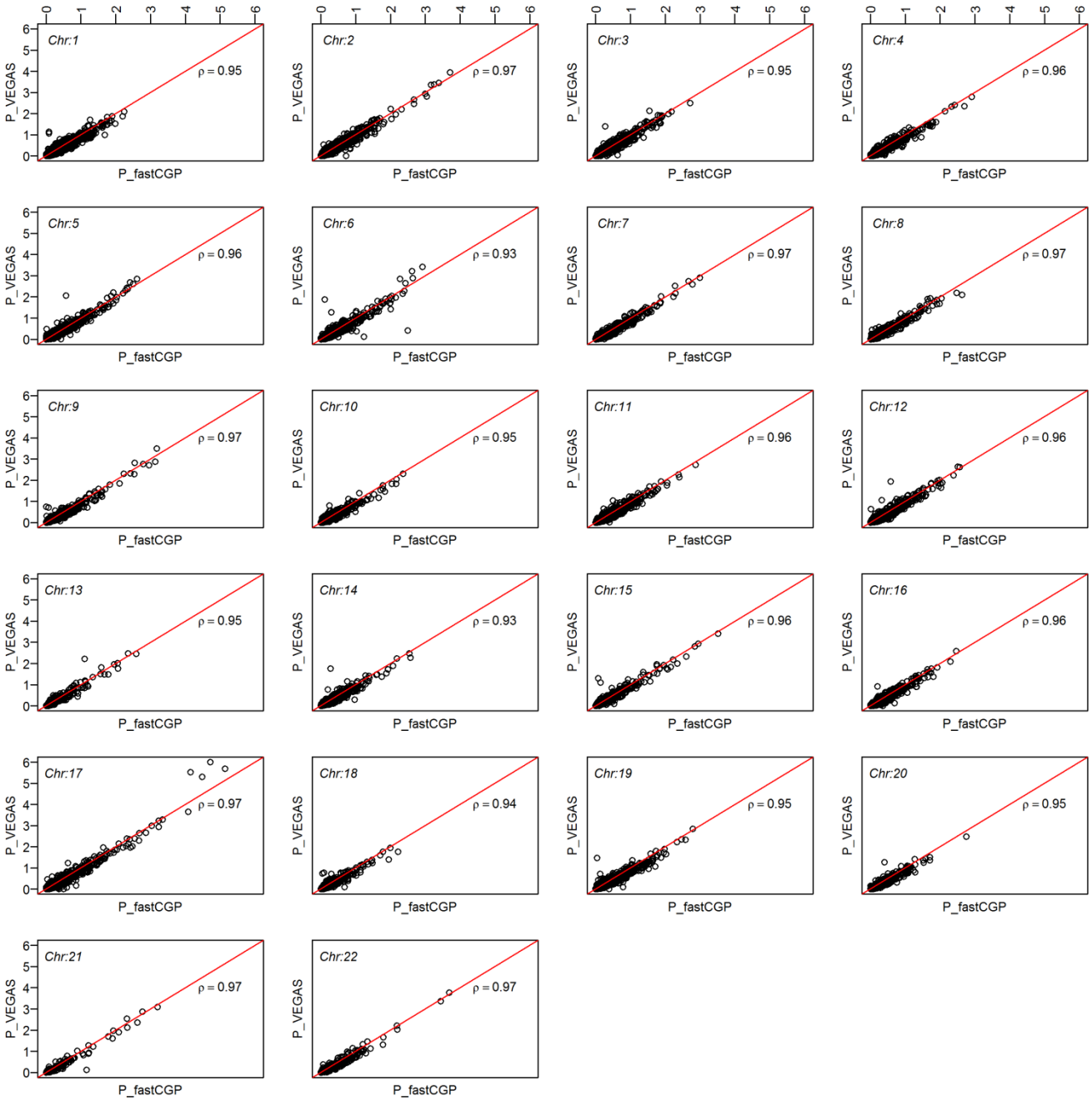
Supplementary Figure S4. Consistent gene modules between META1 and META2. We selected top 10 module pairs showing highest pairwise module similarities among all module pairs between META1 and META2. The involved modules were merged within each dataset, resulting in a subnetwork of 171 genes in META1 (a) and a subnetwork of 201 genes in META2 (b). The intersection of the two subnetworks was retrieved to construct the final module, resulting in a module of 91 genes (nodes in red). The node sizes in this plot are proportional to the gene z-scores



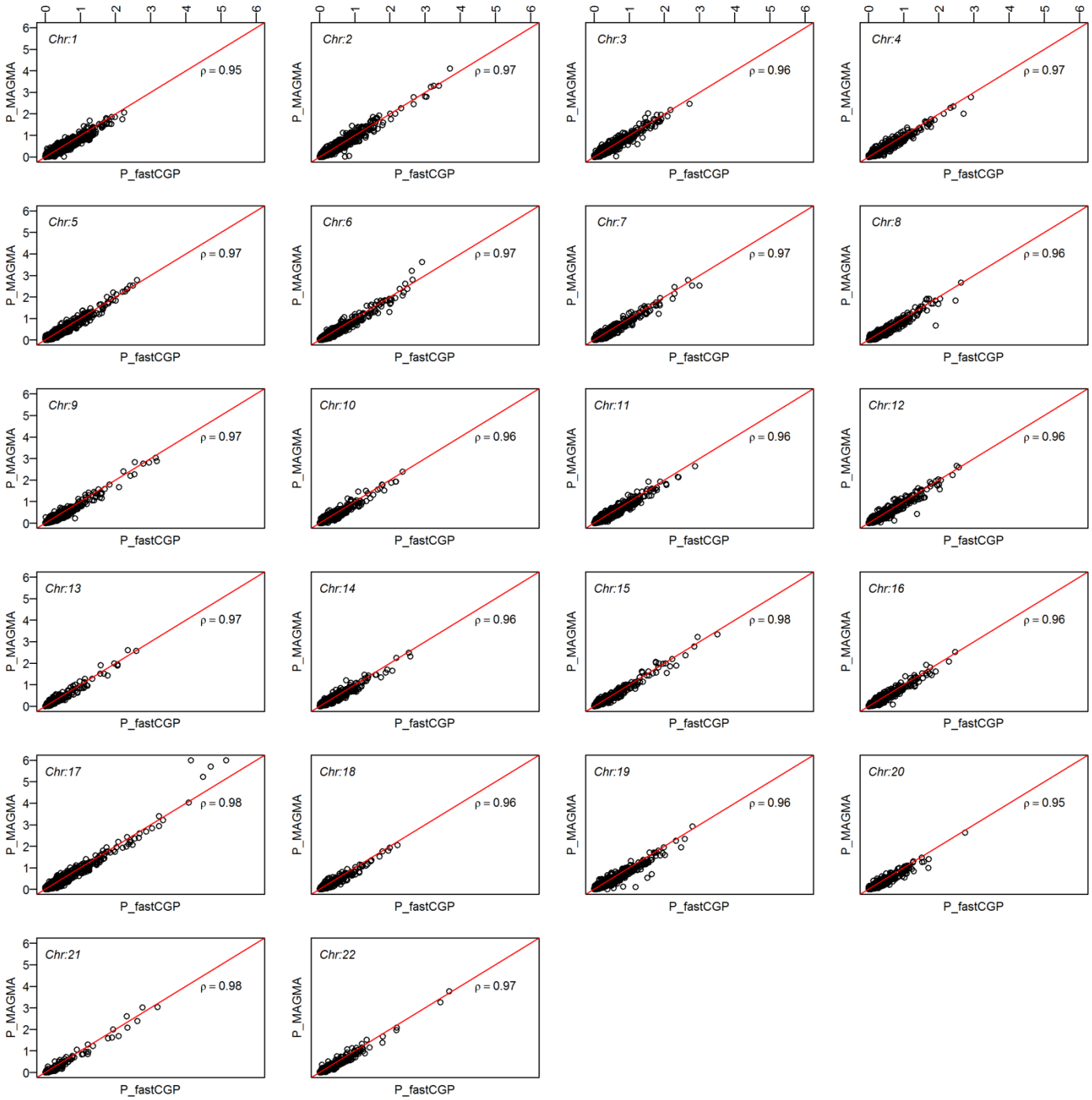
Supplementary Figure S5. An illustrative example of fastCGP. An artificial GWAS result consisting of L=100 SNP *P*-values were created. All *P*-values were set as 0.1 except $P_{11} = 0.05, P_{12} = 0.02, P_{14} = 0.04,$ and $P_{18} = 0.07$. These *P*-values are ordered on a circle according to the chromosomal position of corresponding SNPs. A gene g has three SNPs mapped to its genomic region. Its uncorrected *P*-value P_g is set as the minimum SNP *P*-value among the three mapped SNPs ($P_g = P_{18} = 0.07$). There are four extreme SNP *P*-values on the circle P_{11}, P_{12}, P_{14} and P_{18} ($\leq P_g$). The consecutive extreme *P*-value pairs are $P_{11} \sim P_{12}, P_{12} \sim P_{14}, P_{14} \sim P_{18}, P_{18} \sim P_{11}$



Supplementary Figure S6. Comparison of gene-level *P*-values obtained by fastCGP and VEGAS2[2]. Gene-level *P*-values were computed from asthma META2 dataset using fastCGP and VEGAS2 (*-bestsnp* sub-model). Their results (-log10 (*P*-value)) were compared for each chromosome from chromosome 1 (Chr1) to chromosome 22 (Chr22). The red diagonal lines indicate perfect match (identical) of the two results. ρ represents the Pearson correlation coefficient between the two results



Supplementary Figure S7. Comparison of gene-level *P*-values obtained by fastCGP and MAGMA[3]. Gene-level *P*-values were computed from asthma META2 dataset using fastCGP and MAGMA (*-snp-wise=top,1* sub-model). Their results (-log10 (*P*-value)) were compared for each chromosome from chromosome 1 (Chr1) to chromosome 22 (Chr22). The red diagonal lines indicate perfect match (identical) of the two results. ρ represents the Pearson correlation coefficient between the two results



Supplementary Table S1. Gene-level *P*-values in META1 and META2 datasets for the 91 genes in the final childhood-onset asthma module

Chr	Gene	Start	End	# SNPs	META1			META2		
					Best SNP	Best SNP *P*-value	Corrected Best SNP *P*-value (by fastCGP)	Best SNP	Best SNP *P*-value	Corrected Best SNP *P*-value (by fastCGP)
Genes nominally significant in both META1 and META2										
2	***IL1RL1***	102 927 962	102 968 497	80	rs4988957	2.9E-05	1.5E-03	rs10192157	2.2E-05	5.6E-04
2	***IL18R1***	102 979 097	103 015 218	48	rs3771166	1.9E-05	8.2E-04	rs1974675	2.4E-05	4.1E-04
2	***IL18RAP***	103 035 254	103 069 025	71	rs6543135	2.6E-03	4.5E-02	rs2310300	2.6E-05	6.9E-04
4	*CRMP1*	5 822 491	5 894 785	60	rs13144677	2.7E-03	4.1E-02	rs10011385	1.3E-03	1.7E-02
5	***RAD50***	131 892 630	131 979 599	35	rs2240032	1.2E-03	1.5E-02	rs2897443	3.8E-04	3.9E-03
6	*ZNF192*	28 109 716	28 125 236	21	rs13205911	6.4E-05	1.2E-03	rs2622321	7.4E-03	4.1E-02
6	*RAET1E*	150 209 601	150 212 097	10	rs9371533	1.1E-02	4.5E-02	rs9371533	1.0E-02	3.6E-02
9	***IL33***	6 241 678	6 257 982	20	rs7019575	4.8E-03	3.2E-02	rs7019575	8.7E-05	6.7E-04
9	*CTSL1*	90 340 974	90 346 384	4	rs2378757	1.8E-03	5.6E-03	rs2378757	6.9E-03	1.4E-02
12	*C12orf43*	121 440 848	121 454 300	10	rs3751150	4.4E-04	3.1E-03	rs3751151	1.2E-02	4.3E-02
17	***PNMT***	37 824 507	37 826 728	1	rs876493	1.3E-07	3.1E-05	rs876493	4.0E-05	7.3E-05
17	***ERBB2***	37 844 393	37 884 915	7	rs1058808	3.5E-06	1.6E-04	rs1058808	1.6E-05	8.2E-05
17	***IKZF3***	37 921 198	38 020 441	48	rs907091	2.5E-15	4.6E-05	rs9909593	1.6E-07	7.1E-05
17	***GSDMB***	38 060 848	38 074 903	12	rs9303281	2.6E-16	5.5E-06	rs2305480	7.5E-08	1.9E-05
17	***ORMDL3***	38 077 296	38 083 854	4	rs12603332	4.6E-15	9.7E-06	rs8076131	7.3E-08	7.2E-06
17	***GSDMA***	38 119 226	38 134 019	6	rs7212938	2.4E-13	2.7E-05	rs3902025	3.2E-07	3.3E-05
17	***PSMD3***	38 137 060	38 154 212	21	rs11655264	1.0E-07	9.6E-05	rs12453334	6.9E-05	5.8E-04
17	***MED24***	38 175 350	38 210 889	21	rs12309	3.0E-06	2.5E-04	rs12451897	1.2E-04	9.2E-04
19	*JAK3*	17 935 591	17 958 841	12	rs2110586	5.8E-03	2.8E-02	rs3212701	1.2E-02	4.5E-02
Genes nominally significant in either META1 or META2										
2	*MARCH4*	217 122 585	217 236 750	113	rs1477235	3.0E-02	4.0E-01	rs1510836	1.3E-03	2.8E-02
3	*CCRL1*	132 316 094	132 321 382	2	rs7626622	3.0E-04	8.5E-04	rs7626622	1.7E-01	2.2E-01
6	*BTN3A3*	26 440 763	26 453 643	15	rs13220495	1.8E-04	1.9E-03	rs3846845	1.3E-01	4.3E-01
6	*ZNF165*	28 046 572	28 057 341	9	rs1321505	3.7E-03	1.6E-02	rs203878	2.8E-02	8.8E-02
6	*MICA*	31 371 371	31 383 090	41	rs2844518	3.1E-08	1.0E-04	rs12213831	6.4E-02	4.0E-01
6	*MICB*	31 465 855	31 478 901	26	rs3130614	2.9E-05	7.6E-04	rs2855814	3.5E-01	9.0E-01
6	*BAI3*	69 345 632	70 099 403	770	rs3757043	4.6E-05	1.3E-02	rs17502590	3.7E-02	9.5E-01
6	*ULBP1*	150 285 143	150 294 846	6	rs9478311	1.2E-01	2.8E-01	rs9478311	1.2E-03	3.2E-03
9	*CCL19*	34 689 567	34 691 274	2	rs3176813	4.1E-02	5.5E-02	rs3136658	2.6E-02	3.3E-02
9	*SHC3*	91 628 046	91 793 682	134	rs1331180	2.0E-05	1.6E-03	rs2316280	3.6E-02	4.9E-01

10	*CNNM2*	104 678 114	104 838 241	129	rs943036	2.3E-05	1.8E-03	rs2296569	6.5E-02	6.7E-01
12	*STAT6*	57 489 191	57 505 161	10	rs324015	8.7E-04	5.3E-03	rs1059513	2.0E-02	6.8E-02
17	*CCL8*	32 646 066	32 648 421	2	rs3138036	9.2E-01	9.7E-01	rs3138036	2.2E-02	2.8E-02
18	*ZNF24*	32 912 178	32 924 426	12	rs7239712	4.2E-03	2.2E-02	rs1064753	1.9E-02	7.2E-02
19	*JSRP1*	2 252 252	2 255 344	3	rs7250822	1.1E-02	2.2E-02	rs7250822	2.0E-01	3.2E-01
19	*FCER2*	7 753 643	7 767 032	22	rs2287866	1.4E-03	1.3E-02	rs2303112	1.4E-02	7.8E-02
Other genes										
1	*FBXO6*	11 724 150	11 734 411	2	rs747863	6.1E-01	7.5E-01	rs747863	6.1E-02	7.9E-02
1	*CD48*	160 648 536	160 681 585	14	rs3796502	1.6E-02	7.2E-02	rs10489636	8.2E-02	2.9E-01
1	*CD247*	167 399 877	167 487 847	116	rs864537	2.7E-03	6.7E-02	rs864537	3.0E-03	6.0E-02
1	*CR2*	207 627 670	207 663 240	41	rs17258982	4.6E-02	3.2E-01	rs7543913	9.1E-02	5.1E-01
2	*CEBPZ*	37 428 772	37 458 740	12	rs2239650	3.4E-02	1.3E-01	rs11689186	1.0E-01	3.3E-01
2	*IL1RL2*	102 803 433	102 855 811	134	rs17026782	6.6E-02	6.9E-01	rs12987222	1.4E-02	2.4E-01
2	*DNER*	230 222 345	230 579 286	353	rs10192168	1.6E-03	1.1E-01	rs6726280	3.2E-03	1.6E-01
3	*CCR8*	39 371 197	39 375 171	2	rs2853699	5.4E-02	7.2E-02	rs4676633	2.3E-01	3.0E-01
3	*CCBP2*	42 850 964	42 908 775	36	rs13093968	8.3E-03	7.4E-02	rs4396867	4.4E-01	9.7E-01
3	*CCR9*	45 928 019	45 944 667	12	rs17714101	4.5E-02	1.6E-01	rs6441931	2.6E-01	6.7E-01
3	*RYK*	133 875 978	133 969 586	52	rs4280635	8.4E-02	5.4E-01	rs10935104	7.5E-02	4.9E-01
3	*SEC62*	169 684 580	169 716 161	23	rs9813592	1.8E-02	1.1E-01	rs16854694	3.2E-01	8.5E-01
4	*LNX1*	54 326 437	54 457 724	156	rs2117600	2.6E-02	4.4E-01	rs9312642	1.5E-02	2.9E-01
4	*LPHN3*	62 362 839	62 938 168	336	rs17082520	1.3E-02	4.6E-01	rs1497906	8.9E-03	3.4E-01
4	*CXCL13*	78 432 907	78 532 988	27	rs17406477	2.1E-01	7.3E-01	rs355687	2.8E-02	1.6E-01
5	*PCDHGA11*	140 800 537	140 892 546	57	rs11958830	4.5E-02	3.7E-01	rs1423149	1.8E-01	8.1E-01
6	*RAET1G*	150 238 014	150 244 214	4	rs6927913	2.3E-01	4.2E-01	rs9397070	5.0E-02	9.7E-02
6	*ULBP3*	150 385 743	150 390 202	7	rs12202737	3.0E-01	6.3E-01	rs2010212	2.7E-02	7.3E-02
6	*PARK2*	161 768 590	163 148 834	1776	rs4623220	1.6E-03	3.9E-01	rs11966738	1.9E-03	4.0E-01
7	*TPST1*	65 670 259	65 825 438	68	rs778732	3.8E-01	9.8E-01	rs4149463	1.9E-01	8.6E-01
8	*TERF1*	73 921 097	73 959 987	26	rs12334686	2.4E-01	7.7E-01	rs10107605	5.8E-01	9.9E-01
9	*CCL27*	34 661 893	34 662 689	1	rs11575584	6.3E-01	6.1E-01	rs11575584	3.4E-01	3.1E-01
9	*SPTLC1*	94 793 427	94 877 690	74	rs16908106	9.5E-02	6.6E-01	rs12235495	4.1E-03	5.8E-02
10	*PCDH21*	85 954 517	85 977 122	34	rs12781048	2.1E-01	7.6E-01	rs12781048	4.1E-02	2.5E-01
11	*DCHS1*	6 642 558	6 677 080	28	rs11607376	7.1E-02	3.6E-01	rs997263	2.0E-01	7.1E-01
11	*NCR3LG1*	17 373 279	17 398 868	13	rs6486364	4.9E-02	1.9E-01	rs12791318	1.1E-01	3.7E-01
11	*IL18*	112 013 976	112 034 840	17	rs1834481	1.0E-01	3.9E-01	rs2043055	1.5E-01	5.1E-01
11	*CXCR5*	118 754 541	118 766 971	14	rs12363277	1.0E-01	3.6E-01	rs12363277	5.1E-02	1.9E-01
12	*KLRK1*	10 524 952	10 542 640	20	rs2617149	1.5E-02	8.2E-02	rs12826560	9.0E-02	3.7E-01

12	*KLRC4*	10 559 983	10 562 356	5	rs2734565	1.4E-01	2.9E-01	rs2617170	4.2E-01	7.2E-01
12	*KIF21A*	39 687 030	39 836 918	71	rs11171691	2.4E-02	2.6E-01	rs11172108	4.9E-02	4.3E-01
12	*BIN2*	51 674 822	51 717 938	20	rs4761998	6.4E-02	2.9E-01	rs4761995	5.9E-02	2.6E-01
12	*IFNG*	68 548 550	68 553 521	4	rs1861494	6.6E-01	8.9E-01	rs1861493	3.4E-01	5.8E-01
12	*CMKLR1*	108 681 821	108 733 094	48	rs11113818	8.2E-02	5.1E-01	rs10861889	5.1E-02	3.6E-01
13	*FREM2*	39 261 173	39 461 268	220	rs11618650	4.1E-03	1.5E-01	rs2218722	1.3E-02	3.3E-01
13	*PCDH20*	61 983 818	61 989 655	6	rs3829388	1.8E-01	3.9E-01	rs3812872	3.0E-01	6.1E-01
15	*CD276*	73 976 622	74 006 859	35	rs12591553	3.1E-01	8.9E-01	rs12594595	9.0E-02	4.7E-01
16	*ZNF434*	3 432 085	3 451 025	14	rs28603	8.1E-02	2.9E-01	rs17136367	2.3E-01	6.4E-01
16	*ZNF174*	3 451 190	3 459 364	5	rs39728	3.7E-01	6.6E-01	rs37811	2.1E-01	4.3E-01
16	*ITGAX*	31 366 509	31 394 318	16	rs2929	1.9E-01	6.0E-01	rs8052139	4.7E-02	1.9E-01
17	*CCL13*	32 683 471	32 685 629	2	rs159313	7.2E-01	8.5E-01	rs2072069	8.5E-01	9.4E-01
17	*CCL1*	32 687 399	32 690 252	5	rs3136682	4.4E-02	1.0E-01	rs3136682	2.3E-01	4.5E-01
17	*CCL16*	34 303 535	34 308 523	4	rs2063979	8.2E-01	9.7E-01	rs11080369	3.4E-02	6.6E-02
17	*CCL4*	34 431 220	34 433 014	2	rs1719147	4.4E-02	6.0E-02	rs1634517	5.0E-01	6.3E-01
17	*ZNHIT3*	34 842 473	34 851 662	4	rs2306589	8.1E-02	1.6E-01	rs2277662	2.4E-01	4.3E-01
18	*SERPINB3*	61 322 431	61 329 197	3	rs1065205	4.5E-01	6.7E-01	rs7228687	2.0E-01	3.3E-01
19	*ELAVL1*	8 023 457	8 070 529	24	rs7251814	3.9E-01	9.1E-01	rs12977189	8.3E-02	3.8E-01
19	*CCL25*	8 117 934	8 127 547	8	rs2287936	3.9E-01	7.7E-01	rs2032887	3.5E-01	7.2E-01
19	*C19orf66*	10 196 806	10 203 928	2	rs2232066	3.9E-02	5.3E-02	rs2232066	9.5E-02	1.2E-01
19	*CDKN2D*	10 677 138	10 679 655	1	rs1465701	6.3E-01	6.1E-01	rs1465701	2.7E-01	2.4E-01
19	*ZNF20*	12 242 803	12 251 140	2	rs155955	2.1E-01	2.8E-01	rs12608894	1.9E-01	2.5E-01
19	*RCN3*	50 030 875	50 046 890	5	rs10419198	3.3E-01	6.1E-01	rs8108243	1.9E-01	3.9E-01
20	*SIRPD*	1 514 897	1 538 343	26	rs16995146	5.1E-02	2.7E-01	rs2249673	2.3E-01	7.5E-01
21	*APP*	27 252 861	27 543 446	306	rs7281055	8.2E-03	3.2E-01	rs2830076	3.1E-02	6.8E-01
21	*COL18A1*	46 825 097	46 933 634	82	rs2236483	1.3E-01	7.8E-01	rs17004785	2.0E-01	9.1E-01
22	*TPST2*	26 921 714	26 986 089	94	rs5752349	3.4E-02	4.0E-01	rs4149484	4.7E-02	4.8E-01

Start and end positions of each gene are in accordance with Build 37.1.

The genes at known asthma loci are in bold.

Supplementary Methods 1: computing gene-level *P*-values via fastCGP. We take advantage of the Circular Genomic Permutation (CGP) strategy[4] and propose an efficient and exact method, named fastCGP, to compute gene-level *P*-values from SNP-level *P*-values of a GWAS. CGP is a randomization method that permutes SNP-level statistics in a genomic manner to preserve the genomic structure such as regional linkage disequilibrium (LD), thereby to keep similar patterns of correlation in the permutated data as in the original data. Briefly, it considers the genome to be circular and ordered from chromosome 1 to chromosome 22. SNP-level *P*-values of a GWAS are ordered according to the position of the SNPs on the circle. A CGP sample is generated by rotating the ordered statistics for a random position and reassigning them to each SNP. This randomization strategy has been successfully applied to several studies[5–7], and was shown to have similar performances compared to the gold standard of phenotype permutation in the context of pathway analysis[8].

Our method starts by mapping SNPs to genes (between the start site and 3'-untranslated region of each gene) using dbSNP Build 132 and human Genome Build 37.1 (a user can choose another mapping strategy). The gene-level *P*-value of a gene g, denoted as P_g, is represented by the best SNP *P*-value among all SNPs mapped to the gene. This *P*-value is biased by gene length (amount of mapped SNPs) as genes with more SNPs mapped tend to have a lower best SNP *P*-value by chance. We correct for such bias using a permutation test framework. We define the corrected *P*-value as $P_{\text{corrected}} = 1 - l/(L+1)$, where L is the total number of CGP samples; l is the number of samples with $P_{\pi,g} > P_g$, which we call as *normal CGP samples* (nCGP). Particularly, we include all non-repeating CGP samples so that to obtain the best obtainable *P*-value within this permutation test framework. In this case, L becomes the total amount of SNPs placed on the circle (hence is the number of SNPs in a GWAS), while l can be calculated analytically without generating any CGP sample. For illustration convenience, we call a SNP *P*-value as an extreme *P*-values if it is less than or equal to P_g. We say two extreme *P*-values are consecutive if there is no other extreme *P*-value placed between them on the circle. Note that the SNPs mapped to a gene are consecutive on the circle, hence by rotating the SNP *P*-values, it generates a nCGP only if all *P*-values reassigned to gene g are located between some pair of consecutive extreme *P*-values, say, $P_i \sim P_j$ ($1 \le i, j \le L$ are the SNP positions on the circle). Denote d_{ij} as the number of positions between P_i and P_j on the circle, m as the number of SNPs assigned to the *gene,* $I_{(\cdot)}$ as the indicator function. Then the number of unique rotations with all reassigned *P*-values located within $P_i \sim P_j$ is equal to $\gamma_{ij} = (d_{ij} - m + 1) I_{(d_{ij} \ge m)}$. Since the total amount of non-repeating nCGPs is the summation of γ_{ij} for all pairs of consecutive extreme *P*-

values, this leads to the formula $P_{\text{corrected}} = 1 - \sum_{i \sim j} \gamma_{ij} / (L+1)$. The complete algorithm is summarized below

Algorithm Computation of corrected gene-level *P*-values via fastCGP

Step 1.	Order GWAS SNP *P*-values on a circle according to the genomic positions of SNPs
Step 2.	Map SNPs to genes according to their genomic positions
Step 3.	For a gene g, set P_g as the minimum *P*-value of all SNPs mapped to it
Step 4.	Find all extreme SNP *P*-values on the circle: $\{P_{SNP} \mid P_{SNP} \le P_g\}$
Step 5.	Compute $\gamma_{ij} = (d_{ij} - m + 1) I_{(d_{ij} \ge m)}$ for all pairs of consecutive extreme *P*-value $P_i \sim P_j$
Step 6.	Compute the corrected gene-level *P*-value: $P_{\text{corrected}} = 1 - \sum_{i \sim j} \gamma_{ij} / (L+1)$

In the following we present an illustrative example of fastCGP. We constructed an artificial GWAS result consisting of $L = 100$ SNP *P*-values (Supplementary Fig. S5). We set all *P*-values as 0.1 except $P_{11} = 0.05, P_{12} = 0.02, P_{14} = 0.04$, and $P_{18} = 0.07$. These *P*-values are ordered on a circle according to the chromosomal position of corresponding SNPs. A gene g has $m = 3$ SNPs mapped to its genomic region. Its uncorrected *P*-value P_g is set as the minimum SNP *P*-value among the three mapped SNPs ($P_g = P_{18} = 0.07$). There are four extreme SNP *P*-values on the circle : P_{11}, P_{12}, P_{14} and P_{18}. The consecutive extreme *P*-value pairs are $P_{11} \sim P_{12}, P_{12} \sim P_{14}, P_{14} \sim P_{18}, P_{18} \sim P_{11}$. For each pair, the amount of unique rotations with all SNP *P*-values reassigned to gene g falling into this pair is 0, 0, 1 and 92 respectively. Thereby $P_{corrected} = 1 - (0 + 0 + 1 + 92)/101 = 0.079$.

