

BIOINFORMATICS APPROACHES TO ASSOCIATE SINGLE NUCLEOTIDE
POLYMORPHISMS WITH HUMAN DISEASES ACCORDING TO THEIR
PATHWAY RELATED CONTEXT

by
BURCU GÜNGÖR

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Sabanci University

June 2012

BIOINFORMATICS APPROACHES TO ASSOCIATE SINGLE NUCLEOTIDE
POLYMORPHISMS WITH HUMAN DISEASES ACCORDING TO THEIR
PATHWAY RELATED CONTEXT

APPROVED BY:

Assoc. Prof. O. Uğur Sezerman
(Thesis Supervisor)



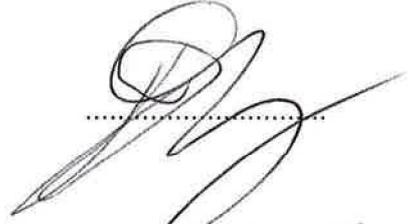
Prof. Uğur Özbek



Prof. Selim Çetiner



Assoc. Prof. Devrim Gözüağık



Assist. Prof. Murat Çokol



DATE OF APPROVAL:05.06.2012.....

© BURCU GÜNGÖR 2012

All rights reserved

ABSTRACT

BIOINFORMATICS APPROACHES TO ASSOCIATE SINGLE NUCLEOTIDE POLYMORPHISMS WITH HUMAN DISEASES ACCORDING TO THEIR PATHWAY RELATED CONTEXT

Burcu Güngör
Biological Sciences and Bioengineering
PhD Thesis, 2012

Prof. O. Ugur Sezerman (Thesis Supervisor)

Keywords: Genome Wide Association Study (GWAS), Single Nucleotide Polymorphism (SNP), human complex diseases, pathways, protein-protein interaction networks

Genome-wide association studies (GWASs) with millions of single nucleotide polymorphisms (SNPs) are popular strategies to reveal the genetic basis of human complex diseases. Despite many successes of GWASs, it is well recognized that new analytical approaches have to be integrated to achieve their full potential. In this thesis, starting with a list of SNPs, found to be associated with disease in GWAS, we have developed a novel methodology to devise functionally important pathways through the identification of SNP targeted genes within these pathways. Our methodology is based on functionalization of important SNPs to identify effected genes and disease related pathways. We have tested our methodology on rheumatoid arthritis, epilepsy, intracranial aneurysm and Behçet's disease datasets. With the whole-genome sequencing on the horizon, we show that the full potential of GWASs can be achieved by integrating prior knowledge from functional properties of a SNP and pathway-oriented analysis via protein-protein interaction networks.

ÖZET

TEK NÜKLEOTİD POLİMORFİZMLERİNİ YOLAKLAR ÜZERİNDEN İNSAN HASTALIKLARI İLE İLİŞKİLENDİRMEK İÇİN BİYOİNFORMATİK YÖNTEMLER

Burcu Güngör
Biyolojik Bilimler ve Biyomühendislik
Doktora Tezi, 2012

Prof. Dr. O. Uğur Sezerman (Tez Danışmanı)

Anahtar Kelimeler: tüm genom bağlantı analizi, tek nükleotid polimorfizmi, karmaşık insan hastalıkları, yolaklar, protein-protein etkileşim ağları

Milyonlarca tek nükleotid polimorfizmlerinin incelendiği tüm genom bağlantı analizleri (TGBA), insan karmaşık hastalıklarının genetik temellerini açığa çıkarmak için popüler stratejilerdir. TGBAların bilinen pek çok başarısına rağmen, onların tüm potansiyellerine ulaşabilmek için yeni analitik yöntemlerin entegre edilmesi gerektiği iyi bilinir. Bu tezde, TGBAda hastalıkla ilişkisi bulunmuş tekli nükleotid polimorfizm (TNP) listesi ile başlayıp, fonksiyonel olarak önemli yolak listesini, yolağın içindeki TNPlar tarafından hedeflenen genleri bularak ortaya çıkaran yeni bir yöntem geliştirdik. Metodumuz, etkinen genlerin ve hastalıkla ilgili yolakların bulunması için önemli TNPlerin fonksiyonel özelliklerinin incelenmesiyle başlar. Yöntemimizi romatizma, epilepsi, anevrizma ve Behçet hastalığı TGBA verilerinde test ettik. Ufukta tüm genom dizilemesi varken, TGBAnın tüm potansiyellerine, TNPlerin fonksiyonel özellikleri ve protein protein etkileşim ağları ile yolak bazlı analizlerden önsel bilgiler katarak erişilebileceğini gösterdik.

To my little son Selim, my husband Çađrı, and my dearest family

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Ugur Sezerman who attracted me to the field of bioinformatics in 1999, and supported me consistently since then. I am deeply indebted to him for all his advices, helps, and for being a role model during both my undergraduate and doctorate studies. Even during our undergraduate years, he showed us being an academician is not only rewarding, but also very enjoyable. I am thankful to him most importantly for encouraging me to go back for my PhD studies, even with having a one year old baby and a full time job. I am also grateful to my former advisor at Georgia Tech, Mark Borodovsky. Through his supervision, he has shown me how research should be done and how an academician should be. I have to thank Howard Jacob and Oya Aran for convincing me that I will be back to academia one day. Next, I need to express my gratitude to my examining committee members Ugur Özbek, Selim Çetiner, Murat Çokol and Devrim Gözüaçık. I am also very grateful to John Bowes and to Simon Potter, for their help with WTCCC data formats. I also would like to thank Scott Saccone, Phil Hyoun Lee, Claude Chelala, Gabriela Bindea for their helps with SPOT, F-SNP, SNPnexus, ClueGO tools; Albert-László Barabási and Michael Cusick for providing us PPI dataset; Christine Nardini, Sergio Baranzini for their valuable discussions. I have to thank Murat Gunel, Katsuhito Yasuno, Ituro Inoue for sharing their GWAS data on aneurysm; Boris Kirschek, Hirofumi Nakaoka for sharing their gene expression data on aneurysm; Dalia Kasperaviciute for sharing their epilepsy data; Akira Meguro, Ahmet Gul for sharing their Behçet's disease data with us. I am grateful to my friends, Tuba Ozbay, Müge Erdoğan-Birlik, Ece Egemen, Bahar Soğutmaz Özdemir, Hande Kaymakçalan-Çelebiler, Süreyya Özoğur-Akyüz. I also have to thank Taşkın Koçak for his support and tolerance. Finally, a special thank you goes to my family. They have always given me their unconditional love and supported me in my life and education. I'd like to give my heartfelt thanks for my husband Çağrı, especially since