We compared the performance of fastCGP with two other popular methods VEGAS2[2] and MAGMA[3]. The META2 asthma dataset in our study was used for the comparison. For both VEGAS2 and MAGMA we implemented their best-SNP sub-model (use *-bestsnp* option for VEGAS2 and *snp-wise=top,1* option for MAGMA). Both methods apply Monte-Carlo simulations to correct the best-SNP *P*-value for the gene length bias. During simulation, the LD patterns between SNPs within a gene are estimated on the basis of the LD structure of a set of reference individuals. As we could not use the original genotypes of our asthma dataset, we used the 1,000 Genomes European population as external reference. We observed that the results obtained by fastCGP were concordant with those obtained using VEGAS2 or MAGMA. At the chromosome level, the Pearson correlation coefficients between the gene-level *P*-values (-log10 transformed) of fastCGP and VEGAS2 range from 0.93 to 0.97, with an average value of 0.96 (Supplementary Fig. S6). The correlation coefficients of gene-level *P*-values between fastCGP and MAGMA range from 0.95 to 0.98, with an average value of 0.97 (Supplementary Fig. S7). As

for computational efficiency, both fastCGP and MAGMA took ~30 minutes on a PC (Intel Core i7 3.40GHz CPU, 8GB RAM); while VEGAS2 took 13 hours to perform the analysis.

Supplementary Methods 2: generating random modules via Metropolis-Hasting Random Walk algorithm. We inherited the Metropolis-Hasting Random Walk (MHRW) algorithm[9] to generate random modules. It has the property that the stationary probability of each node to be sampled follows the uniform distribution, and has been demonstrated to work well in practice[9]. In the beginning, a seed gene V is chosen from the scored-PPI. The next gene W is selected at random from all neighbours of V . W is added to the module and set as the next seed if the degree ratio of V and W is smaller than a random number drawn from uniform distribution $U(0,1)$. Otherwise stay at V and repeat the step. The procedure iterates until the module has the same number of genes as the module under test. To ensure sufficient coverage of the whole scored-PPI, we set each gene in the network as a seed to generate a module. Then, a total of N (equals to the number of genes in the scored-PPI) random modules are generated.

References

1. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363,** 1211–1221 (2010).
2. Mishra, A. & Macgregor, S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res. Hum. Genet.* **18,** 86–91 (2015).
3. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11,** e1004219 (2015).
4. Cabrera, C. P. *et al.* Uncovering networks from genome-wide association studies via circular genomic permutation. *G3 Bethesda Md* **2,** 1067–1075 (2012).
5. Stainton, J. J. *et al.* Detecting signatures of selection in nine distinct lines of broiler chickens. *Anim. Genet.* **46,** 37–49 (2015).
6. Mott, R. *et al.* The architecture of parent-of-origin effects in mice. *Cell* **156,** 332–342 (2014).

7. Chambers, E. V., Bickmore, W. A. & Semple, C. A. Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS Comput. Biol.* **9,** e1003017 (2013).
8. Brossard, M. *et al.* Comparison of permutations strategies to assess gene-set significance in gene-set-enrichment analysis. *22nd annual meeting of the International Genetic Epidemiology Society (IGES), 15-17 Sept. 2013, Chicago (USA). Abstract published online at www.geneticepi.org/meeting-abstracts/.*
9. Gjoka, M., Kurant, M., Butts, C. T. & Markopoulou, A. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. in *2010 Proceedings IEEE INFOCOM* 1–9 (2010). doi:10.1109/INFCOM.2010.5462078

CHAPTER IV. THE SIGMOD METHOD FOR IDENTIFYING DISEASE-ASSOCIATED GENE MODULES

1 Summary

The search for active modules consisting of closely related genes and enriched in high association signals plays an essential role in network-based analysis of GWAS data. This is a difficult task when confronted with the noise in the GWAS outcomes, the heterogeneity of quality of information in gene/protein network, and the huge search space at genome-wide scale. Ideally, a module search method should have as much as these properties: (1) the capability of finding the module having the maximum disease association score; (2) the ability to take the network quality into account; and (3) can be executed in an affordable time. Nonetheless, due to the underlying algorithmic complexity, most existing active module search methods were only able to focus on one or two of these aspects, thus limiting the performance of network-based analysis as a whole.

In this Chapter, I will present a novel efficient and robust active module search method, named SigMod, which has the ability to select a gene module that has the maximum disease-association score and tends to be strongly interconnected. This method was formulated as a binary quadratic optimization problem. We showed that solving this optimization problem is equivalent to finding the min-cut on a graph, hence it can be solved via many available graph-cut algorithms. SigMod has several advantages compared to existing module search algorithms: (1) it can find the module having exactly the maximum disease association score; (2) it allows incorporating edge weights that usually reflect the confidence or strength of connections between genes in the network; (3) its selection path can be computed rapidly hence offers the flexibility for users to select a desirable amount of genes, and (4) the identified gene module tend to have strong connectivity, thus includes genes of close functional relevance.

We first evaluated the performance of SigMod using simulated data. We used the gene network retrieved from the STRING database (introduced in Section 3.3.1 of Chapter I), that

contains various types of gene/protein relationship information including direct (physical) and indirect (functional) interactions. Each edge in the STRING gene network is associated with a weight varying from 0 to 1, which represents the combined confidence of the relationship between two genes derived from multiple sources of information. We chose five strongly interconnected gene modules identified in STRING using CFinder (Adamcsek et al., 2006) as candidate causal modules. The *p*-values of the genes belonging to the causal module were set uniformly distributed between 0 and 10^{-3} (representing signals) whereas *p*-values of other genes were set uniformly distributed between 0 and 1 (representing noise). Experiments performed on these simulated data showed SigMod has the best power and lowest false discovery rate as compared to two state-of-the-art methods—dmGWAS (Jia et al., 2011) and SConES (Azencott et al., 2013). This high performance was preserved when additional noise was intentionally added to both the gene *p*-values and the STRING gene network, demonstrating the robustness of SigMod.

We then applied SigMod to the GABRIEL childhood-onset asthma GWAS outcomes. Using META1 as discovery dataset, we identified a gene module enriched in high association signals and made of 190 genes that are biologically interesting for studying the genetic mechanism underlying asthma. All these genes have a nominally significant *p*-value ($p \leq 0.05$), which shows the ability of SigMod to identify high score genes. When we evaluate this module using META2 dataset, 30 genes were again significant, hence they were significant in both the discovery and replication dataset. Functional clustering analysis using the DAVID tool (Huang et al., 2009) and KEGG pathway enrichment analysis of the genes belonging to the module pinpointed nine functionally closely related gene clusters and 15 enriched pathways. These gene clusters and pathways include biological processes that are known to be related to asthma risk, and also mechanisms that are novel and deserve further investigation of their role in asthma occurrence.

Overall, we proposed an exact and efficient method, named SigMod, for integrative analysis of GWAS data with network-based knowledge. This method enabled to find a relevant gene module enriched with high disease-association signals. It is robust against noise from either the GWAS data or the background network. SigMod can be also applied to any other network-based feature selection problem sharing the same concept.

2 Article published in *Bioinformatics* (doi:10.1093/bioinformatics/btx004)

Bioinformatics, 2017, 1–9
doi: 10.1093/bioinformatics/btx004
Advance Access Publication Date: 9 January 2017
Original Paper

OXFORD

Systems biology

SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network

Yuanlong Liu[1,2,*], Myriam Brossard[1,2], Damian Roqueiro[3], Patricia Margaritte-Jeannin[1,2], Chloé Sarnowski[1,2], Emmanuelle Bouzigon[1,2] and Florence Demenais[1,2]

[1]INSERM, Genetic Variation and Human Diseases Unit, UMR-946, Paris, France, [2]Institut Universitaire d'Hématologie, Université Paris Diderot, Sorbonne Paris Cité, Paris, France and [3]Department of Biosystems Science and Engineering, ETH Zurich, Machine Learning and Computational Biology Lab, Switzerland

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on May 17, 2016; revised on December 19, 2016; editorial decision on January 4, 2017; accepted on January 6, 2017

Abstract

Motivation: Apart from single marker-based tests classically used in genome-wide association studies (GWAS), network-assisted analysis has become a promising approach to identify a set of genes associated with disease. To date, most network-assisted methods aim at finding genes connected in a background network, whatever the density or strength of their connections. This can hamper the findings as sparse connections are non-robust against noise from either the GWAS results or the network resource.

Results: We present SigMod, a novel and efficient method integrating GWAS results and gene network to identify a strongly interconnected gene module enriched in high association signals. Our method is formulated as a binary quadratic optimization problem, which can be solved exactly through graph min-cut algorithms. Compared to existing methods, SigMod has several desirable properties: (i) edge weights quantifying confidence of connections between genes are taken into account, (ii) the selection path can be computed rapidly, (iii) the identified gene module is strongly interconnected, hence includes genes of high functional relevance, and (iv) the method is robust against noise from either the GWAS results or the network resource. We applied SigMod to both simulated and real data. It was found to outperform state-of-the-art network-assisted methods in identifying disease-associated genes. When SigMod was applied to childhood-onset asthma GWAS results, it successfully identified a gene module enriched in consistently high association signals and made of functionally related genes that are biologically relevant for asthma.

Availability and implementation: An R package SigMod is available at: https://github.com/YuanlongLiu/SigMod

Contact: yuanlong.liu@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have achieved considerable success in genetic analysis of complex traits. Thousands of single nucleotide polymorphisms (SNPs) associated with human traits and diseases have been identified since the first GWA study was published (Klein *et al.*, 2005) (http://www.genome.gov/gwastu

dies/). However, the single marker analysis commonly used in GWAS has limitations. Under the very conservative genome-wide significant level of $P = 5 \times 10^{-8}$, only a few of the most significant signals are reported, while many polymorphisms with small marginal effects are missed. The reported SNPs often explain a limited part of the genetic component of a disease or trait (Eichler *et al.*, 2010; Maher, 2008).

To overcome these limitations, a variety of knowledge-based methods have been proposed for integrative and joint analysis of multiple genes. Examples include, but are not limited to the gene set enrichment analysis (GSEA) methods that identify biological pathways enriched in association signals (Subramanian *et al.*, 2005); text-mining methods that build links between genes from scientific literature (Raychaudhuri *et al.*, 2009), etc. Among these approaches stands the network-assisted analyses that overlay gene-level *P*-values onto a gene network (GeneNet) to search for connected genes (also known as gene module) enriched in association signals. The rationale behind this is the principle of 'guilt-by-association', which states that genes (or gene products) connected in a network are usually participating in the same, or related, cellular functions (Li *et al.*, 2015; Oliver, 2000; Wolfe *et al.*, 2005). Although a number of methods have been developed for this purpose (Cabusora *et al.* 2005; Ideker *et al.*, 2002; Jia *et al.*, 2011), they often search modules using heuristic or greedy algorithms, hence cannot guarantee to identify the module enriched in highest signals, and are prone to include biologically irrelevant genes by chance. Also, many of them have the limitation that edge weights are not taken into account during the module searching process, although edge weights represent the confidence or strength of connections between genes and can contain useful information.

Azencott *et al.* (2013) have proposed a module searching method that overcomes some of these limitations. Their method, named SConES, was originally developed for identifying a set of SNPs that are maximally associated with a phenotype and tend to be connected in an underlying network. SConES formulates the module searching task as a binary optimization problem that can be solved exactly and efficiently via graph min-cut algorithms. It also allows incorporating edge weights, making it more robust to false connections. Nevertheless, SConES sets its tuning parameters via a cross-validation strategy that requires using raw genotype data, and therefore cannot be applied to studies in which only summary-level statistics are available, as it is often the case in large genetic consortiums. Also, as indicated in their paper, SConES may select several disconnected subnetworks along with multiple isolated nodes, which may lead to an overall low interconnection among selected nodes. These disconnected subnetworks and especially the isolated nodes, are likely to be less functionally related to the other nodes and the selected module may be less associated with disease as compared to a module whose nodes are strongly connected.

In this article, we propose a novel method SigMod that has the ability to select a Strongly Interconnected Gene MODule maximally associated with the disease. We formulate this module selection task as an optimization problem similar to SConES, but we incorporate a modification in the objective function to explicitly encourage the overall strong interconnection among selected genes. We believe that a set of strongly interconnected genes are more functionally related and biologically relevant. We show that our method has the same advantage as SConES in terms of allowing incorporation of edge weights, and can also be solved exactly and efficiently via graph min-cut algorithms. In addition, we propose an algorithm to compute the module selection path, which provides the ability to trace the selection change and to select a desirable amount of genes. We also develop a parameter setting strategy to identify the optimal selection. Our strategy does not require using raw genotype data, hence can be applied to a broader range of studies than SConES. We evaluated SigMod using both simulated and real data, and made comparisons with SConES and another popular network-based method dmGWAS (Jia *et al.*, 2011). The results showed our method is more powerful in identifying a module made of functionally relevant genes and enriched in consistent association signals.

2 Methods

SigMod aims to identify a disease-associated gene module using two types of input data: a list of gene-level *P*-values obtained from GWAS SNP-level *P*-values, and a GeneNet. To get gene-level *P*-values, SNPs need to be first assigned to genes using dbSNP and RefSeq genes with genomic coordinates in the corresponding genome build, but methods vary according to the choice of gene boundaries that can be strictly limited to the start and stop positions of the genes, or extended beyond these positions up to 500 kb. This SNP to gene assignment issue has been previously debated in Jia and Zhao (2014) and will be further discussed in Section 5. Once SNPs have been assigned to genes, gene-level *P*-values, which represent the significance of gene–disease associations, are computed from GWAS SNP-level *P*-values using any gene-based method that has been previously proposed (e.g., Liu *et al.*, 2010; Lamparter *et al.*, 2016; Li *et al.*, 2011). One of the most popular gene-based methods consists of using the best SNP *P*-value assigned to a gene but this *P*-value needs to be corrected for variation in gene length (as explained in Section 4.2). The GeneNet represents the biological knowledge of gene-gene relationships, such as physical interactions between gene products (proteins), gene co-expression or co-occurrence of gene-related terms in the literature. Each connection can have a weight that measures the confidence or strength of the connection. This type of information can be derived from experiments like co-expression analysis or retrieved from databases such as STRING (Szklarczyk *et al.*, 2014).

In the following sections, we will first introduce the formulation of SigMod, and then provide an efficient and exact algorithm to solve the optimization problem. Afterwards we will present a tuning parameter setting strategy to find the parameters leading to an optimal gene module selection. A flowchart summarizing these steps is shown in Figure 1.

2.1 Formulation of the SigMod method

We first transform gene-level *P*-values into scores by $z = \Phi^{-1}(1 - P)$, where $\Phi^{-1}(\cdot)$ is the inverse normal distribution function. These gene scores are overlaid onto the GeneNet to build a scored GeneNet, denoted as $G = (V, A)$, where V are nodes representing genes, and A is the weighted adjacency matrix representing connections among genes. We define $\boldsymbol{u}$ as a vector of binary variables indicating whether a gene V_p is selected ($\boldsymbol{u}_p = 1$) or not ($\boldsymbol{u}_p = 0$). We formulate this selection task as an optimization problem that maximizes the following objective function:

$$f(\boldsymbol{u}) = \boldsymbol{z}^T\boldsymbol{u} + \lambda \boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} - \eta||\boldsymbol{u}||_0. \quad (1)$$

The first component $\boldsymbol{z}^T\boldsymbol{u}$ defines the joint effect of the gene module on the phenotype (disease) by summing up the scores of its gene members. The second component $\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u}$ quantifies its connection strength as the summed edge weights in the module, since $\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} = \sum_{p,q}\boldsymbol{A}_{pq}\boldsymbol{u}_p\boldsymbol{u}_q$. The third component is the sparsity regularizer controlling the size of the gene module, where the module size is represented by $||\boldsymbol{u}||_0$, i.e., the number of non-zero elements in

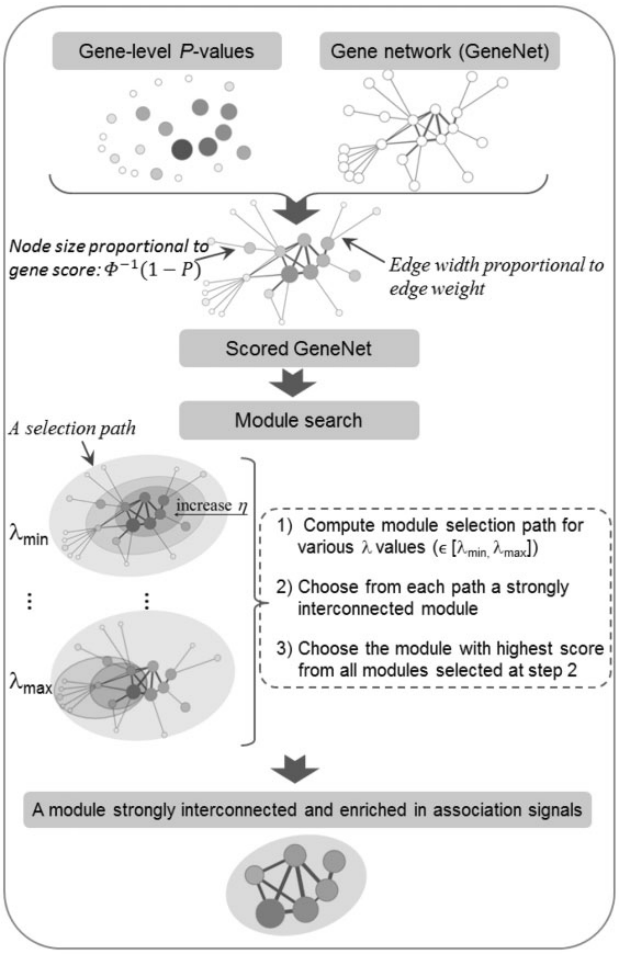


Fig. 1. Workflow of SigMod. SigMod takes a list of gene-level *P*-values computed from genome-wide association studies (GWAS) and a gene network (GeneNet) as input. The gene-level *P*-values are converted into scores and overlaid onto the GeneNet to build a scored network. SigMod identifies a module that is strongly interconnected and enriched in high association signals from this network using a 3-step procedure, as outlined in this figure and detailed in the text (Section 2.3.2). The λ in this figure is the connectivity parameter that controls the balance between module score and module connectivity. The selection path in the figure represents the sequence of distinct modules selected by increasing the sparsity parameter η from a starting value to $+\infty$, as described in Section 2.3.2

$\boldsymbol{u}$. λ and η are positive tuning parameters specifying the importance of the corresponding components. Therefore, we are able to select a strongly interconnected gene module enriched in high association signals, by choosing proper parameters and solving the optimization problem:

$$\arg\max_{\boldsymbol{u}} f(\boldsymbol{u}). \tag{2}$$

Note that our formulation differs from the formulation of SConES (Azencott *et al.*, 2013). SConES selects genes by maximizing the objective function defined as $g(\boldsymbol{u}) = \boldsymbol{z}^T\boldsymbol{u} - \lambda\boldsymbol{u}^T\boldsymbol{L}\boldsymbol{u} - \eta||\boldsymbol{u}||_0$, where $\boldsymbol{L}$ is the Laplacian matrix defined as $L = D - \boldsymbol{A}$, and $\boldsymbol{D}$ is the diagonal matrix of weighted node degrees, i.e. $\boldsymbol{D}_{pp} = \boldsymbol{d}_p := \sum_q \boldsymbol{A}_{pq}$. The difference between the two objective functions $f(\boldsymbol{u})$ and $g(\boldsymbol{u})$ is in the second component, which leads to different behaviors of each method. Specifically, SConES incorporates the Laplacian matrix to encourage adjacent nodes to be selected together. However, this does not guarantee the overall strong interconnection among selected nodes. Contrariwise, the SigMod formulation incorporates the adjacency matrix to explicitly encourage selection of strongly interconnected nodes. More specifically, since $\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{L}$, it has

$$\begin{aligned} f(\boldsymbol{u}) &= \boldsymbol{z}^T\boldsymbol{u} + \lambda\boldsymbol{u}^T(\boldsymbol{D} - \boldsymbol{L})\boldsymbol{u} - \eta||\boldsymbol{u}||_0 \\ &= (\boldsymbol{z} + \lambda\boldsymbol{d})^T\boldsymbol{u} - \lambda\boldsymbol{u}^T\boldsymbol{L}\boldsymbol{u} - \eta||\boldsymbol{u}||_0. \\ &= g(\boldsymbol{u}) + \lambda\boldsymbol{d}^T\boldsymbol{u} \end{aligned}$$

Therefore for each given λ, additional scores $\lambda\boldsymbol{d}$ are added to the nodes in SigMod compared to SConES. Nodes with higher degrees are thereby given more preferences. Additional differences between these two methods will be presented in the following sections.

2.2 Optimization algorithm

We show that the optimization of Equation (2) can be solved exactly using a similar graph min-cut approach as presented in Azencott *et al.* (2013). To achieve this, we construct an augmented network of G (denoted as G_{st}), by first adding two artificial nodes s and t to G, then redefining its adjacency matrix as $\boldsymbol{B}$:

$$\begin{cases} \boldsymbol{B}_{pq} = \lambda\boldsymbol{A}_{pq} \\ \boldsymbol{B}_{sp} = (z_p + \lambda\boldsymbol{d}_p - \eta) \times I(z_p + \lambda\boldsymbol{d}_p \geq \eta) \quad \text{for } 1 \leq p, q \leq n. \\ \boldsymbol{B}_{tp} = (\eta - z_p - \lambda\boldsymbol{d}_p) \times I(z_p + \lambda\boldsymbol{d}_p < \eta) \end{cases} \tag{3}$$

Definition 1. *Given a network $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{A})$, for any $s, t \in \boldsymbol{V}$, a s-t cut $C = \{\boldsymbol{X}, \bar{\boldsymbol{X}}\}$ is defined as a node partition of $\boldsymbol{V}$ such that: (1)$\boldsymbol{X} \cup \bar{\boldsymbol{X}} = \boldsymbol{V}$; (2) $s \in \boldsymbol{X}$ and $t \in \bar{\boldsymbol{X}}$.*

Definition 2. *A s-t cut is called a s-t min-cut if $\kappa(C)$ is minimized, where $\kappa(C)$ is the capacity of a s-t cut C, defined as $\kappa(C) = \sum_{V_p \in \boldsymbol{X}, V_q \in \bar{\boldsymbol{X}}} \boldsymbol{A}_{pq}$.*

Therefore according to Proposition (1) in Azencott *et al.* (2013), if $C^* = (\boldsymbol{X}^*, \overline{\boldsymbol{X}^*})$ is a *s-t* min-cut of G_{st}, then $\boldsymbol{u}^*$ is the solution of the optimization problem of Equation (2), where $\boldsymbol{u}^*_p = 1$ if $\boldsymbol{V}_p \in \boldsymbol{X}^*$, and $\boldsymbol{u}^*_p = 0$ otherwise. Hence solving the optimization problem is equivalent to finding a *s-t* min-cut on the augmented network G_{st}. Thus any *s-t* min-cut algorithm can be applied to find the solution.

2.3 Determination of the tuning parameters η and λ

The SigMod objective function Equation (1) includes two tuning parameters, η and λ, that need to be determined. To find the parameter values leading to an optimal gene module selection, we first propose a path algorithm that allows computing all distinct selections at a given λ while varying η over a range of values. Based on this algorithm, we provide a procedure to find the tuning parameters that can lead to the optimal gene module selection. These different steps are described as follows.

2.3.1 Computing the selection path at any given λ value

For a given value of λ, the module selection by solving Equation (2) has the *nesting property* that $S(\eta_1) \subseteq S(\eta_0)$ if $\eta_1 > \eta_0$, where $S(\eta)$ represents the module selected by setting the sparsity parameter as η (see Supplementary Materials for proof). Therefore increasing η results in removing genes from a previously selected module. To conveniently trace this selection change, we develop the *path algorithm* that allows computing the sequence of distinct modules selected by increasing η from $\eta_{\min}$ to $\eta_{\max}(0 \leq \eta_{\min} < \eta_{\max})$. We denote this sequence as $\mathcal{P} = \langle S(\eta_{\min}), \ldots, S(\eta_{\max})\rangle$ and call it as the selection path over $[\eta_{\min}, \eta_{\max}]$. Note that these modules are nested according to the nesting property, i.e., $S(\eta_{\min}) \supseteq \cdots \supseteq S(\eta_{\max})$. An example of selection path is given in Supplementary Figure S1.

Our path algorithm aims to compute $\mathcal{P}$ efficiently. It is developed by exploring the property of *s-t* min-cut on the augmented graph G_{st}, since computing $S(\eta)$ is equivalent to finding the *s-t* min-cut as stated in Section 2.2. We define the capacity function $\kappa^*(\eta)$ as the capacity of the *s-t* min-cut on G_{st}, where the capacity of a cut is defined in Definition 2 (Section 2.2). It is apparent that $\kappa^*(\eta)$ is a continuous and piecewise linear function of η. Its slope changes at either a break-point or a change-point, where a value of η is a break-point if it leads to the change of selection, and is a change-point if it causes the rewiring of an edge of G_{st} from s to t according to Equation (3). Thus, computing the selection path is equivalent to finding all break-points of $\kappa^*(\eta)$, which can be achieved by correcting $\kappa^*(\eta)$ at each change-point to transform it to a concave function, then applying the iterative contraction algorithm described in Gallo *et al.* (1989). Once all break-points are obtained, the selection path can be computed by setting η at each of the break-points and solving the problem defined by Equation (2). A detailed description of this algorithm is presented in Supplementary Materials.

We also notice that the module selection by solving Equation (2) has the *memoryless property*, that if a gene is not selected by setting η at some value (e.g., $\eta = \eta_{\min}$), then it can be removed from the GeneNet when computing the selection at a η value greater than $\eta_{\min}$. The mathematical description of this property and its proof is given in Supplementary Materials (Proposition 1). This property can be utilized to speed up the computation of selection path over $[\eta_{\min}, \eta_{\max}]$, using the following two-step procedure:

- Step 1: compute $S(\eta_{\min})$ on the complete network G;
- Step 2: compute selection path over $[\eta_{\min}, \eta_{\max}]$ on the subnetwork G_{sub} induced by the genes in $S(\eta_{\min})$.

This speed-up strategy makes the computation of selection path more efficient, especially when the size of $S(\eta_{\min})$ is far less than the total amount of genes in the whole network. It is the case for many studies in which only a small portion of genes are intended to be selected while the majority of genes are left out at the first stage of the selection process.

2.3.2 Hierarchical procedure to find the tuning parameters leading to an optimal gene module selection

As mentioned above, the module selection in SigMod depends on two parameters η and λ. The selection as a function of η can be tracked through the selection path at any given value of λ. The parameter λ, which allows a balance between the module score and module connectivity, needs to be chosen carefully. On one hand, if λ is too small, the selection mainly focuses on gene scores while it ignores the connections among genes. This results in the top scored genes scattered in the network to be selected, whichever their connections. On the other hand, if λ is too big, the network topology dominates the selection, while the gene scores do not influence the module selection. This leads to a set of most strongly interconnected genes to be selected, whichever their association scores. Since the goal of our method is to find a gene module that is strongly interconnected and is enriched in high association signals, we propose the following procedure to set the parameters properly:

- Step 1: do an exhaustive search for k equally spaced λ values in $[\lambda_{\min}, \lambda_{\max}]$. Compute for each λ the selection path $\mathcal{P}(\lambda)$ to collect all modules with module size less than *max_select*;
- Step 2: compute the size difference between consecutive modules in each path $\mathcal{P}(\lambda)$, i.e. $\Delta s_i = |S_i| - |S_{i+1}|$, where S_i is the i^{th} module in the path. Then choose S_{i^*} within each path, where $i^* = \max\{i | \Delta s_i \geq \tau\}$;
- Step 3: remove genes not connected to others in each S_{i^*}. Choose from all resulting S_{i^*} the one with highest standardized score as final selection, denoted as S^*.

In Step 1, we explore the module selection for k different λ values. For each value, we calculate its selection path $\mathcal{P}(\lambda)$ to collect all distinct modules whose number of genes is less than *max_select* (specified by the user). This can be achieved by starting at a trial value $\eta = \eta_0$ and computing the path over the sparsity range $[\eta_0, \infty]$. If $|S(\eta_0)| < max\_select$, decrease η and compute the path in the extended range, until the size of the largest selected module surpasses *max_select*. The range $[\lambda_{\min}, \lambda_{\max}]$ should be broad enough, so that an optimal selection is contained in these paths. Though exhaustive search is potentially expensive, the incorporation of our speed-up path algorithm can largely reduce the computational burden.

In Steps 2, the goal is to find a *local optimum* module within each path, where by *local optimum* we mean the selected module is strongly interconnected and enriched in high scores relative to that path. We identify this *local optimum* by examining the size difference between consecutive modules in $\mathcal{P}$, i.e., $|S_1| - |S_2|$, $|S_2| - |S_3|$, etc. This is because, by our formulation, if the connectivity regularizer does not have an effect, the genes will be removed one by one from the module; while if the regularizer has an effect, some strongly interconnected genes are non-separable and are removed together, which corresponds to a large size jump (τ) between consecutive selections in the selection path, as shown in Supplementary Figure S2. We select the smallest module in the path that contains such non-separable genes (by choosing $i^* = \max\{i | \Delta s_i \geq \tau\}$). We set $\tau = 5$ by default, but it can be adjusted based on actual situation.

In the final step, we first remove the genes that are not connected to any other gene in each *local optimum* module. Then we choose from these local optima the one with highest standardized score, where the standardized score of a module S is defined as

$$z^*(S) = \frac{z(S) - |S| \times \widehat{\mu}}{\sqrt{|S|}\widehat{\sigma}}.$$

Here $z(S) = \sum_{s \in S} z_s$. $\widehat{\mu}$ and $\widehat{\sigma}$ are the sample mean and standard deviation of all gene scores in GeneNet.

A summary of this procedure is shown in Figure 1. Through this hierarchical procedure we increase the possibility to find the true disease-associated gene module.

3 Implementation

We implemented our method in an R package *SigMod* (available at https://github.com/YuanlongLiu/SigMod). SigMod takes a list of gene-level *P*-values and a GeneNet as input. Each connection in the GeneNet can be assigned a weight to quantify the confidence or strength of the connection. When the weight of a connection is unavailable, it can be specified as 1 or 0 to indicate presence or absence of the connection.

The *SigMod* package consists of the main function *select_subnet* to solve the optimization problem of Equation (2); the *selection_path* function to calculate the selection path as described in Section 2.3.1; and additional functions to help identify the optimal module selection. We use the *graph.maxflow* function in R package *igraph 0.7.1* (Csardi and Nepusz, 2006) to find the *s-t* min-cut. It implements the Goldberg-Tarjan Push-Relabel algorithm (Goldberg and Tarjan, 1988), and has the smallest known time complexity of $\mathcal{O}(n_1 n_2 \log(n_1^2/n_2))$, where n_1 is the number of genes in GeneNet and n_2 is the number of connections.

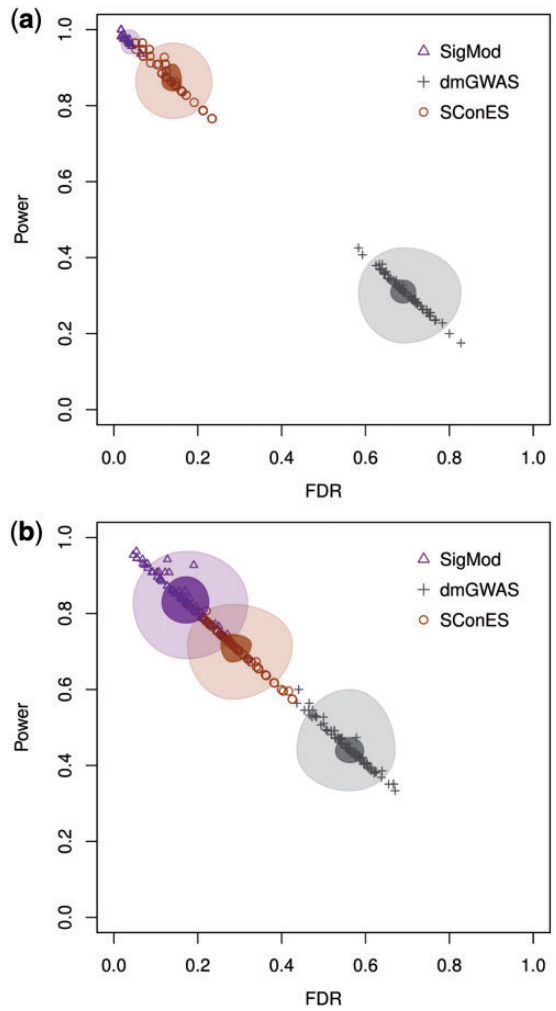


Fig. 2. False discovery rate (FDR) versus power of three network analysis methods applied to simulated data. The results of 20 replicates of five causal modules are aggregated. Five-number statistics (minimum, first quartile, median, third quartile, and maximum) of each quantity are shown by ellipse plot (Tomizono, 2013). Plot (**a**) shows the results without adding noise to the GWAS data or GeneNet. Plot (**b**) shows the results with noise added to both GWAS data and GeneNet

4 Results

We evaluated the performance of SigMod using both simulated and real datasets. We downloaded a comprehensive human GeneNet from the STRING database version 10 (Szklarczyk *et al.*, 2014), which contains information on various types of connections among genes. This GeneNet includes 19 247 genes and 4 274 001 edges. Each edge represents a known or predicted interaction between genes or gene products (proteins), including direct (physical) and indirect (functional) associations derived from four sources including systematic genome comparisons, high-throughput experiments, co-expression and previous knowledge from literature. Each edge in the STRING GeneNet is assigned a weight varying from 0 to 1, which represents the combined confidence of the connection between two genes derived from different sources of information.

4.1 Results of the simulated data

We first conducted simulations using the STRING GeneNet. We chose five strongly interconnected gene modules identified by CFinder (Adamcsek *et al.*, 2006) as candidate causal modules (Supplementary Fig. S3). The sizes of these modules ranged from 47 to 87.

In each simulation, a single module was set as the causal module. We followed the proposal of Rajagopalan and Agarwal (2005) to set *P*-values of the genes belonging to the causal module to be uniformly distributed between 0 and 10^{-3}. *P*-values of other genes were uniformly distributed between 0 and 1. We set $[\lambda_{\min}, \lambda_{\max}] = [0.005, 0.05]$ and computed selection paths for $k = 100$ values of λ in this range. Other parameters were set as $\tau = 5$ and $max\_select = 1000$.

We compared our method with two state-of-the-art module search methods dmGWAS (Jia *et al.*, 2011) and SConES (Azencott *et al.*, 2013). The dmGWAS method identifies gene modules by starting from each gene in the GeneNet and repeatedly adding neighboring genes that generate the maximum increment of the module score ($z(S) = \sum_{s \in S} z_s$). Module growth terminates if adding neighboring genes does not yield more than r% ($r = 10$ by default) increment of the score. As in dmGWAS the number of genes to be selected is determined by the user, we selected approximately the same number of genes as that of the causal module under study. To do so, we first set parameters to their default values to generate raw modules. Then we ordered the raw modules according to their module scores. Top modules were selected sequentially until the cumulative size of these modules exceeded that of the causal module. The SConES method, as described in Section 2.1, selects genes by maximizing the objective function $g(\boldsymbol{u})$. It should be noticed that its original implementation uses a cross-validation approach to set tuning parameters, which does not apply to our study as raw genotype data are not used. Nonetheless, according to the relationship between $f(\boldsymbol{u})$ and $g(\boldsymbol{u})$ described in Section 2.1, it is straightforward that our path algorithm can also be applied to SConES. Thereby, we computed its selection paths using the same λs as for SigMod. In each path, we chose the first module selection whose size exceeded that of the causal module. Among these selections we chose the one with largest standardized score.

We ran 20 repetitions for each of the five candidate gene modules (hence 20×5 experiments for each method). We computed the power (fraction of causal module genes selected) and false discovery rate (FDR, fraction of selected genes that are not causal) of each experiment. SigMod has systematically higher power and lower FDR over all experiments, as presented in Figure 2 (results are aggregated for all experiments; see Supplementary Fig. S4 for individual results). SConES has lower power and higher FDR than SigMod while dmGWAS performs worst in these simulations. We further compared the standardized connection strength of the selected modules, defined as $\rho = 2\omega/m(m-1)$, where m is the module size; ω is the sum of pairwise edge weights in the module. As shown in Figure 3 and Supplementary Figure S5, the connection strengths of gene modules selected by SigMod are much higher than the other two methods.

The performance of these methods against noise was also evaluated. Two sources of noise were considered simultaneously. The first one is standard Gaussian noise added to the scores of the causal module genes. The second noise is added to the topology of GeneNet by randomly rewiring 5% of the edges, where at a rewire step, two edges $V_1 \sim V_2$, $V_3 \sim V_4$ becomes $V_1 \sim V_4$, $V_3 \sim V_2$. This rewire process keeps the distribution of node degree unchanged. We observed that SigMod still has the best performance among the three methods, with an average power of 0.83 and FDR of 0.18 (Fig. 2 and Supplementary Fig. S4). Interestingly, dmGWAS has an improved performance when noise is added (higher power and lower FDR). This is because it selects genes with highest scores.

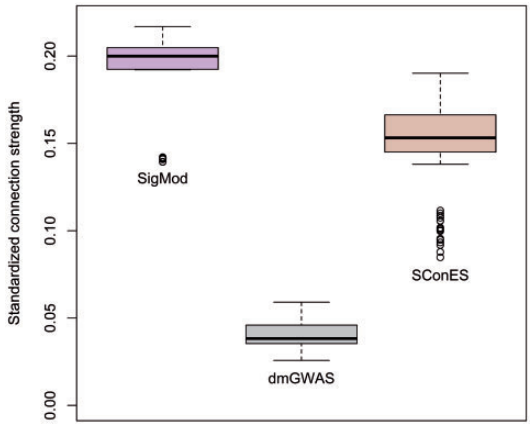


Fig. 3. Box plots of the standardized connection strength (ρ) of gene modules identified by three network analysis methods (SigMod, SConES and dmGWAS). The results of 20 replicates of five causal modules were aggregated

By adding Gaussian noise to the scores, genes with increased score are more likely to be selected.

4.2 Identification of a gene module associated with childhood-onset asthma

We applied SigMod to the meta-analysis results of 18 childhood-onset asthma GWASs, which were part of the European GABRIEL asthma consortium (Moffatt *et al.*, 2010). The data are described in detail elsewhere (Moffatt *et al.*, 2010). In order to check the consistency of results of SigMod, we used a discovery-evaluation scheme. Therefore, the 18 childhood-onset asthma GWASs were randomly split into two groups of nine GWASs while preserving a similar sample size for the two groups: 3031 cases/2893 controls in the first group for discovery and 2679 cases/3364 controls in the second group for evaluation. In each group, a meta-analysis was applied to 2 370 689 single SNP association statistics (Hapmap2-imputed SNPs after quality control), using a random-effects model implemented in STATA V12 (distributed by Stata Corporation, College Station, Texas, USA). The results of these two meta-analyses (SNP-level *P*-values) were respectively named META1 and META2.

To aggregate SNP-level results into genes, SNPs were mapped to genes (between the start site and 3'-untranslated region of each gene) using dbSNP Build 132 and human Genome Build 37.1, making a total of 24 120 genes with at least one SNP mapped. Each gene-level *P*-value was taken as the best SNP *P*-value among all SNPs mapped to the gene, and was further corrected for gene length using permutations. We applied the circular genomic permutation (CGP) approach that can preserve linkage disequilibrium (LD) among SNPs when permuting SNP-level statistics (Cabrera *et al.*, 2012). It was shown to have similar performance to the highly time-consuming gold standard of phenotype permutation (Brossard *et al.*, 2013). These corrected gene-level *P*-values were converted to scores by inverse normal transformation. The scores were mapped to the STRING-based GeneNet to build a scored network, which consisted of 15 724 genes and 3 055 850 edges.

We applied SigMod to the META1 discovery set. We used the same parameter settings as described in simulations, i.e. $\lambda_{\min} = 0.005$, $\lambda_{\max} = 0.05$, $k = 100$, $\tau = 5$ and $max\_select = 1000$. We identified a strongly interconnected gene module of 190 genes and 1295 connections (Supplementary Fig. S6).

4.2.1 Enrichment of the identified gene module in high association signals

The selected gene module has a standardized score of 36.09, which is significantly higher than the scores of 100,000 random modules (each has the same number of genes as in the identified module) sampled from the scored GeneNet ($P < 10^{-5}$; Supplementary Fig. S7). All module genes have significant *P*-values ($P \leq 0.05$), ranging from 5.48×10^{-6} to 1.88×10^{-2}. These *P*-values are ranked at the top of the whole gene list, with highest rank of 1 and lowest rank of 581 (Supplementary Table S1).

We then evaluated whether the selected gene module was enriched in consistent association signals, by computing its score using META2 dataset. The gene module had a standardized score of 5.85, which was again significantly higher than scores of 100 000 randomly generated modules ($P < 10^{-5}$; Supplementary Fig. S7). This shows the ability of SigMod to select a module displaying consistent association signals.

4.2.2 Association of the identified gene module with asthma

The association of the identified module with childhood-onset asthma was evaluated through CGP permutation of SNP *P*-values that can preserve the genomic structure, using META1 and META2 respectively. For each evaluation, a total of 100 000 CGP samples were generated and scores of the identified gene module were recomputed using these samples. The observed score of the identified module was significantly higher than those obtained from the permutation samples ($P < 10^{-5}$ evaluated using either META1 or META2) (Supplementary Fig. S8). This shows the gene module is significantly associated with childhood-onset asthma.

4.2.3 Functional clustering and annotations of genes belonging to the identified gene module

Our method is based on the 'guilt by association' principle. To explore the functional relatedness of genes belonging to the selected module, we used the gene functional classification tool of the DAVID Bioinformatics Resource (Huang *et al.*, 2009). This tool generates a gene-to-gene similarity matrix based on shared functional annotation profiles using over 75 000 terms from 14 annotation sources and classifies highly related genes into functionally related groups. We identified nine functional gene clusters of which seven included genes having strong connections within our selected module (Fig. 4 for these seven groups and Supplementary Figure S9 for the additional two groups). Altogether the nine functionally related groups included 68 out of the 190 module genes (36%). The function of each gene cluster was annotated by the most representative gene ontology (GO) category shared by all genes within a cluster and with highest (or close to highest) enrichment in these genes. For the seven clusters with strong gene–gene connections, these GO categories corresponded to the MHC protein complex, known to be associated with many immune-related diseases including asthma, and potentially novel mechanisms such as nucleosome assembly, regulation of ubiquitin-protein ligase activity, protein catabolic process, zinc ion binding, as well as regulation of transcription (clusters 6 and 7) which plays a key role in autoimmune diseases (Farh *et al.*, 2015) that share susceptibility loci with asthma (Welter *et al.*, 2014).

Finally, we performed KEGG pathway enrichment analysis to further annotate the module genes. We used the enrichKEGG function of the R package clusterProfiler (Yu *et al.*, 2012), which interrogates KEGG on the fly to get the latest pathway information. We found 15 pathways (Table S2) significantly enriched in genes from

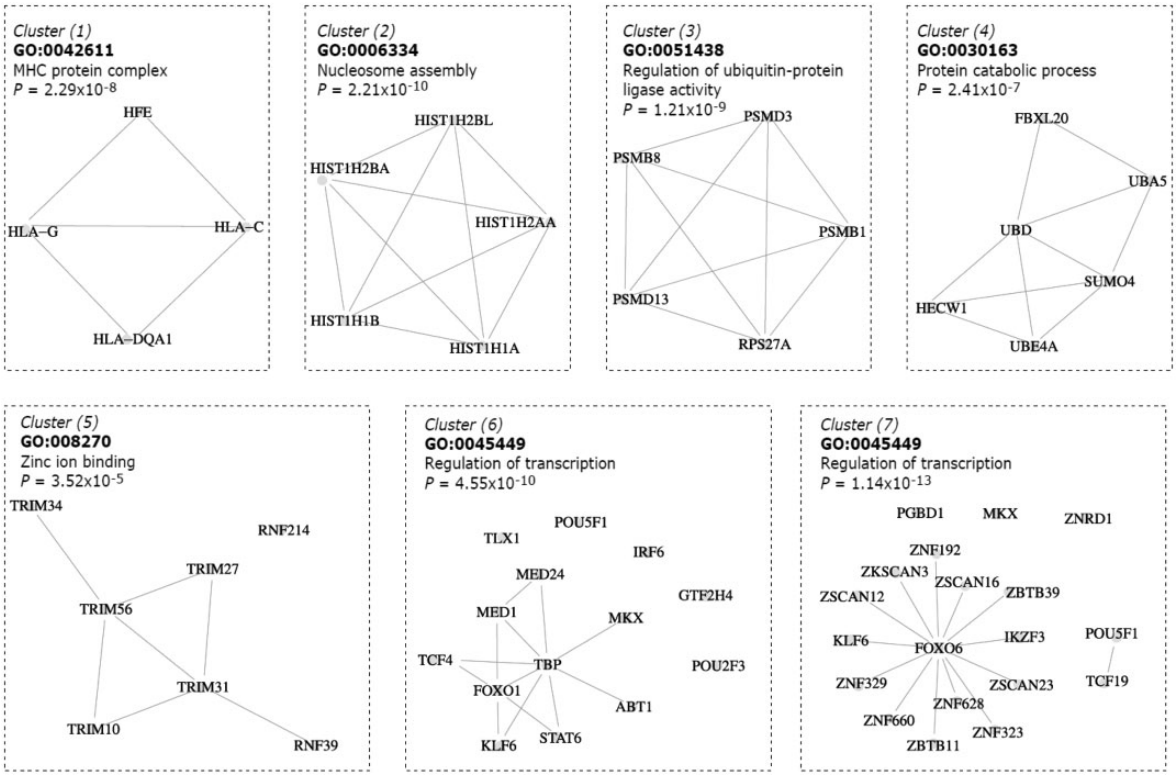


Fig. 4. Seven strongly interconnected functional gene clusters identified by DAVID in the selected gene module associated with childhood-onset asthma. The main function of each cluster is represented by the gene ontology (GO) category that has highest enrichment in the cluster genes. The *P*-values correspond to the significance of enrichment of the shown GO term in the corresponding gene cluster

the identified module (FDR < 0.05). Of particular interest is that five KEGG pathways are related to virus infection, which supports previous findings of the modulating effect of genetic variants associated with asthma at the 17q21 locus on the association of asthma with viral infections (Çalışkan *et al.*, 2013; Smit *et al.*, 2010). Moreover, the antigen presentation pathway was already identified by DAVID as the MHC complex GO, and the Inflammatory Bowel Disease and Type 1 Diabetes pathways represent two auto-immune diseases that share susceptibility loci with asthma (Welter *et al.*, 2014). All of this adds further evidence that the selected gene module includes genes of functional relevance for asthma.

4.2.4 Comparison of results using SigMod, dmGWAS and SConES

For purpose of comparison, we also applied dmGWAS and SConES to the META1 dataset to identify modules. We used the same strategy as described in the simulation study to select approximately the same number of genes as selected by SigMod. We compared the identified modules for their enrichment of association signals (quantified by the module score z^*), and evaluated the replicability of these signals in the independent META2 dataset.

As shown in Table 1, the module identified by SConES has a slightly lower score than the module selected by SigMod. All genes of SigMod and SConES modules have a significant *P*-value ($P \leq 0.05$), hence are likely to be bona fide genes. Comparatively, the module identified by dmGWAS has a score that is twice as small as the SigMod module score. Also only half of its module genes have a significant *P*-value. This shows dmGWAS has a lower ability to identify genes having strong association signals. This is likely because: dmGWAS uses a heuristic search algorithm that does not guarantee the maximization of the module score; while SigMod and SConES use exact algorithms to ensure the maximization.

Table 1. Comparison of the performance of SigMod with dmGWAS and SConES in identifying a gene module associated with childhood-onset asthma

				META1		META2	
	#Genes	#Edges	ρ	z^*	#sig	z^*	#sig
SigMod	190	1295	0.022	36.08	190	5.85	30
SConES	190	232	0.004	35.52	190	4.14	18
dmGWAS	191	679	0.011	17.18	92	3.65	25

Various features of the identified gene module were compared, including the number of genes and edges, the connection strength (ρ), the standardized module score (z^*), and the number of nominally significant genes (#sig) in the module. META1 and META2 are the two datasets consisting of SNP-level *P*-values obtained from meta-analyses of childhood-asthma GWAS.

When these modules were evaluated for replication of results using the independent META2 dataset, the module identified by SigMod again had the highest score (see Table 1). Specifically, 30 genes out of the 190 genes were significant when evaluated from META2 (Supplementary Table S1), hence were significant in both META1 and META2 and are thus of biological interest. These module genes account for almost half of the 70 genes in the GeneNet that are significant in both datasets, demonstrating the ability of SigMod to identify genes displaying consistently high association signals. Comparatively, the signals in the module identified by dmGWAS or by SConES were less replicated, as indicated in Table 1 by the module score and the number of significant genes evaluated using META2. Specifically, 18 out of 190 genes identified by

SConES from META1 remained significant in META2. This lower replication rate (60% of the SigMod replication rate) may be due to the lower overall interconnection among genes selected by SConES. As shown in Table 1, the number and strength of connections between genes in the SConES module are both 18% of the values observed in the SigMod module. These genes with lower overall connection strength are likely to be less functionally relevant, and to have a less consistent joint effect on disease.

As for computational efficiency, all three methods (with SConES using our tuning parameter setting strategy) have comparable run time of ~3 h on a server (2.66 GHz Intel® Xeon® Processor X5650 and 160 GB of RAM).

5 Discussion and conclusion

Network-assisted analysis of GWAS data to identify gene modules enriched in high association signals has received increasing attention over the last decade. In this article, we proposed a novel method SigMod, tailored for such purpose. SigMod takes a gene network and a list of gene-level *P*-values as input. The gene network can be retrieved from databases or derived from experiments that are best suitable to the study. In our application to the asthma data, we chose the STRING network that has the advantage of integrating connection information from various sources. The gene-level *P*-values, which represent the significance of gene–disease association, can be computed from GWAS SNP-level *P*-values using any proper gene-based methods (e.g., Lamparter *et al*., 2016; Liu *et al*., 2010). In our study, gene-level *P*-values were chosen as the best SNP *P*-value in a gene and were corrected for gene length using Circular Genomic Permutation that can preserve the LD pattern between SNPs. One challenge in network-assisted analysis is the assignment of SNPs to genes, as discussed in Jia and Zhao's review (Jia and Zhao, 2014). In our study, we used a stringent definition of gene boundaries, which were represented by the start site and 3'-untranslated region of each gene to reduce false positives. Although gene boundaries can be extended to a few kilobases both upstream and downstream of a gene, it was shown that a change of boundaries from 0 to 250 kb did not significantly affect the power of the related network analysis (Lee *et al*, 2011), although this needs to be further confirmed. Moreover, extension of boundaries to flanking regions of a gene may increase the degree of overlap of nearby genes and thus the number of wrong SNP-to-gene assignments. More sophisticated SNP to gene annotation strategies that take into account functional information, such as gene expression through expression quantitative trait loci (eQTLs), or that define a regulatory domain for each gene (McLean *et al*, 2010), may be considered. However, the performance of such annotation strategies with respect to the classical ones need to be further assessed.

SigMod selects a strongly interconnected gene module enriched in association signals by optimizing a binary quadratic objective function. We showed the optimization problem can be solved exactly through graph min-cut algorithms. We also designed a path algorithm that allows computing the selection path at any given λ value. This provides the flexibility to select an appropriate number of genes. In combination with the path algorithm, we proposed a strategy that enables choosing proper parameters to keep a balance between module score and module connectivity. This strategy does not require using raw genotype data. We believe that a proper parameter setting strategy is as important as the formulation of the objective function, as inappropriate parameters can lead to unwanted results, especially for network-assisted analysis where numerous gene modules can be selected. Comparatively, in the original SConES method the parameters are determined using a cross-validation approach, which cannot be applied to situations where raw genotype data are unavailable, as often encountered.

In comparison to previous approaches that only require the selected genes being connected in a network, SigMod encourages selecting genes having overall strong interconnection. This emphasis is well grounded as the identified module is more robust against noise. In particular, genes that have some false connections in the selected module may still be kept in the module after removing such connections, whereas for a loosely interconnected module, removal of false connections may destroy the module structure. Also, a strong interconnection among genes can reflect close functional relationships, as implied by the 'guilt by association' principle and demonstrated by our application to the asthma dataset.

SigMod has a different focus compared with SConES. Specifically, SConES focuses on co-selection of adjacent nodes rather than the overall strong interconnection among selected nodes. The node preference between SigMod and SConES is also different. SConES favors low degree nodes while SigMod rewards nodes of higher degrees, as indicated in Section 2.1. We believe that rewarding high degree nodes is particularly suitable for some applications. It has been widely observed that many disease-causing genes have high degrees in a gene network, especially those playing a central role in complex diseases (Lee *et al*. 2013; Xiong *et al*., 2014). These genes can even show higher connectivity in an integrated gene network (e.g. STRING) that aggregates connection information from various sources. Although SigMod rewards genes of higher degree, the scale of rewarding is controlled by a tuning parameter λ. This parameter keeps the balance between the module score and the connectivity, which can be chosen properly using our parameter setting strategy. The validity of this strategy was verified in the simulation study and in the application to asthma GWAS data, where all selected genes were nominally significant (after correction for gene length) and were ranked at the top of the gene list in the whole network (Supplementary Table S1). We did not observe any gene was selected just because it is a hub gene even when it had a very low score.

In our simulations, we found SigMod outperforms SConES and another state-of-the-art method dmGWAS. It has the best power and lowest false discovery rate. This high performance was preserved in presence of noise from both GWAS results and network information, demonstrating its robustness. Further application of SigMod to childhood-onset asthma GWAS results successfully identified a gene module significantly associated with disease. The analyses of functional relationships among genes highlighted known asthma-related gene functions and novel ones which allow generating new hypotheses regarding the mechanisms underlying asthma pathogenesis. Though the module identified by SConES was also enriched in high association signals in the META1 discovery dataset, these signals were less well replicated in the independent META2 dataset. A possible explanation is that the genes in the SConES module are less connected than those identified by SigMod, as reflected by the overall connection measure (ρ). They are thus likely to be less functionally related and may have a less consistent joint effect on disease. This emphasizes again the importance of favouring strong interconnection as achieved by SigMod.

To our knowledge, our method is one of the very few methods in related work that both take edge weights into account and can be solved using exact algorithms. As there are emerging approaches to define connections among genes (e.g., physical or functional, experiment verified or computational based interaction), edge weights are

an important indicator of the confidence or strength of the connection. For those methods that do not incorporate edge weight, an arbitrary hard cutoff has to be given to define the presence or absence of a connection, which can lose useful information.

Our current formulation of SigMod did not take into account the LD pattern that may exist among SNPs belonging to adjacent genes or gene clusters in a chromosomal region that may share similar functions. This may cause over selection of genes belonging to such clusters. However, when many genes possess high scores but are in the same LD interval, the algorithm picks automatically those having stronger connections with other genes located in different chromosomal regions. This matches the concept of Taşan *et al.* (2015) that genes with more connections are of higher importance. Nonetheless, SigMod is different from their approach, in that the algorithm decides itself the optimal number of genes to be selected in a LD interval, instead of given a 'prix fixe' constraint to select only one gene from it. We believe this is more rational as it is generally unsure whether there is only one causal gene in a LD interval.

In conclusion, we proposed an exact and efficient method SigMod for integrative analysis of GWAS data with network-based knowledge. Our method enables to find a functionally relevant gene module enriched with high association signals. It is robust against noise from either the GWAS results or the background network. Though our method is especially designed for identifying a gene module associated with disease (or trait), it can be applied to any other network-assisted feature selection problem of the same concept.

Funding

This work was funded by the Marie Curie Initial Training Network MLPM2012, Grant No. 316861. This work was also supported by the French National Agency for Research ANR-USPC-2012-EDAGWAS, ANR-11-BSV1-027-GWIS-AM and ANR-15-EPIG-0004-05 and a contract from the European Commission (018996).

Conflict of Interest: none declared.

References

Adamcsek,B. *et al.* (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**, 1021–1023.

Azencott,C.A. *et al.* (2013) Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, **29**, i171–i179.

Brossard,M. *et al.* (2013) Comparison of permutations strategies to assess gene- set significance in gene-set-enrichment analysis. *Abstracts from the 22nd Annual Meeting of the International Genetic Epidemiology Society*, Page 15.

Cabrera,C.P. *et al.* (2012) Uncovering networks from genome-wide association studies via circular genomic permutation. *G3: Genes | Genomes | Genetics*, **2**, 1067–1075.

Cabusora,L. *et al.* (2005) Differential network expression during drug and stress response. *Bioinformatics*, **21**, 2898–2905.

Çalışkan,M. *et al.* (2013) Rhinovirus wheezing illness and genetic risk of childhood- onset asthma. *New Engl. J. Med.*, **368**, 1398–1407.

Csardi,G., and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal Complex Syst.*, **1695**, 1–9.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Farh,K.K.H. *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.

Gallo,G. *et al.* (1989) A fast parametric maximum flow algorithm and applications. *SIAM J. Comput.*, **18**, 30–55.

Goldberg,A.V., and Tarjan,R.E. (1988) A new approach to the maximum-flow problem. *J. ACM (JACM)*, **35**, 921–940.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*, **4**, 44–57.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18(suppl 1)**, S233–S240.

Jia,P. *et al.* (2011) dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, **27**, 95–102.

Jia,P. and Zhao,Z. (2014) Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Hum. Genet.*, **133**, 125–138.

Klein,R.J. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Lamparter,D. *et al.* (2016) Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.*, **12**, e1004714.

Lee,I. *et al.* (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.

Lee,Y. *et al.* (2013) Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. *J. Am. Med. Informat. Assoc.*, **20**, 619–629.

Li,M.X. *et al.* (2011) GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.*, **88**, 283–293.

Li,Z.C. *et al.* (2015) Identification of drug–target interaction from interactome network with guilt-by-association principle and topology features. *Bioinformatics*, **32**, 1057–1064.

Liu,J.Z. *et al.* (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.*, **87**, 139–145.

Maher,B. (2008) Personal genomes: The case of the missing heritability. *Nature*, **456**, 18–21.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Moffatt,M.F. *et al.* (2010) A large-scale, consortium-based genomewide association study of asthma. *New Engl. J. Med.*, **363**, 1211–1221.

Oliver,S. (2000) Proteomics: guilt-by-association goes global. *Nature*, **403**, 601–603.

Rajagopalan,D., and Agarwal,P. (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, **21**, 788–793.

Raychaudhuri,S. *et al.* (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.*, **5**, e1000534.

Smit,L. *et al.* (2010) 17q21 variants modify the association between early respiratory infections and asthma. *Eur. Respir. J.*, **36**, 57–64.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Szklarczyk,D. *et al.* (2014) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, page gku1003.

Taşan,M. *et al.* (2015) Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nat. Methods*, **12**, 154–159.

Tomizono,S. (2013). *elliplot: Ellipse Summary Plot of Quantiles*. R package version 1.1.1.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

Wolfe,C.J. *et al.* (2005) Systematic survey reveals general applicability of guilt-by-association within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.

Xiong,W. *et al.* (2014) The centrality of cancer proteins in human protein-protein interaction network: a revisit. *Int. J. Computat. Biol. Drug Design*, **7**, 146–156.

Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*, **16**, 284–287.

SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network

Yuanlong Liu, Myriam Brossard, Damian Roqueiro, Patricia Margaritte-Jeannin, Chloé Sarnowski, Emmanuelle Bouzigon, Florence Demenais

1 Supplementary Figures

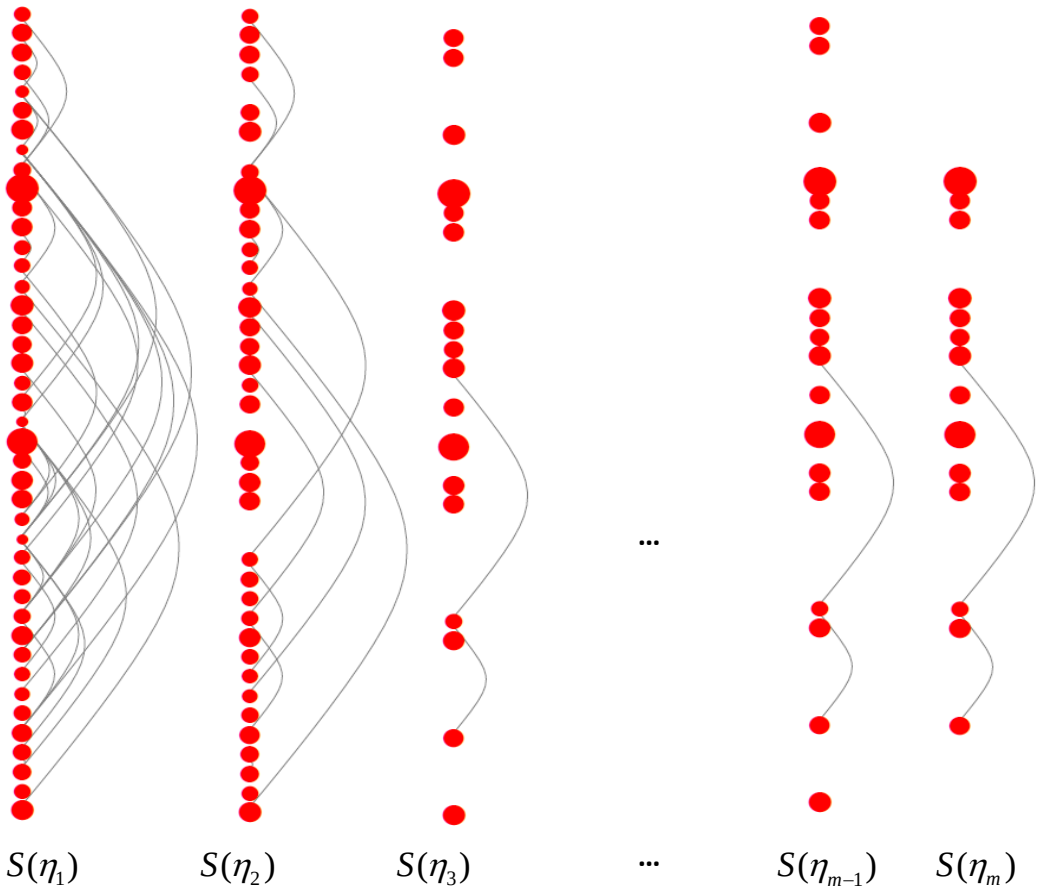


Figure S1: The selection path at a given λ value over the sparsity range $[\eta_{\min}, \eta_{\max}]$, defined as $\mathcal{P} = \langle S(\eta_1), ..., S(\eta_m) \rangle$. Here $S(\eta_1), ..., S(\eta_m)$ are the sequence of distinct modules selected by moving η from $\eta_{\min}$ to $\eta_{\max}$, and $\eta_1, ..., \eta_m$ are some sparsity values leading to these distinct selections ($\eta_{\min} \leq \eta_1 < ... < \eta_m \leq \eta_{\max}$). Red nodes represent genes and the curved lines are their connections. The *nesting property* is reflected by $S(\eta_1) \supseteq \cdots \supseteq S(\eta_m)$.

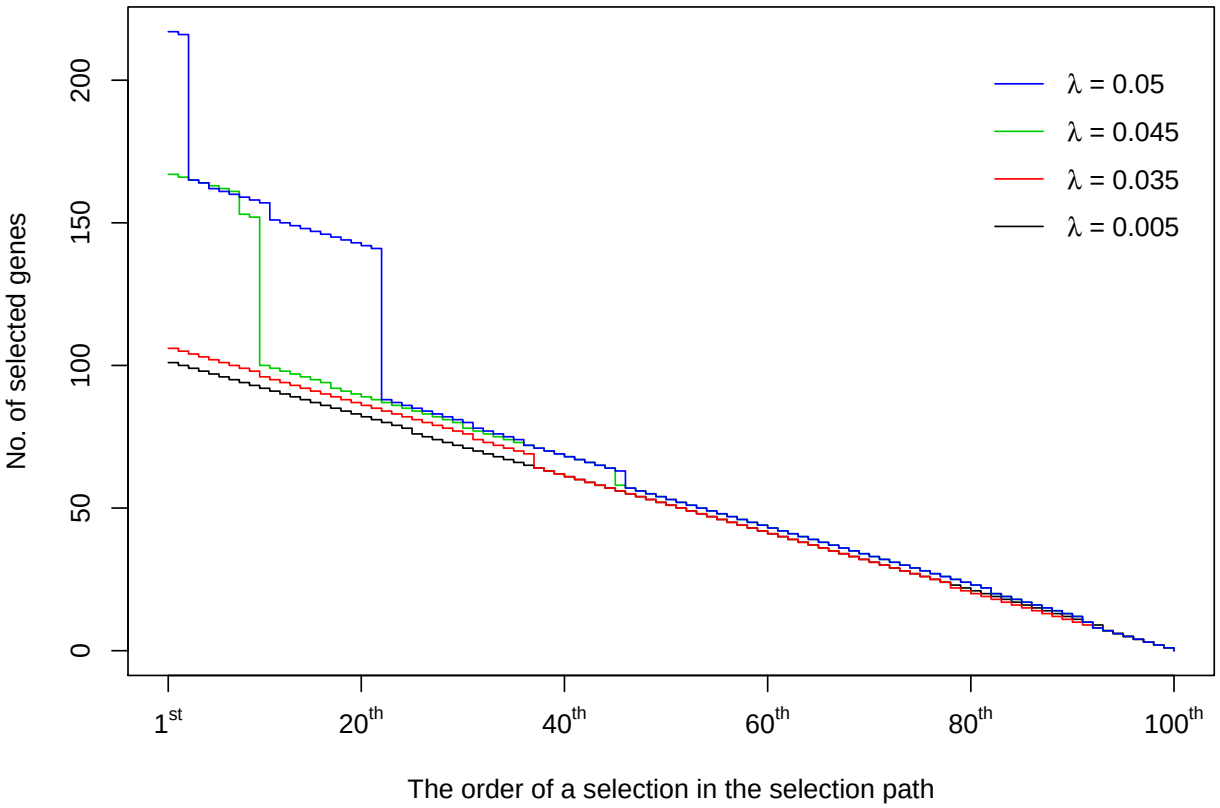


Figure S2: The number of genes in each module (S) from the selection path ($\mathcal{P}$). Four paths corresponding to four λ values are shown. The x-axis represents the order of each module in the path. The y-axis represents the number of genes in a module. Each path contains 100 distinct modules which end with an empty module (no gene selected). In a given path, a big size jump between two consecutive modules indicates some strongly interconnected genes are non-separable and are selected together. For larger λ values, the jumps are usually bigger, such as for $\lambda = 0.05$ (blue line).

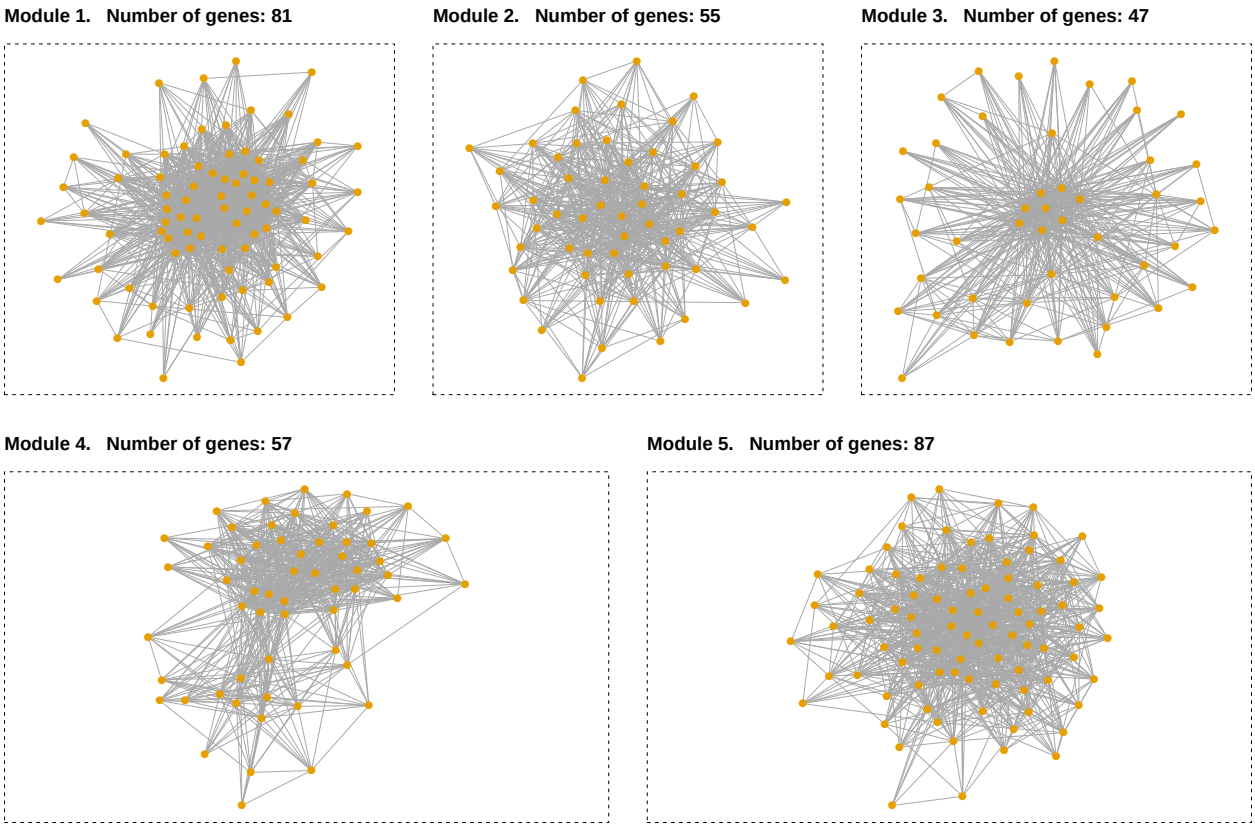


Figure S3: Five strongly interconnected gene modules identified by CFinder were chosen as causal module for simulation study.

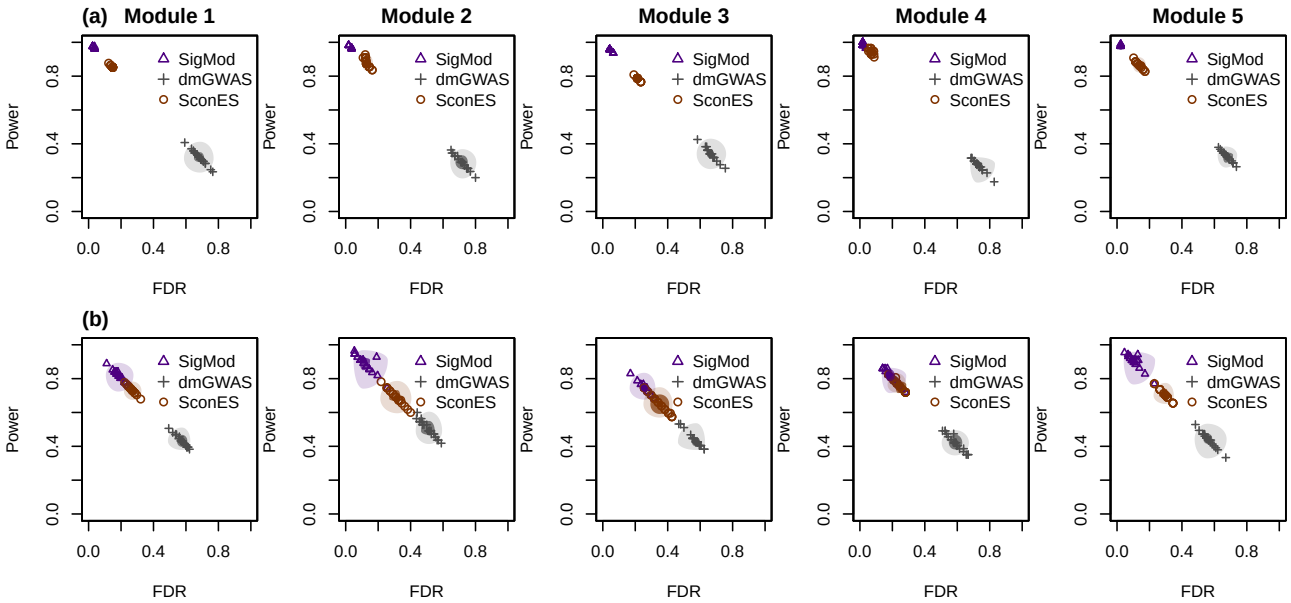


Figure S4: False discovery rate (FDR) versus power of three network analysis methods (SigMod, dmGWAS and SConES) applied to simulated data. In each experiment, one of the five modules identified by CFinder was set as the causal module. Plot (a) shows the results without adding noise to the GWAS data or GeneNet. Plot (b) shows the results with noise added to both gene-level P-values and gene network data.

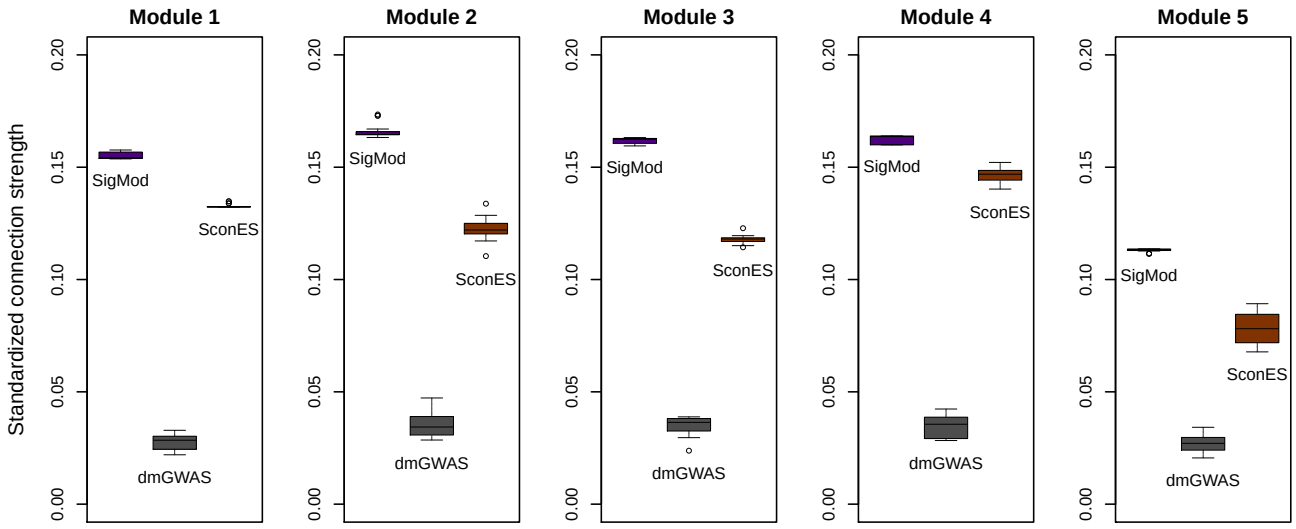


Figure S5: Box plots of the standardized connection strength (ρ) of gene modules identified by three network analysis methods (SigMod, SConES and dmGWAS). In each experiment, one of the five modules identified by CFinder was set as the causal module.

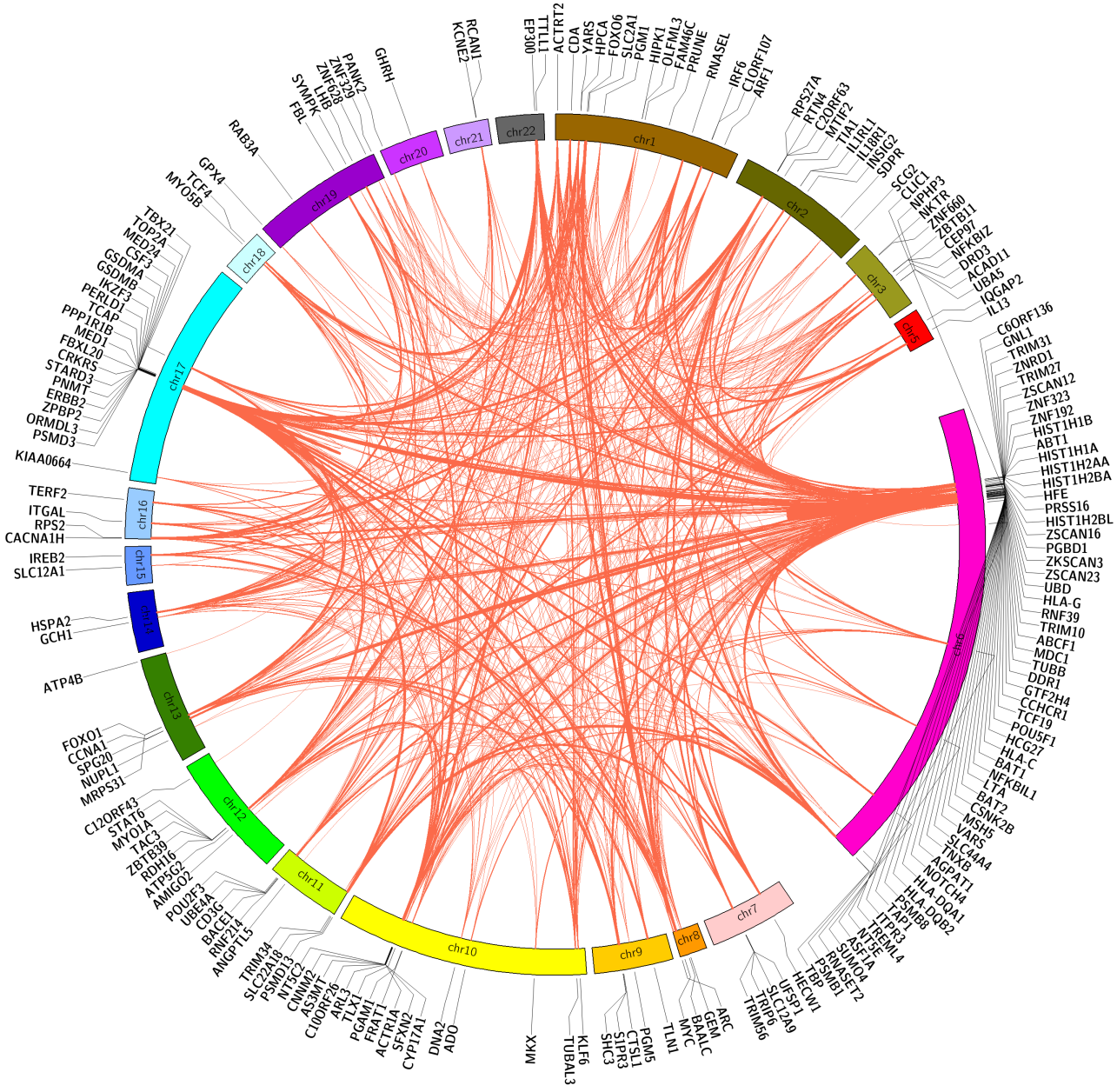


Figure S6: The gene module identified by applying SigMod to asthma META1 dataset. This module contains 190 genes and 1295 connections. In the figure the edge widths are proportional to edge weights, with smallest weight of 0.151 and largest weight of 0.999. Chromosome lengths have been rescaled for better layout.

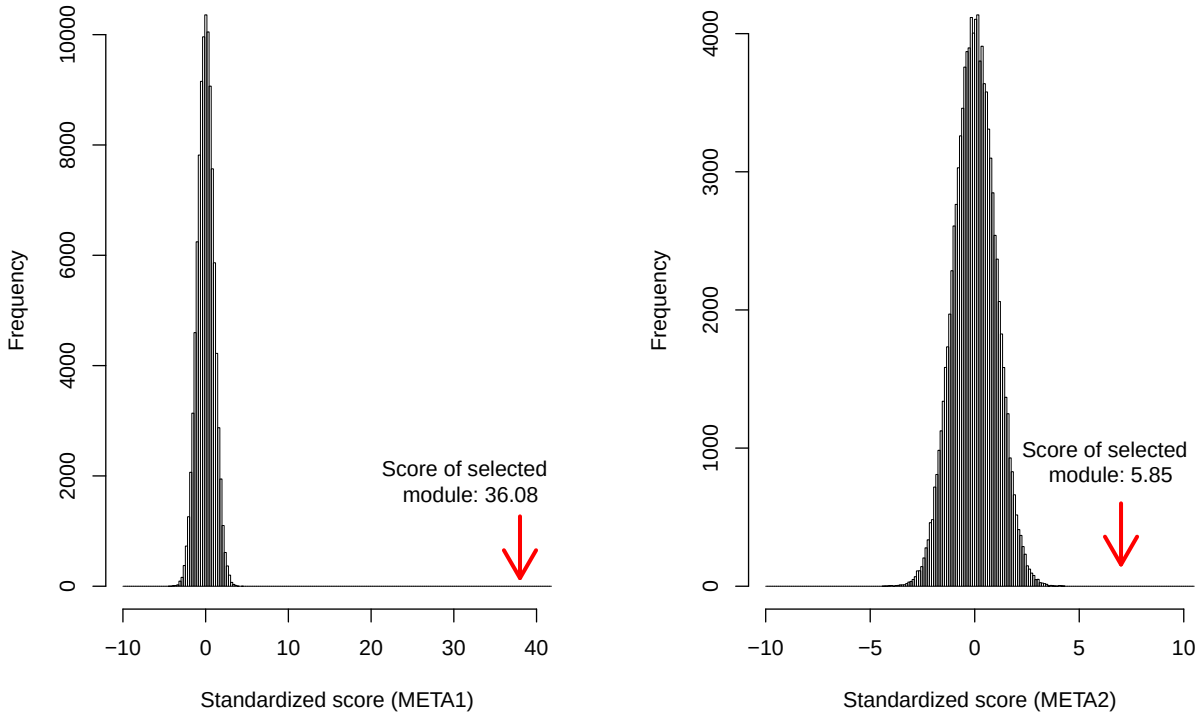


Figure S7: Histogram of the standardized module scores z^* of 100,000 random modules sampled from GeneNet (each module has 190 genes). The selected gene module has a significantly higher score than the random modules ($P < 10^{-5}$), evaluated using META1 and META2 dataset respectively.

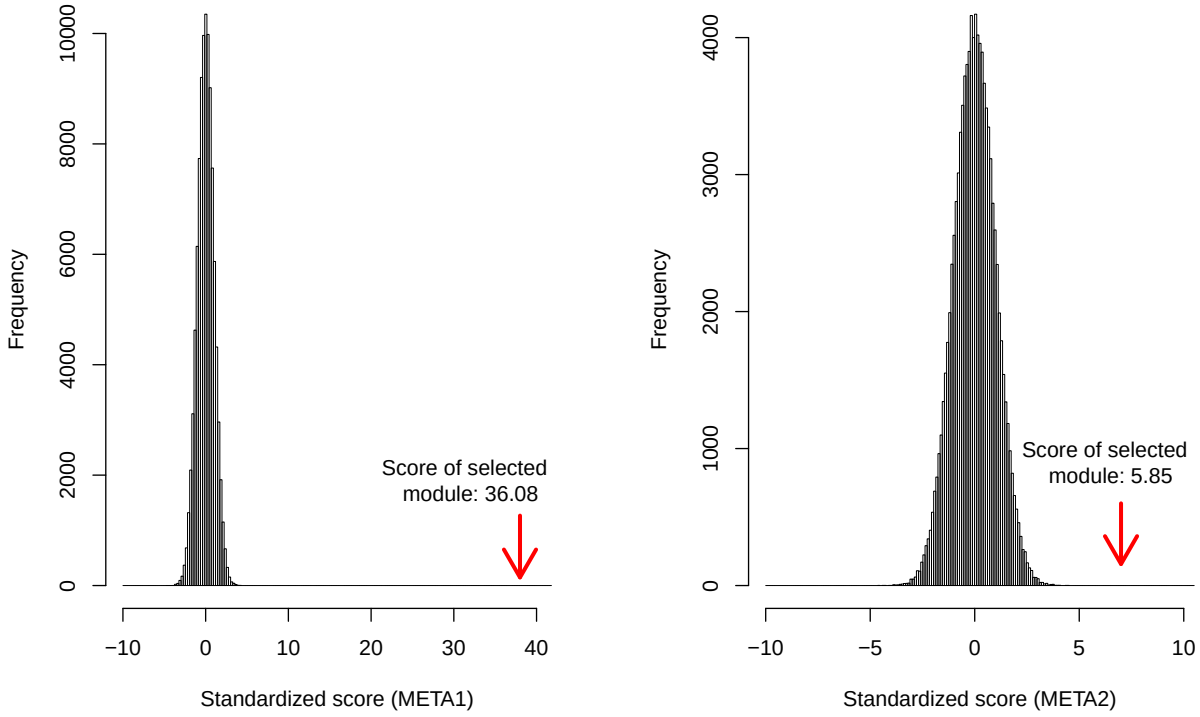


Figure S8: Histogram of the scores of selected gene module based on 100,000 random permutations of SNP-level statistics using Circular Genomic Permutation (CGP). The observed module score is significantly higher than those obtained from the permutation samples ($P < 10^{-5}$), evaluated using META1 and META2 dataset respectively.

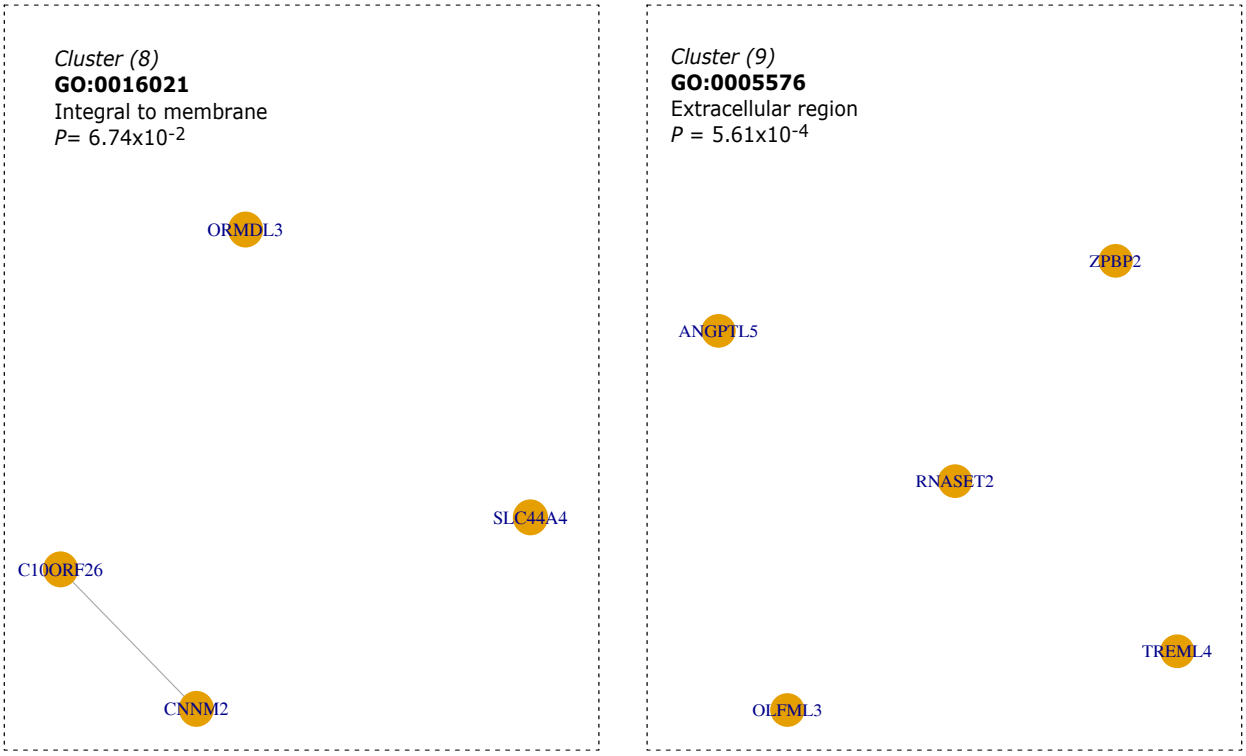


Figure S9: Two additional loosely connected or unconnected functional gene clusters out of nine gene clusters identified by DAVID in the selected gene module. The main function of each cluster is annotated using the gene ontology (GO) category with highest enrichment in these genes. P-values indicate the significance of enrichment of each GO category.

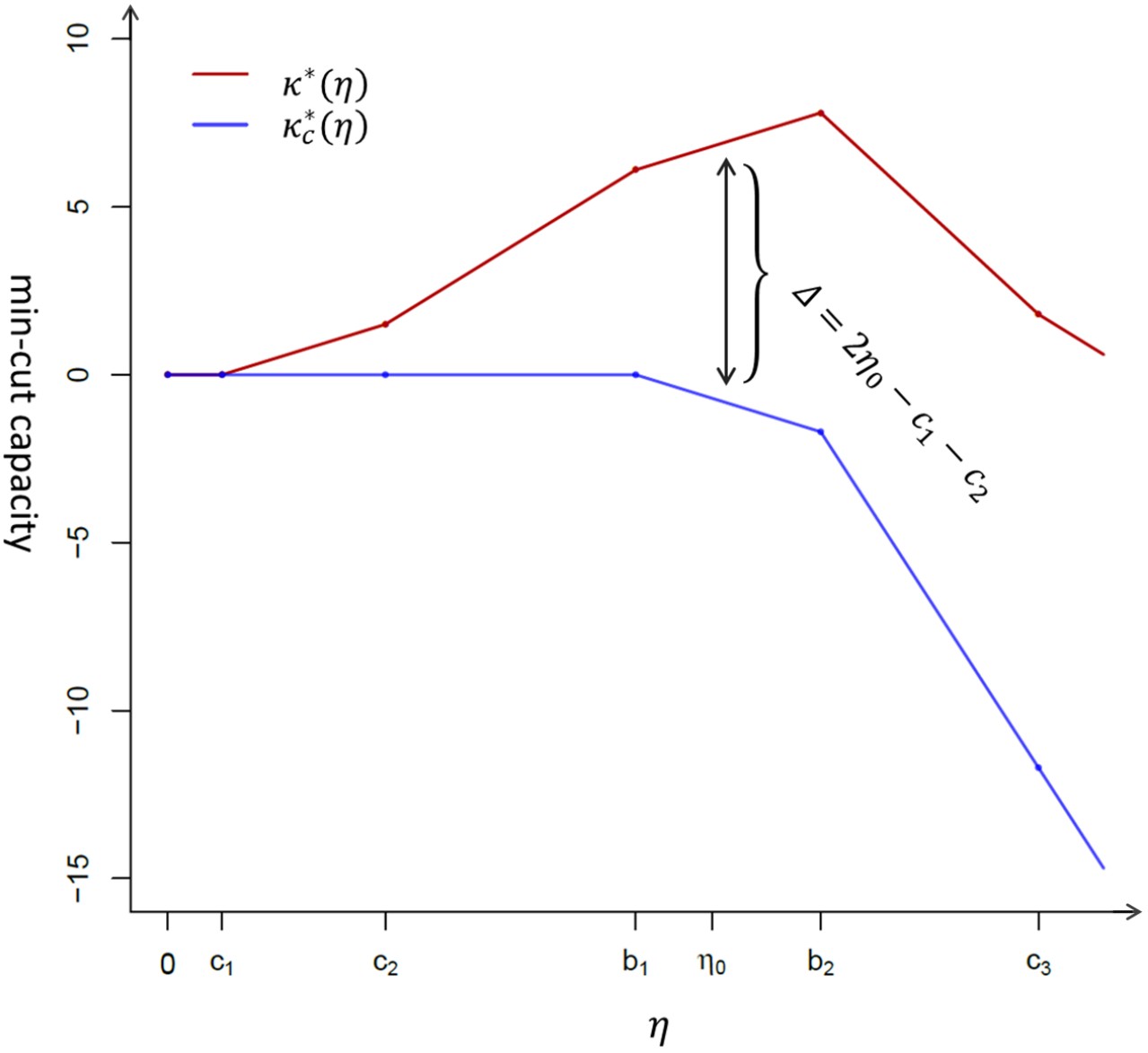


Figure S10: An example of *min-cut* capacity curve $\kappa^*(\eta)$ and its corrected curve $\kappa_c^*(\eta)$. $\kappa^*(\eta)$ has three change-points $\{c_1, c_2, c_3\}$ and two break-points $\{b_1, b_2\}$. It is piece-wise linear and concave between any consecutive change-points. The corrected curve $\kappa_c^*(\eta)$ is concave throughout its domain. The correction at η_0 is $\Delta = 2\eta_0 - c_1 - c_2$ according to Equation (S2).

2 Supplementary Tables

Table S1: Association statistics of 190 genes in the selected gene module for childhood asthma. $P_{\texttt{best-SNP}}$ is the best SNP P-value among all SNPs mapped to a gene. Gene P-value is the best SNP P-value corrected for gene length using CGP-permutation. "Rank" is the rank of a gene P-value among all genes mapped to the GeneNet evaluated using the META1 discovery dataset. The genes having significant P-values ($P \leq 0.05$) in both META1 and META2 datasets are shown in bold.

			META1 (Discovery data)				META2 (Replication data)		
No.	Chr.	Gene	best SNP	best SNP position	$P_{\text{best\_SNP}}$	Gene P-value	$P_{\text{best\_SNP}}$	Gene P-value	Rank
1	1	*FAM46C*	rs17037485	118162124	4.30E-04	1.45E-01	7.35E-03	6.80E-01	272
2	1	*HIPK1*	rs7545300	114503636	2.46E-04	5.32E-02	3.41E-03	2.64E-01	165
3	1	*OLFML3*	rs7364	114524661	3.66E-04	8.76E-01	9.63E-04	9.52E-01	58
4	1	*PGM1*	rs855311	64091739	3.87E-04	7.21E-03	8.28E-03	8.09E-02	294
5	1	***FOXO6***	rs7539614	41831047	1.22E-03	1.76E-04	7.89E-03	8.90E-04	284
6	1	*SLC2A1*	rs841847	43402708	8.64E-04	3.67E-02	1.30E-02	2.64E-01	422
7	1	*YARS*	rs2282294	33246177	2.59E-03	5.08E-01	1.67E-02	9.36E-01	524
8	1	*HPCA*	rs1284371	33352771	4.96E-03	7.86E-01	1.06E-02	9.39E-01	351
9	1	*CDA*	rs818202	20916791	3.63E-04	1.02E-01	3.91E-03	4.02E-01	178
10	1	*ACTRT2*	rs3795262	2938697	2.18E-03	3.06E-01	5.34E-03	4.76E-01	215
11	1	*PRUNE*	rs4970989	151003600	9.11E-04	1.24E-02	8.47E-03	6.48E-02	301
12	1	*RNASEL*	rs486907	182554557	2.15E-03	4.14E-02	1.19E-02	1.41E-01	387
13	1	*C1ORF107*	rs126280	210019824	3.31E-04	4.35E-01	5.72E-03	9.70E-01	227
14	1	*IRF6*	rs674433	209964875	1.14E-03	2.84E-01	1.07E-02	8.03E-01	353
15	1	*ARF1*	rs1188975	228277690	1.13E-03	1.49E-01	6.59E-03	4.17E-01	254
16	2	*TIA1*	rs2706769	70473453	2.93E-04	2.39E-01	3.05E-03	6.82E-01	157
17	2	*RPS27A*	rs2028139	55460017	9.54E-04	6.39E-01	3.26E-03	8.82E-01	161
18	2	*C2ORF63*	rs13032294	55404883	4.16E-04	5.01E-01	7.71E-03	9.90E-01	282
19	2	*MTIF2*	rs2043712	55480022	3.71E-04	3.29E-01	3.84E-03	8.24E-01	177
20	2	*RTN4*	rs11677099	55254165	4.09E-05	4.25E-02	2.12E-03	4.27E-01	124
21	2	***IL18R1***	rs3771166	102986222	1.87E-05	2.36E-05	8.20E-04	4.08E-04	52
22	2	***IL1RL1***	rs4988957	102968075	2.86E-05	2.17E-05	1.49E-03	5.63E-04	84
23	2	*INSIG2*	rs889904	118860471	9.40E-04	2.23E-01	8.66E-03	6.95E-01	309
24	2	*SDPR*	rs4853645	192704044	1.07E-03	2.50E-01	5.50E-03	5.73E-01	220
25	2	*SCG2*	rs2894511	224465827	9.31E-03	8.30E-01	8.56E-03	8.19E-01	305
26	3	*NFKBIZ*	rs604521	101550578	1.26E-03	1.17E-01	7.62E-03	3.55E-01	279
27	3	*ZNF660*	rs939649	44632212	3.27E-03	3.90E-02	7.55E-03	6.46E-02	277
28	3	*NKTR*	rs1062051	42672486	7.97E-04	1.01E-01	9.94E-03	4.88E-01	333
29	3	*CEP97*	rs2554962	101445118	7.15E-04	8.28E-02	4.25E-03	2.42E-01	187
30	3	*ZBTB11*	rs4683854	101385261	1.18E-03	7.80E-01	4.97E-03	9.77E-01	207
31	3	*DRD3*	rs324022	113887298	1.18E-04	5.06E-02	2.86E-03	3.57E-01	151
32	3	*NPHP3*	rs11708051	132410648	1.14E-03	5.87E-02	1.07E-02	2.74E-01	356
33	3	*UBA5*	rs1378807	132385190	1.27E-03	1.85E-01	6.34E-03	4.54E-01	247
34	3	*ACAD11*	rs2305627	132346992	2.84E-04	1.57E-01	4.12E-03	6.26E-01	183
35	5	*IQGAP2*	rs10514071	75994211	4.53E-05	1.54E-02	6.96E-03	5.35E-01	264
36	5	***IL13***	rs848	131996500	2.71E-04	1.02E-03	1.44E-03	2.44E-03	76
37	6	*PRSS16*	rs9393795	27217719	4.17E-04	3.34E-02	1.41E-03	5.54E-02	75
38	6	***ZNF192***	rs13205911	28124114	6.42E-05	7.37E-03	1.19E-03	4.14E-02	67
39	6	*ZNRD1*	rs8321	30032522	1.07E-03	1.89E-01	7.07E-03	5.34E-01	266
40	6	*ZSCAN12*	rs2232423	28366151	4.08E-04	9.53E-03	6.39E-03	7.38E-02	249
41	6	*TREML4*	rs7774363	41198145	2.36E-04	5.65E-02	4.30E-03	3.49E-01	188
42	6	*ABCF1*	rs3132610	30544401	7.59E-04	6.93E-01	5.32E-03	9.82E-01	213
43	6	*AGPAT1*	rs2269423	32145707	1.62E-04	1.28E-01	1.53E-03	3.82E-01	86
44	6	*BAT1*	rs2734583	31505480	3.56E-05	2.62E-01	9.48E-04	8.19E-01	55
45	6	***BAT2***	rs3132450	31596138	7.72E-04	6.16E-03	8.60E-03	4.01E-02	306
46	6	*C6ORF136*	rs9262135	30618906	4.69E-04	7.70E-01	6.79E-04	7.55E-01	44
47	6	*CCHCR1*	rs130073	31111180	3.05E-04	8.34E-02	6.46E-03	5.38E-01	252

48	6	*CLIC1*	rs3131383	31704294	5.20E-04	3.68E-01	1.23E-03	4.76E-01	70
49	6	*CSNK2B*	rs9267531	31636742	5.29E-04	2.83E-01	2.31E-03	5.39E-01	129
50	6	*DDR1*	rs1049633	30867527	4.07E-05	1.71E-01	8.85E-04	6.01E-01	54
51	6	*GNL1*	rs3130247	30515043	9.78E-04	7.82E-01	3.83E-03	9.70E-01	175
52	6	*GTF2H4*	rs1264308	30879987	5.96E-05	1.55E-01	7.56E-04	4.31E-01	46
53	6	*HCG27*	rs3132511	31170020	3.92E-04	6.01E-02	4.52E-03	2.79E-01	195
54	6	***HFE***	rs1045537	26096748	7.39E-04	7.24E-03	4.92E-03	2.73E-02	206
55	6	*HLA-C*	rs2001181	31236998	5.32E-04	3.91E-01	4.73E-03	8.67E-01	205
56	6	*HLA-DQA1*	rs9272853	32610705	3.09E-05	2.31E-01	3.04E-04	4.57E-01	23
57	6	*HLA-DQB2*	rs9276584	32730835	6.99E-04	3.83E-02	8.24E-03	2.16E-01	293
58	6	*HLA-G*	rs1610696	29798803	1.25E-03	6.60E-01	8.39E-03	9.78E-01	297
59	6	*ITPR3*	rs3736893	33639760	2.51E-05	2.02E-01	1.44E-03	9.22E-01	78
60	6	*LTA*	rs909253	31540313	1.52E-03	1.40E-01	4.71E-03	2.64E-01	203
61	6	*MDC1*	rs3094093	30679628	2.67E-03	3.46E-01	9.87E-03	6.62E-01	330
62	6	*MSH5*	rs3117574	31725230	4.35E-04	7.34E-02	4.72E-03	3.16E-01	204
63	6	*NFKBIL1*	rs2071592	31515340	7.20E-05	4.85E-02	1.04E-03	1.85E-01	60
64	6	*NOTCH4*	rs436388	32186264	2.40E-04	1.07E-01	6.29E-03	6.75E-01	245
65	6	*POU5F1*	rs13409	31132140	6.15E-04	3.29E-02	6.05E-03	1.58E-01	237
66	6	*PSMB8*	rs9276810	32810443	4.43E-04	4.09E-01	2.55E-03	7.66E-01	142
67	6	*RNF39*	rs9261290	30038647	1.03E-03	3.14E-01	5.75E-03	6.92E-01	229
68	6	*SLC44A4*	rs9267659	31846234	1.69E-04	1.30E-01	2.00E-03	4.60E-01	117
69	6	*TAP1*	rs6457684	32814447	1.09E-04	3.08E-01	1.81E-03	8.37E-01	101
70	6	*TCF19*	rs7750641	31129310	2.14E-05	3.52E-02	4.99E-04	1.59E-01	31
71	6	*TNXB*	rs3130286	32042322	3.28E-04	2.53E-01	6.68E-03	8.91E-01	259
72	6	*TRIM10*	rs2523735	30122657	4.04E-04	3.24E-02	5.17E-03	1.79E-01	211
73	6	*TRIM31*	rs1116221	30071330	6.70E-04	3.52E-02	8.10E-03	2.05E-01	290
74	6	*TUBB*	rs8233	30692965	8.13E-04	3.59E-01	2.87E-03	6.03E-01	152
75	6	*UBD*	rs404240	29523957	4.22E-04	4.36E-01	2.68E-03	8.14E-01	148
76	6	*VARS*	rs915652	31749142	3.83E-04	1.66E-01	2.46E-03	4.16E-01	133
77	6	*ABT1*	rs4634439	26598004	1.69E-03	9.26E-02	7.37E-03	2.36E-01	273
78	6	*ZNF323*	rs13217619	28306671	3.45E-04	1.22E-02	6.25E-03	1.03E-01	244
79	6	*ZSCAN16*	rs4713140	28097193	2.09E-03	6.35E-02	9.56E-03	1.79E-01	324
80	6	*ZSCAN23*	rs7766356	28400538	3.06E-04	3.15E-02	2.27E-03	9.84E-02	128
81	6	*TRIM27*	rs3135293	28877247	4.96E-05	1.60E-01	7.80E-04	4.87E-01	50
82	6	*HIST1H1A*	rs16891235	26017542	4.19E-04	2.46E-01	6.34E-04	2.17E-01	35
83	6	*HIST1H1B*	rs17763089	27835218	6.04E-04	6.24E-01	1.38E-03	7.58E-01	72
84	6	*HIST1H2BL*	rs200485	27775697	3.76E-03	7.01E-01	6.42E-03	8.28E-01	250
85	6	*PGBD1*	rs13211507	28257377	2.51E-04	1.66E-02	3.69E-03	1.01E-01	171
86	6	*ZKSCAN3*	rs6921919	28325201	5.77E-04	1.22E-02	6.43E-03	7.14E-02	251
87	6	*HIST1H2AA*	rs4711095	25726774	4.69E-03	4.21E-01	1.00E-02	6.25E-01	336
88	6	*HIST1H2BA*	rs9358872	25727517	4.18E-03	5.56E-01	1.11E-02	8.18E-01	366
89	6	*NT5E*	rs3812139	86196990	2.62E-04	5.67E-02	4.34E-03	3.26E-01	191
90	6	*ASF1A*	rs7772912	119218470	2.63E-03	1.73E-01	1.25E-02	4.52E-01	406
91	6	*SUMO4*	rs237025	149721690	5.64E-03	4.78E-01	5.44E-03	4.47E-01	219
92	6	*PSMB1*	rs12207633	170847181	2.53E-04	7.94E-02	4.16E-03	4.15E-01	185
93	6	*RNASET2*	rs1077453	167360024	4.28E-04	1.23E-01	5.33E-03	5.15E-01	214
94	6	*TBP*	rs12200657	170872108	1.34E-03	8.92E-02	9.76E-03	3.25E-01	325
95	7	*HECW1*	rs2330785	43276428	3.90E-05	2.26E-02	7.58E-03	7.27E-01	278
96	7	*TRIM56*	rs6948536	100731829	8.86E-04	6.55E-01	1.91E-03	7.87E-01	110
97	7	*SLC12A9*	rs314370	100453208	2.93E-04	2.25E-02	2.46E-03	7.99E-02	132
98	7	***TRIP6***	rs6706	100471044	3.32E-04	1.88E-02	1.20E-03	3.11E-02	68
99	7	*UFSP1*	rs12666989	100486754	2.02E-04	3.55E-01	6.35E-04	4.60E-01	36
100	8	*GEM*	rs1050616	95262253	9.92E-05	5.39E-01	1.40E-03	9.54E-01	74
101	8	*BAALC*	rs2454014	104154965	2.34E-04	1.97E-02	8.87E-03	2.86E-01	315
102	8	*MYC*	rs4645956	128750212	7.80E-03	5.32E-02	1.89E-02	1.04E-01	581
103	8	*ARC*	rs10097505	143694184	3.65E-03	6.81E-02	3.78E-03	5.54E-02	173
104	9	*TLN1*	rs10972567	35728019	1.42E-03	4.26E-02	1.10E-02	1.82E-01	363
105	9	*PGM5*	rs7874438	71114809	1.51E-04	7.30E-02	6.17E-03	6.58E-01	240
106	9	***CTSL1***	rs2378757	90343780	1.84E-03	6.89E-03	5.58E-03	1.37E-02	222

107	9	*S1PR3*	rs7858626	91612639	5.42E-06	1.50E-01	2.18E-04	4.03E-01	19
108	9	*SHC3*	rs1331180	91630548	1.98E-05	3.57E-02	1.63E-03	4.86E-01	91
109	10	*MKX*	rs2637277	27969032	3.54E-04	7.69E-03	8.82E-03	9.77E-02	314
110	10	*KLF6*	rs17731	3821561	1.24E-03	2.48E-01	4.64E-03	4.85E-01	200
111	10	*TUBAL3*	rs7097775	5435918	6.60E-04	3.12E-01	6.00E-03	8.04E-01	235
112	10	*ADO*	rs9990	64567938	3.35E-03	3.13E-01	9.24E-03	5.41E-01	317
113	10	*DNA2*	rs10998202	70224310	4.06E-04	3.75E-02	6.95E-03	2.64E-01	263
114	10	*ARL3*	rs2251772	104470918	1.52E-04	2.94E-01	1.74E-03	7.53E-01	98
115	10	*ACTR1A*	rs3781290	104244948	7.32E-04	3.68E-01	6.99E-03	8.76E-01	265
116	10	*FRAT1*	rs3781373	99080585	1.38E-03	6.45E-01	2.73E-03	7.77E-01	149
117	10	*CYP17A1*	rs10786712	104596396	4.21E-05	3.61E-01	7.25E-04	8.11E-01	45
118	10	*C10ORF26*	rs284858	104573936	6.52E-05	7.40E-02	2.59E-03	5.86E-01	145
119	10	*CNNM2*	rs943036	104836047	2.34E-05	6.52E-02	1.76E-03	6.72E-01	99
120	10	*NT5C2*	rs746293	104897254	1.78E-05	5.67E-02	9.50E-04	4.69E-01	56
121	10	*SFXN2*	rs1475644	104491164	2.21E-04	1.59E-01	2.26E-03	4.98E-01	127
122	10	*TLX1*	rs2235128	102896801	1.89E-03	5.71E-02	1.01E-02	1.81E-01	337
123	10	*AS3MT*	rs1591915	104646849	2.73E-05	3.42E-01	6.64E-04	8.71E-01	39
124	10	*PGAM1*	rs764223	99191935	9.68E-03	2.07E-01	8.87E-03	1.80E-01	316
125	11	*TRIM34*	rs3740998	5664831	4.94E-04	4.19E-02	7.55E-03	2.70E-01	276
126	11	***PSMD13***	rs577298	248002	9.13E-04	2.63E-03	1.35E-02	2.49E-02	430
127	11	*SLC22A18*	rs11024581	2939705	4.02E-04	1.48E-01	5.59E-03	6.17E-01	224
128	11	*ANGPTL5*	rs7109121	101766771	4.72E-04	1.48E-02	5.05E-03	7.64E-02	209
129	11	*CD3G*	rs3212264	118216234	3.83E-04	6.50E-01	1.83E-03	9.14E-01	104
130	11	*POU2F3*	rs7484249	120142118	7.26E-05	2.13E-02	2.38E-03	2.02E-01	131
131	11	*RNF214*	rs655023	117127620	1.09E-03	6.22E-02	9.43E-03	2.68E-01	322
132	11	*UBE4A*	rs12576486	118235490	2.42E-04	1.20E-01	3.12E-03	4.71E-01	159
133	11	*BACE1*	rs490460	117163765	1.01E-03	1.20E-01	8.31E-03	4.33E-01	295
134	12	*STAT6*	rs324015	57490100	8.73E-04	2.01E-02	5.35E-03	6.78E-02	216
135	12	*AMIGO2*	rs854889	47471037	4.42E-03	5.58E-01	7.31E-03	6.91E-01	270
136	12	*ATP5G2*	rs12422531	54067827	2.09E-03	4.83E-01	1.02E-02	8.70E-01	340
137	12	*RDH16*	rs901068	57346805	5.45E-04	2.77E-02	2.67E-03	6.94E-02	146
138	12	*TAC3*	rs733629	57406444	6.95E-04	7.30E-01	2.50E-03	9.36E-01	137
139	12	***ZBTB39***	rs4016338	57399015	3.46E-03	1.37E-03	9.48E-03	2.73E-03	323
140	12	*MYO1A*	rs17119344	57422934	7.69E-04	2.90E-01	6.63E-03	7.68E-01	257
141	12	***C12ORF43***	rs3751150	121442214	4.36E-04	1.25E-02	3.15E-03	4.30E-02	160
142	13	*NUPL1*	rs9551192	25904797	6.09E-04	7.89E-02	4.41E-03	2.65E-01	192
143	13	*CCNA1*	rs4245378	37007040	3.71E-03	2.62E-01	1.29E-02	5.40E-01	420
144	13	*SPG20*	rs9547247	36892264	2.99E-04	5.33E-01	6.17E-03	9.94E-01	241
145	13	*FOXO1*	rs7323267	41204015	5.74E-05	1.80E-01	2.55E-03	8.84E-01	143
146	13	*MRPS31*	rs9549281	41330171	1.73E-04	4.16E-01	1.59E-03	8.41E-01	90
147	13	*ATP4B*	rs11164142	114309226	5.49E-03	6.28E-01	8.79E-03	7.62E-01	313
148	14	*GCH1*	rs10498472	55354869	7.14E-04	1.21E-01	1.00E-02	5.86E-01	334
149	14	*HSPA2*	rs17101919	65007547	1.25E-02	3.02E-01	1.83E-02	3.94E-01	562
150	15	*SLC12A1*	rs11857986	48571605	2.74E-04	1.02E-01	6.60E-03	6.39E-01	255
151	15	*IREB2*	rs11630228	78736325	2.12E-04	4.52E-02	4.66E-03	3.39E-01	202
152	16	***ITGAL***	rs2230434	30518096	1.64E-04	2.28E-03	2.58E-03	1.58E-02	144
153	16	*CACNA1H*	rs4984637	1261282	1.02E-03	5.14E-01	7.44E-03	9.34E-01	275
154	16	*RPS2*	rs6366	2014031	1.38E-02	6.58E-01	1.23E-02	6.35E-01	397
155	16	*TERF2*	rs3785073	69401937	3.33E-03	9.11E-02	1.18E-02	2.16E-01	381
156	17	*KIAA0664*	rs11078312	2600186	9.92E-04	3.44E-02	5.59E-03	1.07E-01	223
157	17	***IKZF3***	rs907091	37921742	2.49E-15	1.55E-07	4.56E-05	7.09E-05	7
158	17	***CSF3***	rs25645	38173143	2.41E-06	1.19E-04	1.24E-04	4.50E-04	12
159	17	***ERBB2***	rs1058808	37884037	3.47E-06	1.62E-05	1.61E-04	8.19E-05	16
160	17	***STARD3***	rs11869286	37813856	4.57E-07	2.28E-04	9.08E-05	8.64E-04	9
161	17	*CRKRS*	rs11658678	37680096	1.91E-04	4.66E-02	2.54E-03	2.26E-01	141
162	17	*MED1*	rs10445306	37591422	4.30E-04	4.99E-02	2.50E-03	1.33E-01	136
163	17	***ORMDL3***	rs12603332	38082807	4.59E-15	7.31E-08	9.71E-06	7.18E-06	2
164	17	***PERLD1***	rs903502	37829604	5.32E-07	7.03E-05	1.55E-04	6.33E-04	15
165	17	***PNMT***	rs876493	37824545	1.34E-07	4.00E-05	3.08E-05	7.35E-05	6

166	17	***PPP1R1B***	rs2271309	37784990	2.20E-04	1.01E-03	1.07E-03	2.09E-03	62
167	17	***TCAP***	rs1053651	37822311	8.12E-06	3.65E-03	6.80E-05	2.68E-03	8
168	17	***ZPBP2***	rs12936231	38029120	1.03E-14	1.01E-07	2.24E-05	1.94E-05	4
169	17	*FBXL20*	rs8069451	37504933	4.12E-04	6.10E-02	7.84E-03	4.18E-01	283
170	17	***GSDMA***	rs7212938	38122680	2.42E-13	3.22E-07	2.66E-05	3.25E-05	5
171	17	***GSDMB***	rs9303281	38074046	2.59E-16	7.53E-08	5.49E-06	1.94E-05	1
172	17	***MED24***	rs12309	38175462	3.03E-06	1.20E-04	2.53E-04	9.25E-04	21
173	17	***PSMD3***	rs11655264	38138995	1.00E-07	6.93E-05	9.63E-05	5.80E-04	10
174	17	***TOP2A***	rs2586112	38572366	2.39E-04	7.33E-03	1.89E-03	2.44E-02	107
175	17	***TBX21***	rs12721470	45822579	1.91E-03	9.12E-03	1.14E-02	3.60E-02	372
176	18	***MYO5B***	rs7237973	47514403	6.74E-05	4.26E-04	1.21E-02	4.78E-02	389
177	18	*TCF4*	rs10515970	52980635	9.62E-05	2.95E-02	6.79E-03	5.24E-01	261
178	19	*RAB3A*	rs2271882	18309365	2.89E-03	1.38E-01	5.17E-03	1.79E-01	210
179	19	*GPX4*	rs713041	1106615	1.20E-03	4.01E-01	1.46E-03	3.70E-01	79
180	19	*FBL*	rs11083539	40325680	1.85E-03	3.51E-01	9.28E-03	7.41E-01	320
181	19	*LHB*	rs1800447	49519905	1.02E-02	5.45E-01	9.27E-03	5.17E-01	318
182	19	*SYMPK*	rs7258364	46356548	1.97E-04	8.16E-02	3.42E-03	4.42E-01	167
183	19	*ZNF329*	rs159667	58648359	2.79E-05	4.68E-01	4.01E-04	8.58E-01	28
184	19	*ZNF628*	rs4801677	55990975	3.54E-03	5.80E-01	8.04E-03	7.92E-01	289
185	20	*PANK2*	rs6052169	3896449	3.75E-04	2.88E-01	3.33E-03	7.33E-01	163
186	20	*GHRH*	rs4988492	35882698	6.86E-04	3.29E-01	1.54E-03	4.28E-01	88
187	21	*KCNE2*	rs1010668	35738158	4.47E-04	2.22E-01	1.48E-03	3.55E-01	83
188	21	*RCAN1*	rs1557270	35957254	2.52E-05	7.41E-02	1.96E-03	7.29E-01	113
189	22	*TTLL1*	rs5759126	43455665	3.11E-04	1.15E-01	6.81E-03	6.59E-01	262
190	22	*EP300*	rs1569857	41540934	8.82E-04	7.32E-02	1.29E-02	4.38E-01	419

Table S2: 15 KEGG pathways enriched in the selected gene module for childhood asthma

Enriched KEGG	*P*-value	Ratio
Epstein-Barr virus infection	3.28E-04	12/202
Herpes simplex infection	1.57E-03	10/185
Thyroid hormone signaling pathway	1.57E-03	8/118
HTLV-I infection	3.34E-03	11/258
Antigen processing and presentation	3.34E-03	6/77
Inflammatory bowel disease (IBD)	9.14E-03	5/65
Type I diabetes mellitus	1.15E-02	4/43
Proteasome	1.15E-02	4/44
Phagosome	1.50E-02	7/154
Viral carcinogenesis	1.78E-02	8/205
Viral myocarditis	2.42E-02	4/59
Glucagon signaling pathway	3.07E-02	5/101
Central carbon metabolism in cancer	3.20E-02	4/67
Allograft rejection	3.55E-02	3/38
Graft-versus-host disease	4.08E-02	3/41

Note: *P*-values are FDR adjusted; Ratio is the number of genes from the 190 module genes that map to the pathway divided by the total number of genes that map to the canonical pathway.

3 Supplementary Notes

3.1 Computing the selection path at any given λ value

In our formulation, a module is selected by solving the optimization problem

$$\arg\max_{\boldsymbol{u}} \boldsymbol{z}^\intercal \boldsymbol{u} + \lambda \boldsymbol{u}^\intercal A \boldsymbol{u} - \eta \|\boldsymbol{u}\|_0. \tag{S1}$$

Therefore the selection can vary for different values of λ and η. To conveniently trace the selection change, we develop the path algorithm that enables computing all distinct module selections at any fixed λ value. Specifically, we define the selection path of a given λ value over a sparsity range $[\eta_{\min}, \eta_{\max}]$ as the sequence of distinct modules selected by moving η from $\eta_{\min}$ to $\eta_{\max}$, denoted as $\mathcal{P} = \langle S(\eta_1), ..., S(\eta_m) \rangle$ (as described in the main paper). Note that the entire selection path can be obtained by setting $\eta_{\min} = 0$ and $\eta_{\max} = +\infty$.

To compute $\mathcal{P}$, we define the capacity function $\kappa^*(\eta)$ as the capacity of the *s-t min-cut* on G_{st} (*s-t min-cut*, capacity and G_{st} are defined in the main paper). Note that $\kappa^*(\eta)$ can be expressed as $\kappa^*(\eta) = (k_1 - k_2) \times \eta + C$, where k_1 is the number of edges connecting the selected nodes and the sink node t in G_{st}; k_2 is the number of edges connecting the unselected nodes and the source node s; C is some constant independent of η. Vary the value η will not change the cut edges unless: (1) η is set as a value that causes the rewiring of an edge of G_{st} from s to t according to Equation (3) (given in the main paper), or (2) η is set a a value that leads to the change of a selection. Therefore $\kappa^*(\eta)$ is a continuous and piece-wise linear function of η. Its slope changes at either a break-point or a change-point, where a η value is called a break-point if it leads to the change of selection, and called a change-point if it causes the rewiring of an edge.

Thus, computing the selection path is equivalent to finding the break-points of $\kappa^*(\eta)$. As our formulation satisfies all conditions of *parametric maximum flow* algorithm (Gallo *et al.*, 1989), hence $\kappa^*(\eta)$ is concave between any consecutive change-points (Gallo *et al.*, 1989). Therefore we can transform $\kappa^*(\eta)$ into a concave function throughout its domain, by correcting at each of its change-points so that the slope does not change at these values (as shown in Figure 10). We achieve this by using the following correction function

$$\kappa_c^*(\eta) = \kappa^*(\eta) - \sum_{c_p \in \mathcal{C}; c_p \leq \eta} (\eta - c_p), \tag{S2}$$

where $\mathcal{C}$ represents the change-points that can be obtained from Equation (3) in the main paper, simply, $\mathcal{C} = \{c_p \mid c_p > 0; c_p = z_p + \lambda d_p; 1 \leq p \leq n\}$.

The corrected capacity function $\kappa_c^*(\eta)$ has no change-points but the same break-points as $\kappa^*(\eta)$. It is also piece-wise linear and strictly concave throughout its domain. Thereby all break-points can be calculated by applying the iterative contraction algorithm described in Gallo *et al.* (1989). When all break-points are obtained, the selection path can be computed by setting η at each of the break-points and solving the corresponding optimization problem defined by Equation (S1). The pseudocode is given in Algorithm 1.

3.2 The nesting property and memoryless property of our module selection method

The module selection by solving Equation (S1) has the nesting property that $S(\eta_1) \subseteq S(\eta_0)$ if $\eta_1 > \eta_0$. It also has the memoryless property, that if a gene is not selected by setting η at some value (e.g., $\eta = \eta_0$), then it can be removed from the GeneNet when computing the selection at a η value greater than η_0, as demonstrated by the following proposition. Here $S(\eta)$ represents the module selected by setting the sparsity parameter as η.

Proposition 1. *Denote the selected genes by setting the sparsity parameter at η_0 as $S(\eta_0)$. The subnetwork of G induced by $S(\eta_0)$ is denoted as G_{sub}. Then, for any $\eta_1 > \eta_0$,*

$$S(\eta_1) = S_{sub}(\eta_1),$$

where $S_{sub}(\cdot)$ represents the selection by solving Equation (S1) *defined on the subnetwork G_{sub}.*

Proof. As the module selection via Equation (S1) satisfies all conditions of *parametric maximum flow* algorithm (Gallo *et al.*, 1989), therefore the selection has the nesting property (according to Lemma 2.4 in Gallo *et al.* (1989)):

$$S(\eta_1) \subseteq S(\eta_0) \text{ for any } \eta_1 > \eta_0.$$

Denote S_0 as the node indices of $S(\eta_0)$, then maximizing $f(\boldsymbol{u}, \eta_1)$ is equivalent to maximizing $f\Big(\big[\boldsymbol{u}_{S_0}^\intercal, \mathbf{0}^\intercal\big]^\intercal, \eta_1\Big)$. This is because according to the nesting property, the nodes unselected by setting the sparsity parameter at η_0 will not be selected by setting at a larger value η_1 neither. Note $f\Big(\big[\boldsymbol{u}_{S_0}^\intercal, \mathbf{0}^\intercal\big]^\intercal, \eta_1\Big) \equiv f(\boldsymbol{u}_{S_0}, \eta_1)$, where the latter is the objective function defined on the induced network $G_{\texttt{sub}}$. This leads to $S(\eta_1) = S_{\texttt{sub}}(\eta_1)$, hence Proposition 1 holds. □

Algorithm 1 Compute the selection path $\mathcal{P}$ over a sparsity range $[\eta_{\min}, \eta_{\max}]$

function SELECTION_PATH($\eta_{\min}, \eta_{\max}$)
  Compute $S(\eta_{\min})$ and $S(\eta_{\max})$ ▷ As described in the main paper
  if $S(\eta_{\min}) == S(\eta_{\max})$ **then** ▷ No more selection between them
    $path \leftarrow S(\eta_{\min})$
  else ▷ Use the divide-and-conquer strategy
    Compute $\kappa^*(\eta)$ at $\eta_{\min}$ and $\eta_{\max}$
    Compute $\kappa_c^*(\eta)$ at $\eta_{\min}$ and $\eta_{\max}$ according to Equation (S2)
    Compute the tangent lines of $\kappa_c^*(\eta)$ at $\eta_{\min}$ and $\eta_{\max}$
    $\eta_{\text{mid}} \leftarrow \eta \in [\eta_{\min}, \eta_{\max}]$ where two tangent lines intersect ▷ The middle point
    $path\_head \leftarrow$ SELECTION_PATH($\eta_{\min}, \eta_{\text{mid}}$)
    $path\_tail \leftarrow$ SELECTION_PATH($\eta_{\text{mid}}, \eta_{\max}$)
    $path \leftarrow path\_head \cup path\_tail$ ▷ Put two sub-paths together
  end if
  $path \leftarrow$ order the selections in $path$ by the size (number of nodes) of each selection
  return $path$
end function

References

Gallo, G. *et al.* (1989). A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, **18**(1), 30–55.

CHAPTER V. DISCUSSION, PERSPECTIVES, AND CONCLUSION

Genome-wide association studies of multifactorial diseases, such as asthma, have identified many genetic variants associated with these diseases, but these variants usually explain only a part of the whole genetic component. The limitations of current methods for GWAS motivated us to investigate in alternate approaches that can complement the conventionally used single-marker approach. As stated in the introduction section, the primary goal of this thesis is to explore network-based analysis strategies that combine GWAS outcomes with biological networks to identify functionally-relevant gene modules underlying disease, and to design novel methods to facilitate similar studies. The studies conducted in this thesis involved majorly five issues to be addressed at different stages: (1) SNP to gene mapping; (2) combining SNP *p*-values into gene *p*-values; (3) choosing the gene network; (4) searching for active modules; and (5) evaluation and interpretation of results. In the following, I will focus on point (1)-(4) to discuss the strategies and improvements we made to address related issues. I will also talk about the perspectives that can be made to optimize the performance at each stage.

1 Mapping SNPs to genes

A GWAS is performed at SNP-level whereas the network knowledge is generally given at gene/protein-level. To bridge this gap, SNPs need to be first mapped to genes. While an intragenic SNP that locates between the start site and 3'-untranslated region of a gene is usually mapped to that gene, methods can vary when considering SNPs outside this region. The most commonly used approach is to define a "flanking window" that extends the start and stop positions of a gene by some kilobases (kb). The SNP database dbSNP (Sherry et al., 2001), for example, uses an upstream extension of 2 kb and a downstream extension of 0.5 kb to define gene boundaries. Several studies have used larger extensions of 100 kb or up to 500 kb (Luo et al., 2010; Wang et al., 2007). When a SNP can be mapped to multiple genes as a result of overlapping windows, the closest gene is usually chosen (Lehne et al., 2011; Wang et al., 2007). A few studies also considered mapping intergenic SNPs to genes if a SNP falls within a LD block spanning the gene (Chapman et al., 2003). Hybrid approaches that combine the flanking window strategy and the LD strategy were also considered (Hong et al., 2009; Lehne et al., 2011).

In this thesis work, we used a stringent intragenic mapping criterion that does not consider any extension of the gene boundaries. This criterion has been used in the analysis of Ballard et al. (2009) and Peng et al. (2010), among others. Our choice may lead to a loss in power but may reduce false positives. Although a flanking window approach can be used to extend a few kilobases both upstream and downstream of a gene, it was shown in a comparison study that varying the window size from 0 to 250 kb did not significantly affect the power of the related network analysis (Lee et al., 2011). Moreover, the extension of boundaries to flanking regions of a gene may increase the degree of overlap of nearby genes and thus increase the number of inappropriate SNP to gene mapping.

On the other hand, this SNP to gene mapping issue reveals the limitation of analyzing GWAS dataset at the gene-level, and, by extension, the limitation of network-based analysis and other gene-based analyses as a whole. It was reported that the majority of SNPs in a GWAS were found to fall in intergenic regions (Macintyre et al., 2014; Schork et al., 2013). The genes which these SNPs have an effect on can be difficult to determine (Macintyre et al., 2014; Witte, 2010). In our study of asthma GWAS data, 1,388,029 out of a total of 2,370,689 SNPs were not mapped to any gene (57%). Though the flanking window or the LD-based mapping

strategy may help link some of these SNPs to nearby genes, their contributions are relatively small. This results in many GWAS hits left with no link to any gene and are discarded for downstream network analysis. More sophisticated SNP to gene mapping strategies that extend the distance-based mapping to function-based mapping, such as based on gene expression through expression quantitative trait loci (eQTLs), or by defining a more precise regulatory domain for each gene (McLean et al., 2010), may be considered and have the potential to boost the performance of network-based analysis.

An alternate strategy that can circumvent the SNP to gene mapping issue and include more SNPs into the analysis is to use a SNP network instead of a gene/protein network, although this is less feasible at present because SNP-SNP interactions are poorly characterized and the established knowledge is relatively sporadic. Nonetheless, there exist several studies that described the construction of SNP network at genome-wide level. Azencott et al. (2013) constructed a SNP network by linking SNPs if they are adjacent on the genomic sequence, or near the same gene (within a specified distance), or near two interacting genes in a gene network. Liu et al. (2012) described the construction of a trait-specific SNP network by testing statistical interaction between SNPs for all pairwise SNP combinations. Given the interaction statistics, an unweighted network can be built based on a user-defined significance cut-off, whereas a weighted network can also be constructed with the edge weights revealing the interaction significance. The construction of such a trait-specific SNP interaction network will become more practical along with the progress of techniques for testing genome-wide SNP-SNP interactions (Lin et al., 2016; Prabhu et al., 2012; Wan et al., 2010), thus will make network-based analysis at SNP-level more feasible. An additional benefit to performing network-analysis at SNP-level is that the SNP statistics are overloaded directly onto the SNP network. Therefore there is no more need to combine SNP *p*-values into gene *p*-values. The potential false positives/negatives caused by methodological limitations of *p*-value combination methods can be avoided. However, the biological interpretation of the analysis results at SNP-level will require complex investigations.

2 Combining SNP *p*-values to gene *p*-values

After mapping SNPs to genes, a statistical question is how to aggregate *p*-values at SNP-level into gene-level. Our study described in Chapter III included a novel, exact and efficient algorithm named fastCGP that was designed for this purpose. Although there exist various methods addressing this question, they require individual genotype data or an external SNP reference panel to compute the correlation among SNP *p*-values (or the intermediate metrics transformed from SNP *p*-values). It is of note that the use of an external reference panel may not always reflect the correlation structure in the dataset under survey, especially in those obtained from large genetic consortiums which are usually composed of different populations. Advantageously, fastCGP utilizes the LD pattern existing in the dataset that is under survey, thus can be more appropriate in some cases.

Given the advantages of fastCGP, our next goal is to explore whether it can be applied to compute the *p*-value of a general genetic entity that is defined as an arbitrary collection of genes. This includes pathways and gene modules. Unfortunately, the current design of the analytical computational algorithm has a restriction that the SNPs annotated to the genetic entity should be consecutive on the genome, i.e., the entity represents a continuous segment of the chromosome. This is true for genes, but not the case for pathways or gene modules in general. Yet, adopting the definition of the corrected *p*-value and other terminologies as described in the fastCGP algorithm presented in Chapter III, it is possible to adapt our analytical algorithm to compute the *p*-value in the general case. We will investigate the details in the future.

Currently, the fastCGP method takes only the best SNP *p*-value (*top 1*) to represent the gene. We will further seek the feasibility of combining *top x%* of the SNP *p*-values to represent the gene *p*-value. This *top x%* option can be more powerful when the trait is highly polygenic (Li et al., 2011), and has been implemented in several of the gene-based methods, like MAGMA, VEGAS, and VEGAS2. There is no question that the top *x%* option can be combined with the CGP strategy, but a brute-force implementation needs to be used to compare the observed gene representative statistic (T) with that of CGP samples, which, its time complexity is similar to that of VEGAS and thus will lose its "fast" nature. A more efficient strategy implementing this functionality deserves to be further explored.

3 Choosing the gene/protein network

The performance of network-based analysis depends heavily on the network data used for the analysis. Both the extensive coverage and high accuracy of interaction information are important to success. Yet, it is widely recognized that our knowledge of protein interaction is far from complete (Kelly et al., 2012; Menche et al., 2015). The interaction information stored in different databases also have relatively low overlap with each other (Mbiyavanga, 2014). These databases are likely to provide complementary network knowledge instead of replicable knowledge. Another concern is the quality of protein interaction information. Unlike pathway resources that are acquired through laboratory studies or careful manual curation by domain experts, protein interactions are identified via diverse techniques including in vivo/in vitro experiments and also in silico predictions, thus can be more heterogeneous in data quality.

Some early studies used network information retrieved from a single primary database (Tu et al., 2006; Wu et al., 2009) , while the rapid growth of available network collections has allowed recent studies to recruit multiple databases (or a meta-database) to increase quality and/or coverage, such as conducted in Hillenmeyer et al. (2016); Jia et al. (2011); Wang et al. (2015), and in our study as described in Chapter III and IV. A useful level of information provided in some databases is the interaction score summarised from various sources of evidence. This score quantifies the quality/confidence of an interaction and has the advantage that it can be updated constantly and is expected to converge to the ground truth. Integration of this level of information into network-based analysis, as enabled by our SigMod method, can improve the performance for finding meaningful results.

Another concern of network data is that most available resources are static and lack the temporal and spatial dependence of interactions in real biological systems. Ignoring context information can affluence the study performance since gene/protein expression levels and interactions are known to vary across cells and tissue types (Barshir et al., 2012; Kotlyar et al., 2016). Taking the tissue-specific information into account in a network-based analysis is essential to find bona-fide disease mechanisms and can reduce false positives. In recent years, tissue-specific networks are accumulating. For example, the TissueNet database (Barshir et al., 2012) and the IID database (Kotlyar et al., 2016) provide tissue-specific protein-protein interactions for various tissues of human and other module organisms. We foresee tissue-specific network data will become more and more abundant. In the future, more network-

based analysis will be performed in its tissue-specific context that can best capture the biological activities relevant to the disease condition.

4 Searching for active modules

The search for gene modules consisting of related genes and enriched in high and replicable association signals is the main objective of a module search method. Given the noise in the GWAS dataset, the heterogeneity of quality over the interaction information in a gene/protein network, and the huge search space at genome-wide scale, this objective can be difficult to reach. In this thesis, we have designed two strategies to best approach this goal. They are discussed below.

4.1 The DMS-based bi-directional module search strategy

We have first exploited the DMS module search strategy implemented in the dmGWAS package (Jia et al., 2011). DMS searches module in a greedy manner, and has been used in several network-based GWAS analyses (Han et al., 2013; Jia et al., 2012). One limitation of DMS is that it has a heuristic nature, thus does not guarantee to find the module having the highest association score but can include irrelevant genes by chance. To reduce the amount of irrelevant findings, previous studies have considered using two or more datasets, and implemented cross-evaluation strategies to identify modules possessing consistent association signals across datasets (Han et al., 2013; Jia et al., 2012). Such approaches were shown to be able to identify reliable and replicable results. This motivated our choice, as described in Chapter III, to use two large datasets, which resulted from a meta-analysis of nine asthma GWAS each, and implemented the bi-directional approach to search for consistent gene module(s). Our strategy used the DMS algorithm to generate raw gene modules within each dataset and then selected modules that had consistent gene compositions between the two datasets. Unlike previous studies that defined consistency based on module score, we quantified it in terms of gene composition. This consistency measure takes into account the module topology, thus leads to select modules that share genes having high association signals and other genes closely connected to these genes—both types of genes may play a role in childhood-onset asthma susceptibility.

We demonstrated our bi-directional strategy is more appropriate than a pooled analysis strategy performed on a single dataset made of the meta-analysis results of all 18 childhood-onset asthma GWAS. The genes selected by the later strategy have less replicable signals in the subdatasets, and were less functionally related based on DAVID functional clustering

analysis. This demonstrated the advantage of using a bi-directional strategy at least in our study.

4.2 The SigMod strategy

Through the application of the DMS method and the exploration of many other active module search algorithms, we realized their limitations in general. Most of them do not guarantee to find the global optimum result, do not utilize information on the strength of protein-protein interaction (when it exists), and are prone to be affected by noise from input GWAS and network data. Though the design of a sophisticated strategy by using multiple discovery datasets could improve the performance, such implementation increases the analysis complexity, and is restricted to a few studies that include more than one dataset. Ideally, a module search method should hold as many of these properties: the capability of taking the network quality into account, the ability to find the exact or close-to-exact solution, and can be executed in affordable time. The awareness of both the importance of these properties and the limitations of current tools motivated us to design a novel module search method named SigMod, as described in chapter IV and detailed in Liu et al. (2017).

In SigMod, the active module search task was formulated as a binary optimization problem. Its goal was to select a set of genes that are enriched in high association signals and tend to be strongly interconnected. An exact and efficient algorithm was proposed to solve the optimization problem. This novel method has several advantages compared to existing methods, including the ability to find the exact solution, to incorporate edge weights, and its robustness to background noise. This method was applied to both simulated and real data, and was shown to outperform two state-of-the-art methods in terms of increased power and decreased false discovery rate. When SigMod was applied to childhood-onset asthma data, it successfully identified a gene module enriched in consistently high association signals and made of functionally related genes that are biologically relevant for asthma. This method was implemented in an R package and is available to the public.

One future direction of updating SigMod is to improve the strategy for setting its tuning parameters. Setting the tuning parameters is a common issue involved in many active module search methods, including the popular methods DMS and jActiveModules. There are two tuning parameters in SigMod: the interconnectivity parameter λ and the sparsity parameter η.

As was discussed previously, η controls the number of genes to be selected. Increasing η results in removing genes that have less contribution to the module quality. Our selection path algorithm enables to compute all distinct selections for varying η values. In the next step, we will implement a visualization function, either within the R package or in a Cytoscape application, so that users can easily trace the selection change. This will facilitate the user to select a desirable amount of genes based on interactive result diagnosis, hence serves as an additional level of prior information for selecting modules according to "expert knowledge" instead of using a fixed parameter setting strategy that may be suitable in some situations but may not be optimal in other situations.

Our principle of determining the interconnectivity parameter λ is to set it as large as possible, as long as the selected module is enriched in high association signals. This encourages selecting strongly interconnected genes, for which the advantages have been detailed in our article (Liu et al., 2017) and discussed in Chapter III. The strategy we implemented is based on an empirical observation, that some strongly interconnected genes will be selected/unselected simultaneously when λ is big enough. Yet this strategy has limitations. It involves the computation of selection path for various λ values that will multiply the computational time, and requires careful inspection of size jump in each selection path. The results can also vary according to the range and different numbers of λ values to be computed. We will explore further strategies that make the determination of λ easier.

4.3 Possible extensions of the DMS-based method and the SigMod method

As described in the introduction to this thesis, there are two categories of network-based methods for analysing GWAS in general—the active module search methods that search for signal enriched modules in a scored network, and the seed gene oriented methods that explore the topological feature of the network with respect to a list of seed genes (seed genes are genes that have disease-association evidence reported from previous studies or summarized from the current GWAS outcomes). Yet, there is rarely a method which combines these two directions, i.e., using seed genes to guide the selection of active modules in a scored network. Such approaches are biologically rational since genes both interacting with known disease genes and having high association scores are very likely to be associated with disease risk.

We are conceiving such strategies either in the DMS framework or in the SigMod framework. In the DMS framework, we could first apply the DMS algorithm to generate raw modules (the number of raw modules to be generated is approximate to the number of genes in the network). Then modules are selected so that they contain most of the seed genes. This task is closely related to the well-known budgeted maximum coverage problem: given a collection of raw modules M generated by DMS, a number q as the maximum number of genes allowed to be selected, and a list of seed genes, choose a subset of raw modules $M' \subseteq M$ such that the number of unique genes included in M' is less or equal to q, and the number of seed genes covered by M' is maximized. Variations of this problem that takes into account the module score can also be considered, and are related to several extensions of the budgeted maximum coverage problem (Khuller et al., 1999). These problems can be readily solved by existing algorithms, which ensures the feasibility of our strategy.

In the SigMod framework, using seed genes to guide active module search can be done in a way by setting the indicator variable corresponding to the seed gene as 1 ($u_i = 1$) in the SigMod objective function $f(\mathbf{u}) = \mathbf{z}^T\mathbf{u} + \lambda\mathbf{u}^T\mathbf{A}\mathbf{u} - \eta \| \mathbf{u} \|_0$, so that the seed genes are selected by default. We have already figured out that, with this modification, the optimization problem has a similar form as that of SigMod, thus can be readily solved using the same algorithm as the one described in Chapter IV and detailed in Liu et al. (2017).

We foresee such seed-gene-guided active module search methods will help identify more genes and functional processes underlying a multifactorial disease such as asthma, and will improve the performance of network-based analysis as a whole. This is because our knowledge of disease-gene association is accumulating rapidly thanks to the advancement of genetic technologies and years of study efforts. These gene-disease links serve as an additional level of prior information, which can help to narrow down the genome-wide search space and focus on the network area of interest.

5 Conclusion

In this thesis, we developed several methods and strategies to facilitate network-based analysis of GWAS data. Application of them to the asthma GWAS data identified biological meaning processes and prioritized new candidate genes related to asthma. Our work revealed the value of performing network-based analysis in unveiling the genetic components underlying complex diseases. We foresee such methods will become more and more useful, especially in the post-GWAS era when a large amount of GWAS datasets are accumulated after years of research efforts, and when various types of biological network are comprehensively characterized.

We believe network-based analysis will further benefit from the methodological improvements with respect to the various steps involved in the analysis, including the use of more appropriate SNP to gene mapping strategy, to design of more powerful approaches to compute gene *p*-values, the choice of appropriate biological network fitting the disease context, and the development of analysis methods to mine useful network knowledge. Our thesis work has contributed to some of these aspects but there is still large space to progress.

Overall, we foresee with ever more GWAS and network data available, and with the right analysis strategies placed in the proper biological context, we can definitely understand the genetic mechanisms underlying asthma and many other complex diseases.

APPENDIX: RESUME DE LA THESE EN LANGUE FRANÇAISE

1 Introduction

Des efforts considérables ont été mis en œuvre pour caractériser les facteurs génétiques de l'asthme, notamment des études de gènes candidats, des études de liaison génétique et, plus récemment, des études d'association pan-génomiques (appelées « GWAS » dans la littérature anglo-saxonne). Bien que ces études aient permis d'identifier avec succès un certain nombre de loci associés à l'asthme, les facteurs génétiques identifiés, à ce jour, n'expliquent qu'une partie de la composante génétique de l'asthme

Les études d'association pan-génomiques ont principalement recherché des associations de la maladie avec des SNPs considérés individuellement sur l'ensemble du génome. Seuls les SNPs les plus fortement associés à la maladie et atteignant un seuil strict de significativité statistique qui prend en compte les tests multiples (classiquement, $p < 5\times10^{-8}$) sont rapportés. Les études d'association pan-génomiques manquent donc de puissance pour détecter des variants génétiques qui ont un effet marginal faible et qui agissent conjointement ou en interaction avec d'autres variants dans la susceptibilité génétique aux maladies. Pour compléter les analyses classiques simple-marqueurs, des analyses plus sophistiquées, qui intègrent les connaissances biologiques aux résultats de « GWAS », ont été proposées pour permettre de détecter un ensemble des gènes qui influencent conjointement le risque de maladie.

Parmi ces approches, les analyses basées sur les réseaux de gènes, qui intègrent les résultats de « GWAS » avec les réseaux d'interaction protéine-protéine (PPI), permettent d'identifier des modules de gènes (sous-réseaux) enrichis en signaux d'association avec la maladie. Le principe sous-tendant ces approches est le « guilt-by-association », qui stipule que les gènes (ou produits de ces gènes) qui sont connectés ont tendance à participer aux mêmes fonctions cellulaires ou mêmes processus biologiques. Par conséquent, les analyses basées sur les réseaux de gènes constituent une approche prometteuse pour identifier les gènes ayant des relations fonctionnelles qui ont un effet marginal faible mais agissent conjointement dans la susceptibilité à la maladie. Les principaux objectifs de cette thèse sont de développer des

méthodes d'analyse basée sur les réseaux de gènes et de les appliquer aux résultats d'études d'association pan-génomiques de l'asthme, afin d'identifier de nouveaux gènes candidats et des processus biologiques impliqués dans l'asthme.

2 Données du consortium GABRIEL sur la génétique de l'asthme utilisées dans cette thèse

Les données pan-génomiques de l'asthme utilisées dans cette thèse proviennent du consortium GABRIEL, une étude pluridisciplinaire pour identifier les facteurs génétiques et environnementaux de l'asthme dans la Communauté Européenne (financement européen).

Les données pan-génomiques du consortium GABRIEL concernaient un total de 10 365 cas d'asthme et 16 110 témoins provenant de 23 études. Ces sujets étaient tous d'origine européenne. Les détails de ces études sont fournis dans l'article de Moffatt et collègues (Moffatt et al., 2010). Les données sur les cas et les témoins proviennent d'études cliniques et d'études en population générale en Europe (études de cohortes et études transversales). Plusieurs études sont des études familiales, et des cas et témoins d'origine européenne ont également été recrutés par des études réalisées au Canada et en Australie. L'asthme a été défini comme un asthme survenu au cours de la vie et diagnostiqué par un médecin. L'asthme de l'enfant a été défini par l'apparition de l'asthme chez une personne de moins de 16 ans, et l'asthme de l'adulte a été défini par l'apparition de l'asthme à 16 ans ou après 16 ans.

Tous les sujets du consortium GABRIEL, à l'exception de ceux des études MRCA et MAGICS, ont été génotypés au Centre National de Génotypage (CNG, Evry, France) à l'aide de la puce Illumina Human610-Quad. Les sujets des études MRCA et MAGICS ont été génotypés avec les puces Illumina Sentrix Human-1 et Sentrix HumanHap300 BeadChips, dans le cadre du premier « GWAS » de l'asthme (Moffatt et al., 2007). Le contrôle de qualité des individus et des SNPs génotypés avec la puce Illumina 610K a suivi le même protocole. En résumé, les individus ont été retirés de l'analyse s'ils n'étaient pas de descendance européenne (basé sur l'analyse en composantes principales de chaque jeu de données avec les populations du projet HapMap) ou avaient plus de 5% de données génotypiques manquantes. Les SNPs ayant plus de 5% de données manquantes, une fréquence de l'allèle mineure (MAF) inférieure à 1% et/ou un test de l'équilibre de Hardy-Weinberg significatif au seuil de 10^{-4} ont été supprimés. Le contrôle de qualité des études MRCA et MAGICS a été détaillé dans l'article de Moffatt et collègues (Moffatt et al, 2007).

Dans chaque jeu de données, des imputations génotypiques ont été effectuées en utilisant le logiciel MACH 1.0 et les données HapMap Phase 2 (version 21) comme panel de référence.

Les SNP dont le score de qualité d'imputation (rsq) était supérieur à 0,5 et la fréquence de l'allèle mineur était supérieure à 0,01 ont été inclus dans les analyses, conduisant à un total de 2,370,689 SNPs analysés.

Les analyses d'association de l'asthme avec chacun des 2.37 millions de SNPs ont été réalisées, dans chaque étude, en utilisant un modèle de régression logistique qui incluait le dosage allélique pour chaque SNP et les composantes principales pour tenir compte des problèmes de stratification de population à l'aide du logiciel Stata version 10 (distributed by Stata Corporation, College Station, Texas, USA). Pour les études familiales, les dépendances familiales ont été prises en compte par une estimation robuste de la variance et en spécifiant l'option cluster dans Stata. Les résultats des analyses des « GWAS » de toutes les études GABRIEL sont hébergées au sein de l'unité UMR-946 (http://genestat.inserm.fr/fr/).

Dans cette thèse, nous nous sommes focalisés sur l'asthme de l'enfant, car il représente une entité plus homogène. Nous avons réparti aléatoirement les 18 études GABRIEL sur l'asthme de l'enfant en deux groupes de neuf études chacun, avec une taille d'échantillon totale similaire dans chaque groupe. Au total, 3 031 cas / 2 893 témoins étaient dans le premier groupe et 2 679 cas / 3 364 témoins dans le deuxième groupe. La répartition des « GWAS » de l'asthme de l'enfant en deux groupes a été motivée par la nécessité d'utiliser deux séries de résultats de GWAS pour les analyses de réseaux effectuées dans le cadre de cette thèse. Cela permettait de disposer de deux échantillons comme fichiers d'entrée pour l'analyse de réseau et ainsi de s'assurer de la cohérence des résultats obtenus par cette analyse à partir de deux échantillons.

Les méta-analyses des effets des SNPs sur l'asthme estimés dans chacune des études d'association pan-génomiques ont été réalisées dans chacun des deux groupes de 9 études et dans l'ensemble des 18 études. Nous avons utilisé des modèles à effets fixes et à effets aléatoires. Dans le modèle à effets aléatoires, l'estimation de la variance entre les études a utilisé la méthode de Der Simonian et Laird (Higgins et Thompson, 2002). Le test de l'effet combiné de chaque SNP sur l'asthme est basé sur le test de Wald. Les tests d'hétérogénéité entre les études dans chaque groupe sont basés sur le test Q de Cochran. Pour minimiser les résultats faussement positifs, nous avons examiné les effets des SNPs pour lesquels au moins deux tiers des études dans chaque groupe contribuaient à la méta-analyse (Moffatt et al., 2010). Etant donné que certaines régions du génome montraient une hétérogénéité

significative des effets des SNPs entre études, nous avons utilisé les valeurs de p obtenues à partir des méta-analyses effectuées avec des modèles à effets aléatoires. Les résultats des méta-analyses (*p*-valeurs des tests d'association pour chaque SNP) des deux groupes de 9 études ont été nommés META1 et META2, respectivement, et correspondent aux deux fichiers d'entrée pour les analyses de réseaux de gènes.

3 Résumé du premier travail de thèse

Mon premier projet de thèse, présenté dans le chapitre III, a consisté à étendre une méthode d'analyse de réseau de gènes, l'algorithme dmGWAS (Jia et al., 2011), et à l'appliquer aux résultats des méta-analyses de GWAS de l'asthme de l'enfant. Nous avons proposé une solution exacte de la méthode CGP « Circular Genomic Permutation », appelée fastCGP, pour calculer les valeurs de p au niveau du gène à partir des *p*-valeurs des SNPs assignés au gène. Nous avons aussi proposé une stratégie de recherche bidirectionnelle du module de gènes associé à l'asthme à partir des deux échantillons META1 et META2: les deux modules sélectionnées à partir de META1 et META2 devaient montrer une cohérence en terme de proportion de gènes partagés par ces modules; le module final était formé par l'intersection de ces deux modules.

L'application de cette méthodologie aux résultats des méta-analyses de « GWAS » de l'asthme de l'enfant (META1 et META2) a permis de détecter un module de 91 génes significativement associé à l'asthme de l'enfant ($p \leqslant 10^{-5}$, en effectuant 100 000 permutations des *p*-valeurs des SNPs sur le génome). Ce module a une structure intéressante, constituée d'un réseau central et de cinq réseaux périphériques. Parmi les 91 gènes appartenant au module sélectionné, 19 gènes ont une valeur de p significative à 5% dans les deux jeux de données META1 et META2. Ces 19 gènes incluaient 13 gènes situés au sein de 4 loci connus et rapportés par l'analyse pan-génomique du consortium GABRIEL (2q12, 5q31, 9p24.1, 17q12-q21; Moffatt et al, 2010), et six gènes à six nouveaux loci qui sont des candidats pertinents pour l'asthme : *CRMP1* (4p16.1), *ZNF192* (6p22 .1), *RAET1E* (6q24.3), *CTSL1* (9p21.33), *C12orf43* (12q24.31) et *JAK3* (19p13-p12). De plus, les gènes connus associés à l'asthme et appartenant tous au réseau central étaient liés au gène *APP* (codant pour la protéine précurseur bêta-amyloïde), gène qui prédispose aux formes familiales de la maladie d'Alzheimer. Cette maladie a une composante inflammatoire et il existe un faisceau d'arguments d'ordre épidémiologique, génétique, épigénétique et même thérapeutique qui suggèrent des liens entre asthme et maladie d'Alzheimer. La connexion observée entre *APP* et gènes associés à l'asthme dans notre module renforce l'hypothèse que ces deux pathologies partageraient des mécanismes génétiques communs. Par ailleurs, l'analyse fonctionnelle des gènes du module issu de notre analyse de réseau à l'aide de DAVID (Huang et al., 2007) a révélé quatre clusters fonctionnels de gènes qui correspondent aux fonctions suivantes:

immunité innée et adaptative, chimiotaxie, adhésion cellulaire et régulation de la transcription. Au total, cette étude a permis de mettre en évidence de nouveaux gènes candidats pour l'asthme et d'apporter un nouvel éclairage sur les relations fonctionnelles entre gènes de susceptibilité à l'asthme de l'enfant.

4 Résume du deuxième travail de thèse

Mon deuxième projet de thèse, présenté au chapitre IV, a consisté à développer une nouvelle méthode de réseau de gènes, nommée SigMod, qui a le potentiel d'améliorer la performance des analyses de réseaux, en général. SigMod vise à sélectionner un module de gènes enrichis en signaux d'association avec la maladie et montrant de fortes inter-connexions. Par rapport aux méthodes d'analyse de réseau précédemment proposées, SigMod a plusieurs avantages, notamment la robustesse au bruit de fond, la capacité de prendre en compte une pondération sur les liens entre gènes (ou produits de gènes) selon le type d'information utilisé, et rend les résultats plus facilement interprétables.

Nous avons d'abord évalué la performance de SigMod sur des données simulées. Nous avons utilisé un réseau de gènes issu de la base de données STRING (présenté dans la section 3.3.1 du chapitre I), qui contient divers types d'informations sur les relations entre gènes et protéines, y compris des interactions directes (physiques) et indirectes (fonctionnelles). A chaque lien entre deux gènes dans le réseau est associé un poids variant de 0 à 1, qui représente la force de la relation entre ces gènes, estimée à partir de multiples sources d'information. Nous avons effectué cinq jeux de simulations (20 répétitions par jeu de simulation), en choisissant, pour chaque jeu de simulations, un module de gènes fortement interconnectés identifié dans STRING en utilisant CFinder (Adamcsek et al., 2006) ; chacun des modules représentait un module causal. Les valeurs p des gènes appartenant à un module causal ont été uniformément réparties entre 0 et 10^{-3} (représentant ainsi des signaux d'association avec la maladie), alors que les valeurs de p d'autres gènes du réseau ont été uniformément réparties entre 0 et 1 (représentant le bruit). Les analyses effectuées sur ces données simulées ont montré que SigMod a la meilleure puissance et le taux de faux positifs le plus bas, comparés à deux autres méthodes récemment proposées—dmGWAS (Jia et al., 2011) et SConES (Azencott et al., 2013). Cette bonne performance de SigMod a été préservée lorsque du bruit supplémentaire a été ajouté à la fois aux valeurs de p du gène et au réseau de gènes issue de STRING, démontrant ainsi la robustesse de SigMod.

Nous avons ensuite appliqué SigMod aux résultats des méta-analyses de « GWAS » de l'asthme de l'enfant (META1 et META2). En utilisant META1 comme jeu de données de découverte, nous avons identifié un module de gènes enrichis en signaux d'association avec l'asthme et composé de 190 gènes pertinents sur le plan biologique. Tous ces gènes ont une

valeur de p significatives au seuil nominal de 5%, ce qui montre la capacité de SigMod à identifier des gènes avec un score élevé. Lorsque ce module a été évalué en utilisant le jeu de données META2, 30 gènes étaient significatifs au seuil de 5% et, donc, dans les deux jeux de données. L'analyse fonctionnelle des gènes du module associé à l'asthme de l'enfant à l'aide de DAVID (Huang et al. 2009) et l'analyse de pathways utilisant la base de données KEGG ont permis de mettre en évidence neuf clusters de gènes ayant des relations fonctionnelles et 15 pathways enrichis en gènes du module; plusieurs de ces pathways sont en lien avec la réponse aux infections virales, qui sont récemment apparus comme jouant un rôle de plus en plus important dans l'asthme. Les clusters de gènes et les pathways correspondent à des processus biologiques connus pour être impliqués dans l'asthme, ainsi qu'à des processus nouveaux qui méritent d'être étudiés de manière plus approfondie pour mieux comprendre leur rôle dans l'asthme.

5 Discussion et conclusion

Les études d'association pan-génomiques (GWAS) de maladies multifactorielles, comme l'asthme, ont identifié de nombreuses variants génétiques associées à ces maladies, mais ces variants n'expliquent qu'une partie de la composante génétique de ces maladies. Les limites des méthodes actuelles des études d'association pan-génomiques nous ont motivés à explorer d'autres approches qui peuvent être considérées comme complémentaires de l'analyse classique simple-marqueur.

Comme indiqué dans l'introduction de cette thèse, notre objectif principal était d'explorer des stratégies d'analyse basées sur les réseaux de gènes qui combinent les résultats de « GWAS » avec des réseaux biologiques entre gènes (ou produits de gènes) issus de bases de données pour identifier des modules de gènes associés à la maladie. Nous avons développé plusieurs méthodes et stratégies pour faciliter cette analyse de réseau. L'application de ces méthodes aux données GWAS de l'asthme ont permis d'identifier des processus biologiques pertinents et ont mis en évidence de nouveaux gènes candidats pour l'asthme. Notre travail a montré l'intérêt d'effectuer des analyses de réseaux de gènes afin de révéler des ensembles de gènes influençant conjointement le risque de maladie. Nous anticipons que ce type d'approches sera de plus en plus utilisé, en particulier dans l'ère actuelle post-GWAS, alors que des résultats de GWAS se sont accumulés au cours des dernières années et que les réseaux biologiques sont de mieux en mieux caractérisés.

Nous pensons que l'analyse basée sur les réseaux de gènes pourra bénéficier de nouvelles améliorations sur le plan méthodologique en ce qui concerne les différentes étapes impliquées dans l'analyse, y compris l'utilisation de stratégies plus appropriées pour assigner les SNPs aux gènes, d'approches plus puissantes pour calculer les valeurs de P des gènes, le choix du réseau biologique adapté au contexte de la maladie et le développement de méthodes d'analyse pour extraire de manière optimale les informations à partir du réseau. Notre travail de thèse a contribué à certains de ces aspects, mais de nouvelles extensions restent à faire.

Au total, nous anticipons qu'avec une quantité croissante de données disponibles à la fois sur les « GWAS » et les réseaux biologiques, et avec des stratégies d'analyse appropriées et adaptées aux connaissances biologiques, il sera possible de progresser rapidement dans la

connaissance des mécanismes génétiques impliqués dans l'asthme et dans de nombreuses autres maladies complexes.

REFERENCES

Abdulrazzaq, Y. M., Bener, A., & DeBuse, P. (1994). Association of allergic symptoms in children with those in their parents. *Allergy*, *49*(9), 737-743.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. [10.1038/nprot.2010.116]. *Nat. Protocols*, *5*(9), 1564-1573. doi: http://www.nature.com/nprot/journal/v5/n9/abs/nprot.2010.116.html#supplementary-information

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., and others. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25-29.

Askland, K., Read, C., O'Connell, C., & Moore, J. H. (2012). Ion channels and schizophrenia: a gene set-based analytic approach to GWAS data for biological hypothesis testing. *Human Genetics*, *131*(3), 373-391. doi: 10.1007/s00439-011-1082-x

Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y., & Borgwardt, K. M. (2013). Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, *29*(13), i171-i179.

Backes, C., Rurainski, A., Klau, G. W., Müller, O., Stöckel, D., Gerasch, A., and others. (2012). An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic acids research*, *40*(6), e43-e43.

Bakir-Gungor, B., Egemen, E., & Sezerman, O. U. (2014). PANOGA: a web-server for identification of SNP targeted pathways from genome-wide association study data. *Bioinformatics*, btt743.

Bakshi, A., Zhu, Z., Vinkhuyzen, A. A., Hill, W. D., McRae, A. F., Visscher, P. M., & Yang, J. (2016). Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Scientific reports*, *6*, 32894.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*(10), 781-791.

Ballard, D. H., Aporntewan, C., Lee, J. Y., Lee, J. S., Wu, Z., & Zhao, H. (2009, 2009). *A pathway analysis applied to Genetic Analysis Workshop 16 genome-wide rheumatoid arthritis data*.

Bardana, E. J. (2008). 10. Occupational asthma. *Journal of Allergy and Clinical Immunology*, *121*(2), S408-S411.

Barrett, J. H., Taylor, J. C., Bright, C., Harland, M., Dunning, A. M., Akslen, L. A., and others. (2015). Fine mapping of genetic susceptibility loci for melanoma reveals a

mixture of single variant and multiple variant regions. *International journal of cancer, 136*(6), 1351-1360.

Barshir, R., Basha, O., Eluk, A., Smoly, I. Y., Lan, A., & Yeger-Lotem, E. (2012). The TissueNet database of human tissue protein–protein interactions. *Nucleic Acids Research, 41*(D1), D841-D844.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., and others. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics, 25*(8), 1091-1093.

Bønnelykke, K., Sleiman, P., Nielsen, K., Kreiner-Møller, E., Mercader, J. M., Belgrave, D., and others. (2014). A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nature genetics, 46*(1), 51-55.

Bouzigon, E., Corda, E., Aschard, H., Dizier, M.-H., Boland, A., Bousquet, J., and others (2008). Effect of 17q21 Variants and Smoking Exposure in Early-Onset Asthma. *New England Journal of Medicine, 359*(19), 1985-1994. doi: 10.1056/NEJMoa0806604

Bouzigon, E., Forabosco, P., Koppelman, G. H., Cookson, W. O., Dizier, M.-H., Duffy, D. L., and others (2010). Meta-analysis of 20 genome-wide linkage studies evidenced new regions linked to asthma and atopy. *European Journal of Human Genetics, 18*(6), 700.

Bøvelstad, H. M., Nyg\a ard, S. a. a., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., & Lingj\a erde, O. C. (2007). Predicting survival from microarray data—a comparative study. *Bioinformatics, 23*(16), 2080-2087.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brossard, M. e. a. (2013). [Comparison of permutations strategies to assess gene-set significance in gene-set-enrichment analysis].

Brückner, A., Polge, C., Lentze, N., Auerbach, D., & Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences, 10*(6), 2763-2788.

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., & Sabatti, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics, 205*(1), 61-75.

Bush, W. S., & Moore, J. H. (2012). Genome-wide association studies. *PLoS Comput Biol, 8*(12), e1002822.

Cabrera, C. P., Navarro, P., Huffman, J. E., Wright, A. F., Hayward, C., Campbell, H., and others (2012). Uncovering networks from genome-wide association studies via

circular genomic permutation. *G3 (Bethesda, Md.), 2*(9), 1067-1075. doi: 10.1534/g3.112.002618

Cattaert, T., Calle, M. L., Dudek, S. M., Mahachie John, J. M., Van Lishout, F., Urrea, V., and others (2011). Model ‐ Based Multifactor Dimensionality Reduction for detecting epistasis in case ‐ control data in the presence of noise. *Annals of human genetics, 75*(1), 78-89.

Chambers, E. V., Bickmore, W. A., & Semple, C. A. (2013). Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS computational biology, 9*(4), e1003017. doi: 10.1371/journal.pcbi.1003017

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., and others. (2007). Replicating genotype–phenotype associations. *Nature, 447*(7145), 655-660.

Chapman, J. M., Cooper, J. D., Todd, J. A., & Clayton, D. G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human heredity, 56*(1-3), 18-31.

Chen, H., Meigs, J. B., & Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology, 37*(2), 196-204.

Chen, S.-H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., and others. (2008). A support vector machine approach for detecting gene-gene interaction. *Genetic epidemiology, 32*(2), 152-167.

Chen, X., Wang, L., Hu, B., Guo, M., Barnard, J., & Zhu, X. (2010). Pathway-based analysis for genome-wide association studies using supervised principal components. *Genetic epidemiology, 34*(7), 716-724.

Chen, Y., Dales, R., Tang, M., & Krewski, D. (2002). Obesity may increase the incidence of asthma in women but not in men: longitudinal observations from the Canadian National Population Health Surveys. *American Journal of Epidemiology, 155*(3), 191-197.

Chimusa, E. R., Mbiyavanga, M., Mazandu, G. K., & Mulder, N. J. (2015). ancGWAS: a Post Genome-wide Association Study Method for Interaction, Pathway, and Ancestry Analysis in Homogeneous and Admixed Populations. *Bioinformatics*, btv619.

Chung, R.-H., & Chen, Y.-E. (2012). A two-stage random forest-based pathway analysis method. *PLOS ONE, 7*(5), e36662.

Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., and others (2011). PINA v2. 0: mining interactome modules. *Nucleic Acids Research*, gkr967.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., and others. (2013). The Reactome pathway knowledgebase. *Nucleic Acids Research, 42*(D1), D472-D477.

Curtis, D., Vine, A. E., & Knight, J. (2008). A simple method for assessing the strength of evidence for association at the level of the whole gene. *Advances and applications in bioinformatics and chemistry: AABC, 1,* 115-120.

De Vlaming, R., & Groenen, P. J. F. (2015). The current and future use of ridge regression for prediction in quantitative genetics. *BioMed research international, 2015*.

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics, 55*(4), 997-1004.

Ding, L., Abebe, T., Beyene, J., Wilke, R. A., Goldberg, A., Woo, J. G., and others. (2013). Rank-based genome-wide analysis reveals the association of ryanodine receptor-2 gene variants with childhood asthma among human populations. *Human genomics, 7*(1), 16.

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., & Müller, T. (2008). Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics, 24*(13), i223-i231.

Dizier, M.-H., Demenais, F., & Mathieu, F. (2017). Gain of power of the general regression model compared to Cochran-Armitage Trend tests: simulation study and application to bipolar disorder. *BMC genetics, 18*(1), 24. doi: 10.1186/s12863-017-0486-6

Dong, X., Hao, Y., Wang, X., & Tian, W. (2016). LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Scientific reports, 6*.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., and others (2007). A systems biology approach for pathway level analysis. *Genome research, 17*(10), 1537-1545.

Dudbridge, F., & Koeleman, B. P. C. (2003). Rank truncated product of P-values, with application to genomewide association scans. *Genetic Epidemiology, 25*(4), 360-366.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics, 10*(1), 48.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics, 11*(6), 446-450.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., and others (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics, 23*(22), 5866-5878.

Fang, Z., Tian, W., & Ji, H. (2012). A network-based gene-weighting approach for pathway analysis. *Cell research, 22*(3), 565-580.

Fernández-Suárez, X. M., Rigden, D. J., & Galperin, M. Y. (2013). The 2014 nucleic acids research database issue and an updated NAR online molecular biology database collection. *Nucleic Acids Research, 42*(D1), D1-D6.

Ferreira, M. A. R., Matheson, M. C., Duffy, D. L., Marks, G. B., Hui, J., Le Souëf, P., and others. (2011). Identification of IL6R and chromosome 11q13. 5 as risk loci for asthma. *The Lancet, 378*(9795), 1006-1014.

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society, 85*(1), 87-94.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., and others. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*(7164), 851-861.

Galanter, J. M., Gignoux, C. R., Torgerson, D. G., Roth, L. A., Eng, C., Oh, S. S., and others. (2014). Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. *Journal of Allergy and Clinical Immunology, 134*(2), 295-305.

Garcia, O., Saveanu, C., Cline, M., Fromont-Racine, M., Jacquier, A., Schwikowski, B., & Aittokallio, T. (2006). GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics, 23*(3), 394-396.

Genin, E., Feingold, J., & Clerget-Darpoux, F. (2008). Identifying modifier genes of monogenic disease: strategies and difficulties. *Human genetics, 124*(4), 357.

Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., and others. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology, 5*(10), R80.

Gergen, P. J., Fowler, J. A., Maurer, K. R., Davis, W. W., & Overpeck, M. D. (1998). The burden of environmental tobacco smoke exposure on the respiratory health of children 2 months through 5 years of age in the United States: Third National Health and Nutrition Examination Survey, 1988 to 1994. *Pediatrics, 101*(2), e8-e8.

Gilliland, F. D., Berhane, K., Islam, T., McConnell, R., Gauderman, W. J., Gilliland, S. S., and others (2003). Obesity and the risk of newly diagnosed asthma in school-age children. *American journal of epidemiology, 158*(5), 406-415.

Gilliland, F. D., Li, Y.-F., & Peters, J. M. (2001). Effects of maternal smoking during pregnancy and environmental tobacco smoke on asthma and wheezing in children. *American journal of respiratory and critical care medicine, 163*(2), 429-436.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., & Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics, 28*(18), i451-i457.

Gold, D. R. (2000). Environmental tobacco smoke, indoor allergens, and childhood asthma. *Environmental health perspectives, 108*(Suppl 4), 643.

Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC genetics, 11*(1), 49.

Guo, X., & Wang, X.-F. (2009). Signaling cross-talk between TGF-β/BMP and other pathways *Cell research* (Vol. 19, pp. 71-88).

Guo, Y.-F., Li, J., Chen, Y., Zhang, L.-S., & Deng, H.-W. (2009). A new permutation strategy of pathway-based approach for genome-wide association study. *BMC bioinformatics, 10*, 429. doi: 10.1186/1471-2105-10-429

Gwinner, F., Boulday, G., Vandiedonck, C., Arnould, M., Cardoso, C., Nikolayeva, I., and others. (2016). Network-based analysis of omics data: The LEAN method. *Bioinformatics*, btw676.

Hahn, L. W., Ritchie, M. D., & Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics, 19*(3), 376-382. doi: 10.1093/bioinformatics/btf869

Halapi, E., & Hakonarson, H. (2004). Recent development in genomic and proteomic research for asthma. *Current opinion in pulmonary medicine, 10*(1), 22-30.

Han, S., Yang, B.-Z., Kranzler, H. R., Liu, X., Zhao, H., Farrer, L. A., and others (2013). Integrating GWASs and human protein interaction networks identifies a gene subnetwork underlying alcohol dependence. *American Journal of Human Genetics, 93*(6), 1027-1034. doi: 10.1016/j.ajhg.2013.10.021

Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics, 48*(10), 1279-1283.

He, X., & Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLOS Genet, 2*(6), e88.

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta ‐ analysis. *Statistics in medicine, 21*(11), 1539-1558.

Hill, W., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics, 38*(6), 226-231.

Hillenmeyer, S., Davis, L. K., Gamazon, E. R., Cook, E. H., Cox, N. J., & Altman, R. B. (2016). STAMS: STRING-Assisted Module Search for Genome Wide Association Studies and Application to Autism. *Bioinformatics*, btw530. doi: 10.1093/bioinformatics/btw530

Himes, B. E., Hunninghake, G. M., Baurley, J. W., Rafaels, N. M., Sleiman, P., Strachan, D. P., and others. (2009). Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. *The American Journal of Human Genetics, 84*(5), 581-593.

Hirota, T., Takahashi, A., Kubo, M., Tsunoda, T., Tomita, K., Doi, S., and others. (2011). Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. *Nature genetics, 43*(9), 893-896.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics, 6*(2), 95-108.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature methods, 10*(11), 1108-1115.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLOS Genet, 4*(7), e1000130.

Holberg, C. J., Morgan, W. J., Wright, A. L., & Martinez, F. D. (1998). Differences in familial segregation of FEV1 between asthmatic and nonasthmatic families: role of a maternal component. *American journal of respiratory and critical care medicine, 158*(1), 162-169.

Holden, M., Deng, S., Wojnowski, L., & Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics, 24*(23), 2784-2785.

Hong, M.-G., Pawitan, Y., Magnusson, P. K. E., & Prince, J. A. (2009). Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Human Genetics, 126*(2), 289-301. doi: 10.1007/s00439-009-0676-z

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44-57. doi: 10.1038/nprot.2008.211

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., and others (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology, 8*, R183. doi: 10.1186/gb-2007-8-9-r183

Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics, 18*(suppl 1), S233-S240.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., & Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics, 92*(6), 841-853.

Jannot, A.-S., Ehret, G., & Perneger, T. (2015). P< 5$\times$ 10- 8 has emerged as a standard of statistical significance for genome-wide association studies. *Journal of clinical epidemiology, 68*(4), 460-465.

Jia, P., Wang, L., Fanous, A. H., Pato, C. N., Edwards, T. L., Consortium, T. I. S., & Zhao, Z. (2012). Network-Assisted Investigation of Combined Causal Signals from Genome-Wide Association Studies in Schizophrenia. *PLOS Computational Biology, 8*(7), e1002587. doi: 10.1371/journal.pcbi.1002587

Jia, P., Zheng, S., Long, J., Zheng, W., & Zhao, Z. (2011). dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics (Oxford, England), 27*(1), 95-102. doi: 10.1093/bioinformatics/btq615

Jin, L., Zuo, X.-Y., Su, W.-Y., Zhao, X.-L., Yuan, M.-Q., Han, L.-Z., and others (2014). Pathway-based analysis tools for complex diseases: a review. *Genomics, Proteomics & Bioinformatics, 12*(5), 210-220. doi: 10.1016/j.gpb.2014.10.002

Jorde, L. B., & Wooding, S. P. (2004). Genetic variation, classification and'race'. *Nature genetics, 36*, S28-S33.

Junker, B. H., & Schreiber, F. (2011). *Analysis of biological networks* (Vol. 2): John Wiley & Sons.

Kabesch, M. (2005). Candidate gene association studies and evidence for gene-by-gene interactions. *Immunology and allergy clinics of North America, 25*(4), 681-708.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2009). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research, 38*(suppl_1), D355-D360.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research, 44*(D1), D457-462. doi: 10.1093/nar/gkv1070

Kelly, W. P., & Stumpf, M. P. (2012). Assessing coverage of protein interaction data using capture–recapture models. *Bulletin of mathematical biology, 74*(2), 356-374.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., and others (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science (New York, N.Y.), 245*(4922), 1073-1080.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Comput Biol, 8*(2), e1002375.

Khoury, M. J., & Yang, Q. (1998). The future of genetic studies of complex human diseases: an epidemiologic perspective. *Epidemiology, 9*(3), 350-354.

Khuller, S., Moss, A., & Naor, J. S. (1999). The budgeted maximum coverage problem. *Information Processing Letters, 70*(1), 39-45.

Kim, J., Sohn, I., Kim, D. D. H., & Jung, S.-H. (2013). SNP selection in genome-wide association studies via penalized support vector machine with MAX test. *Computational and mathematical methods in medicine, 2013*.

King, O. D., Lee, J. C., Dudley, A. M., Janse, D. M., Church, G. M., & Roth, F. P. (2003). Predicting phenotype from patterns of annotation. *Bioinformatics, 19*(suppl 1), i183-i189.

Klammer, M., Godl, K., Tebbe, A., & Schaab, C. (2010). Identifying differentially regulated subnetworks from phosphoproteomic data. *BMC bioinformatics, 11*(1), 351.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., and others. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308*(5720), 385-389.

Köhler, S., Bauer, S., Horn, D., & Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics, 82*(4), 949-958.

Kotlyar, M., Pastrello, C., Sheahan, N., & Jurisica, I. (2016). Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Research, 44*(D1), D536-D541.

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., & Bergmann, S. (2016). Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLOS Comput Biol, 12*(1), e1004714. doi: 10.1371/journal.pcbi.1004714

Lasky-Su, J., Himes, B. E., Raby, B. A., Klanderman, B. J., Sylvia, J. S., Lange, C., and others. (2012). HLA-DQ strikes again: Genome-wide association study further confirms HLA-DQ in the diagnosis of asthma among adults. *Clinical & Experimental Allergy, 42*(12), 1724-1733.

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., & Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research, 21*(7), 1109-1121.

Lee, S., Kong, S., & Xing, E. P. (2016). A network-driven approach for genome-wide association mapping. *Bioinformatics, 32*(12), i164-i173.

Leeuw, C. A. d., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology, 11*(4), e1004219. doi: 10.1371/journal.pcbi.1004219

Lehne, B., Lewis, C. M., & Schlitt, T. (2011). From SNPs to genes: disease association at the gene level. *PloS one, 6*(6), e20133.

Lehne, B., & Schlitt, T. (2009). Protein-protein interaction databases: keeping up with growing interactomes. *Human genomics, 3*(3), 291.

Lettre, G., Lange, C., & Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic epidemiology, 31*(4), 358-362.

Levy, H., Raby, B. A., Lake, S., Tantisira, K. G., Kwiatkowski, D., Lazarus, R., and others. (2005). Association of defensin β-1 gene polymorphisms with asthma. *Journal of allergy and clinical immunology, 115*(2), 252-258.

Li, B., Zhang, Y., Yu, Y., Wang, P., Wang, Y., Wang, Z., & Wang, Y. (2015). Quantitative assessment of gene expression network module-validation methods. *Scientific reports, 5*.

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity, 95*(3), 221-227.

Li, J., Malley, J. D., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2016). Detecting gene-gene interactions using a permutation-based random forest method. *BioData mining, 9*(1), 14.

Li, M.-X., Gui, H.-S., Kwan, J. S. H., & Sham, P. C. (2011). GATES: a rapid and powerful gene-based association test using extended Simes procedure. *The American Journal of Human Genetics, 88*(3), 283-293.

Li, Y., Agarwal, P., & Rajagopalan, D. (2008). A global pathway crosstalk network. *Bioinformatics, 24*(12), 1442-1447.

Li, Z.-C., Huang, M.-H., Zhong, W.-Q., Liu, Z.-Q., Xie, Y., Dai, Z., & Zou, X.-Y. (2016). Identification of drug-target interaction from interactome network with 'guilt-by-association' principle and topology features. *Bioinformatics (Oxford, England), 32*(7), 1057-1064. doi: 10.1093/bioinformatics/btv695

Lin, H.-Y., Chen, D.-T., Huang, P.-Y., Liu, Y.-H., Ochoa, A., Zabaleta, J., and others (2016). SNP interaction pattern identifier (SIPI): an intensive search for SNP–SNP interaction patterns. *Bioinformatics, 33*(6), 822-833.

Lin, P.-L., Yu, Y.-W., & Chung, R.-H. (2016). Pathway analysis incorporating protein-protein interaction networks identified candidate pathways for the seven common diseases. *PLOS ONE, 11*(9), e0162910.

Listgarten, J., Damaraju, S., Poulin, B., Cook, L., Dufour, J., Driga, A., and others (2004). Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clinical cancer research, 10*(8), 2725-2737.

Litonjua, A. A., Carey, V. J., Burge, H. A., Weiss, S. T., & Gold, D. R. (1998). Parental history and the risk for childhood asthma: does mother confer more risk than father? *American journal of respiratory and critical care medicine, 158*(1), 176-181.

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., and others (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. *American Journal of Human Genetics, 87*(1), 139-145. doi: 10.1016/j.ajhg.2010.06.009

Liu, Y., Zhou, J., Liu, Z., Chen, L., & Ng, M. K. (2012). Construction and analysis of genome-wide SNP networks. *2012 IEEE 6th International Conference on Systems Biology (ISB)*, 327-332. doi: 10.1109/isb.2012.6314158

Los, H., Koppelman, G., & Postma, D. (1999). The importance of genetic influences in asthma. *European Respiratory Journal, 14*(5), 1210-1227.

Lunetta, K. L., Hayward, L. B., Segal, J., & Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics, 5*(1), 32.

Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C. I., & Xiong, M. (2010). Genome-wide gene and pathway analysis. *European Journal of Human Genetics, 18*(9), 1045-1053.

Ma, H., Schadt, E. E., Kaplan, L. M., & Zhao, H. (2011). COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics, 27*(9), 1290-1298.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., and others. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research, 45*(D1), D896-D901.

Macintyre, G., Yepes, A. J., Ong, C. S., & Verspoor, K. (2014). Associating disease-related genetic variants in intergenic regions to the genes they impact. *PeerJ, 2*, e639.

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics, 21*(16), 3448-3449.

Mägi, R., & Morris, A. P. (2010). GWAMA: software for genome-wide association meta-analysis. *BMC bioinformatics, 11*(1), 288.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., and others. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753.

March, M. E., Sleiman, P. M., & Hakonarson, H. (2013). Genetic polymorphisms and associated susceptibility to asthma. *International Journal of General Medicine, 6*, 253.

Martinez, F. D., & Vercelli, D. (2013). Asthma. *The Lancet, 382*(9901), 1360-1372. doi: 10.1016/s0140-6736(13)61536-6

Maslov, S., & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science, 296*(5569), 910-913.

Mbiyavanga, M. (2014). *Network-based approach for post genome-wide association study analysis in admixed populations.* University of Cape Town.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., and others (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology, 28*(5), 495-501.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science, 347*(6224), 1257601.

Merico, D., Isserlin, R., Stueker, O., Emili, A., & Bader, G. D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLOS ONE, 5*(11), e13984.

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols, 8*(8), 1551-1566.

Mishra, A., & Macgregor, S. (2015). VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Research and Human Genetics, 18*(1), 86-91. doi: 10.1017/thg.2014.79

Mitra, K., Carvunis, A.-R., Ramesh, S. K., & Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics, 14*(10), 719-732.

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., and others (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology, 4*, 278.

Moffatt, M. F., Gut, I. G., Demenais, F., Strachan, D. P., Bouzigon, E., Heath, S., and others. (2010). A large-scale, consortium-based genomewide association study of asthma. *New England Journal of Medicine, 363*(13), 1211-1221.

Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., and others. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature, 448*(7152), 470-473.

Moorman, J. E., Zahran, H., Truman, B. I., Molla, M. T., Control, C. f. D., & Prevention. (2011). Current asthma prevalence-United States, 2006-2008. *MMWR Surveill Summ, 60*(Suppl), 84-86.

Mott, R., Yuan, W., Kaisaki, P., Gan, X., Cleak, J., Edwards, A., and others (2014). The architecture of parent-of-origin effects in mice. *Cell, 156*(1-2), 332-342. doi: 10.1016/j.cell.2013.11.043

Nacher, J. C., Hayashida, M., & Akutsu, T. (2009). Emergence of scale-free distribution in protein–protein interaction networks based on random selection of interacting domain pairs. *Biosystems, 95*(2), 155-159.

Nacu, Ş., Critchley-Thorne, R., Lee, P., & Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics, 23*(7), 850-858.

National Asthma Education Prevention Program. (2007). Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma-Summary Report 2007. *The Journal of allergy and clinical immunology, 120*(5 Suppl), S94.

Neale, B. M., & Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. *The American Journal of Human Genetics, 75*(3), 353-362.

Nibbe, R. K., Koyutürk, M., & Chance, M. R. (2010). An integrative-omics approach to identify functional sub-networks in human colorectal cancer. *PLOS Comput Biol, 6*(1), e1000639.

Nijman, S. (2011). Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS letters, 585*(1), 1-6.

Noguchi, E., Sakamoto, H., Hirota, T., Ochiai, K., Imoto, Y., Sakashita, M., and others. (2011). Genome-wide association study identifies HLA-DP as a susceptibility gene for pediatric asthma in Asian populations. *PLOS Genet, 7*(7), e1002170.

Ober, C., & Vercelli, D. (2011). Gene–environment interactions in human disease: nuisance or opportunity? *Trends in genetics, 27*(3), 107-115.

Ogris, C., Helleday, T., & Sonnhammer, E. L. L. (2016). PathwAX: a web server for network crosstalk based pathway annotation. *Nucleic Acids Research, 44*(W1), W105-W109.

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012, 2012). *Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions*.

Oh, S., Lee, J., Kwon, M.-S., Weir, B., Ha, K., & Park, T. (2012). A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. *BMC bioinformatics, 13*(9), S5.

Okada, H., Kuhn, C., Feillet, H., & Bach, J. F. (2010). The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. *Clinical & Experimental Immunology, 160*(1), 1-9.

Oliver, S. (2000). Proteomics: Guilt-by-association goes global. *Nature, 403*(6770), 601-603. doi: 10.1038/35001165

Oualkacha, K., Dastani, Z., Li, R., Cingolani, P. E., Spector, T. D., Hammond, C. J., and others (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic Epidemiology*, *37*(4), 366-376.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814-818.

Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T., & Iliopoulos, I. (2015). Protein–protein interaction predictions using text mining methods. *Methods*, *74*, 47-53.

Patnala, R., Clements, J., & Batra, J. (2013). Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC genetics*, *14*(1), 39.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLOS Genet*, *2*(12), e190.

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., & Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, *29*(21), 2757-2764. doi: 10.1093/bioinformatics/btt471

Pearson, E. S. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, *30*(1/2), 134-148.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., and others. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics*, *18*(1), 111-117.

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., and others (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications*, *6*, 5890.

Petsko, G. A. (2009). Guilt by association. *Genome Biology*, *10*, 104. doi: 10.1186/gb-2009-10-4-104

Piovesan, D., Giollo, M., Ferrari, C., & Tosatto, S. C. E. (2015). Protein function prediction using guilty by association from interaction networks. *Amino Acids*, *47*(12), 2583-2592. doi: 10.1007/s00726-015-2049-3

Porollo, A., & Meller, J. l. (2007). Prediction-based fingerprints of protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, *66*(3), 630-645.

Prabhu, S., & Pe'er, I. (2012). Ultrafast genome-wide scan for SNP–SNP interactions in common complex disease. *Genome research*, *22*(11), 2230-2240.

Pulst, S. M. (1999). Genetic linkage analysis. *Archives of neurology*, *56*(6), 667-672.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., and others. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*(3), 559-575.

Qi, Y., Suhail, Y., Lin, Y.-y., Boeke, J. D., & Bader, J. S. (2008). Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome research, 18*(12), 1991-2004.

Qiu, Y.-Q., Zhang, S., & Zhang, X.-S. (2008, 2008). *Uncovering differentially expressed pathways with protein interaction and gene expression data*.

R development core team. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*.

Rajagopalan, D., & Agarwal, P. (2004). Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics, 21*(6), 788-793.

Ramasamy, A., Kuokkanen, M., Vedantam, S., Gajdos, Z. K., Alves, A. C., Lyon, H. N., and others. (2012). Genome-wide association studies of asthma in population-based cohorts confirm known and suggested loci and identify an additional association near HLA. *PLOS ONE, 7*(9), e44008.

Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. (2014). Protein-protein interaction detection: methods and analysis. *International journal of proteomics, 2014*.

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., & Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in medicine, 34*(28), 3769-3792.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., & Vilo, J. (2016). g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, gkw199.

Reiss, D. J., Baliga, N. S., & Bonneau, R. (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC bioinformatics, 7*(1), 280.

Riordan, J. R., Rommens, J. M., Kerem, B.-s., Alon, N., Rozmahel, R., & others. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science, 245*(4922), 1066.

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273*(5281), 1516-1517.

Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K., & Hakonarson, H. (2011). Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic acids research, 39*(9), e62-e62.

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., and others. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet, 7*(1), e1001273.

Ruano, D., Abecasis, G. R., Glaser, B., Lips, E. S., Cornelisse, L. N., de Jong, A. P. H., and others (2010). Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. *American Journal of Human Genetics, 86*(2), 113-125. doi: 10.1016/j.ajhg.2009.12.006

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., and others (2012). A travel guide to Cytoscape plugins. *Nature Methods, 9*(11), 1069-1076.

Saxena, M., Williams, S., Taskén, K., & Mustelin, T. (1999). Crosstalk between cAMP-dependent kinase and MAP kinase through a protein tyrosine phosphatase. *Nature cell biology, 1*(5), 305-310.

Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., and others. (2013). All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLOS Genet, 9*(4), e1003449.

Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development, 19*(3), 212-219.

Schwartz, J. (2004). Air pollution and children's health. *Pediatrics, 113*(4 Suppl), 1037-1043.

Schwarz, D. F., König, I. R., & Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics, 26*(14), 1752-1758. doi: 10.1093/bioinformatics/btq257

Schwarz, D. F., Szymczak, S., Ziegler, A., & König, I. R. (2007, 2007). *Picking single-nucleotide polymorphisms in forests*.

Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnology, 18*(12), 1257-1261.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., and others (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research, 13*(11), 2498-2504.

Shen, X., & Carlborg, Ö. (2013). Beware of risk for increased false positive rates in genome-wide association studies for phenotypic variability. *Frontiers in genetics, 4*, 93.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research, 29*(1), 308-311.

Singh, R., Park, D., Xu, J., Hosur, R., & Berger, B. (2010). Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. *Nucleic acids research*, gkq481.

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics, 9*(6), 477-485.

Sleiman, P. M. A., Flory, J., Imielinski, M., Bradfield, J. P., Annaiah, K., Willis-Owen, S. A. G., and others. (2010). Variants of DENND1B associated with asthma in children. *New England Journal of Medicine, 362*(1), 36-44.

Stainton, J. J., Haley, C. S., Charlesworth, B., Kranis, A., Watson, K., & Wiener, P. (2015). Detecting signatures of selection in nine distinct lines of broiler chickens. *Animal Genetics, 46*(1), 37-49. doi: 10.1111/age.12252

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., and others. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102*(43), 15545-15550.

Sun, Y. V. (2012). Integration of biological networks and pathways with genetic association studies. *Human genetics, 131*(10), 1677-1686.

Sun, Y. V., Cai, Z., Desai, K., Lawrance, R., Leff, R., Jawaid, A., and others (2007, 2007). *Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests*.

Sun, Y. V., Shedden, K. A., Zhu, J., Choi, N.-H., & Kardia, S. L. R. (2009, 2009). *Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression*.

Svishcheva, G. R., Belonogova, N. M., & Axenovich, T. I. (2014). FFBSKAT: fast family-based sequence kernel association test. *PLOS ONE, 9*(6), e99407.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., and others. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research, 45*(D1), D362-D368.

Szymczak, S., Holzinger, E., Dasgupta, A., Malley, J. D., Molloy, A. M., Mills, J. L., and others (2016). r2VIM: A new variable selection method for random forests in genome-wide association studies. *BioData mining, 9*(1), 7.

Tabor, H. K., Risch, N. J., & Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics, 3*(5), 391-397.

Taşan, M., Musso, G., Hao, T., Vidal, M., MacRae, C. A., & Roth, F. P. (2015). Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature Methods, 12*(2), 154-159. doi: 10.1038/nmeth.3215

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74.

Thomas, W. (1989). *Multiple comparison procedures, by Yosef Hochberg and Ajit C. Tamhane: Wiley, New York, 1987, XXII+ 450 pp., price US $44.95, ISBN 0-471-82222-1*: Elsevier.

Tian, W., Zhang, L. V., Taşan, M., Gibbons, F. D., King, O. D., Park, J., and others (2008). Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. *Genome Biology, 9*(Suppl 1), S7. doi: 10.1186/gb-2008-9-s1-s7

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the royal statistical society. Series B (Methodological)*, 267-288.

Tishkoff, S. A., & Kidd, K. K. (2004). Implications of biogeography of human populations for'race'and medicine. *Nature genetics, 36*, S21-S27.

Torgerson, D., Elizabeth, J. A., Grace, Y. C., W James, G., Christopher, R. G., & others. (2011). Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nature genetics, 43*(9), 887-892.

Tu, Z., Wang, L., Arbeitman, M. N., Chen, T., & Sun, F. (2006). An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics, 22*(14), e489-e496. doi: 10.1093/bioinformatics/btl234

Ulitsky, I., & Shamir, R. (2009). Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics, 25*(9), 1158-1164.

Vandin, F., Clay, P., Upfal, E., & Raphael, B. J. (2012). Discovery of mutated subnetworks associated with clinical data in cancer *Biocomputing 2012* (pp. 55-66).

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics, 4*, 270.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., & Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics, 87*(3), 325-340.

Wan, Y. I., Shrine, N. R. G., Artigas, M. S., Wain, L. V., Blakey, J. D., Moffatt, M. F., and others. (2012). Genome-wide association study to identify genetic determinants of severe asthma. *Thorax*, thoraxjnl-2011.

Wang, K., Li, M., & Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics, 81*(6), 1278-1283.

Wang, L., Matsushita, T., Madireddy, L., Mousavi, P., & Baranzini, S. E. (2014). PINBPA: Cytoscape app for network analysis of GWAS data. *Bioinformatics, 31*(2), 262-264.

Wang, L., Mousavi, P., & Baranzini, S. E. (2015). iPINBPA: an integrative network-based functional module discovery tool for genome-wide association studies. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 255-266.

Wang, Q., Yu, H., Zhao, Z., & Jia, P. (2015). EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics*, btv150.

Wang, X., & Cairns, M. J. (2014). SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. *Bioinformatics*, btu090.

Wang, Y., & Xia, Y. (2008). Condition specific subnetwork identification using an optimization model. *Proc Optim Syst Biol, 9*, 333-340.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., and others. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research, 42*(D1), D1001-D1006.

Weng, L., Macciardi, F., Subramanian, A., Guffanti, G., Potkin, S. G., Yu, Z., & Xie, X. (2011). SNP-based pathway enrichment analysis for genome-wide association studies. *BMC bioinformatics, 12*(1), 99.

Wenzel, S. E. (2012). Asthma phenotypes: the evolution from clinical to molecular approaches. [10.1038/nm.2678]. *Nat Med, 18*(5), 716-725.

WHO media center. (2017). Asthma fact sheets, from http://www.who.int/mediacentre/factsheets/fs307/en/

Willemsen, G., Van Beijsterveldt, T. C. E. M., Van Baal, C. G. C. M., Postma, D., & Boomsma, D. I. (2008). Heritability of self-reported asthma and allergy: a study in adult Dutch twins, siblings and parents. *Twin Research and Human Genetics, 11*(02), 132-142.

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics, 26*(17), 2190-2191. doi: 10.1093/bioinformatics/btq340

Wilson, S. (1980). A note on the correct definition of additive deviation and dominance deviation. *Annals of human genetics, 44*(1), 113-115.

Witte, J. S. (2010). Genome-wide association studies and beyond. *Annual review of public health, 31*, 9-20.

Wu, Michael C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics, 89*(1), 82-93. doi: 10.1016/j.ajhg.2011.05.029

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics, 25*(6), 714-721.

Wu, X., Liu, Q., & Jiang, R. (2009). Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics, 25*(1), 98-104. doi: 10.1093/bioinformatics/btn593

Wu, Z., Zhao, X., & Chen, L. (2009). Identifying responsive functional modules from protein-protein interaction network. *Molecules and Cells, 27*(3), 271-277. doi: 10.1007/s10059-009-0035-x

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology, 16*(5), 284-287. doi: 10.1089/omi.2011.0118

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., & Weir, B. S. (2002). Truncated product method for combining P-values. *Genetic Epidemiology, 22*(2), 170-185. doi: 10.1002/gepi.0042

Zhang, H., Wang, M., & Chen, X. (2009). Willows: a memory efficient tree and forest construction package. *BMC bioinformatics, 10*(1), 130.

Zhang, K., Cui, S., Chang, S., Zhang, L., & Wang, J. (2010). i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic acids research, 38*(suppl 2), W90-W95.

Zou, D., Ma, L., Yu, J., & Zhang, Z. (2015). Biological databases for human research. *Genomics, Proteomics & Bioinformatics, 13*(1), 55-63.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.

Zou, L., Huang, Q., Li, A., & Wang, M. (2012). A genome-wide association study of Alzheimer's disease using random forests and enrichment analysis. *Science China Life Sciences, 55*(7), 618-625.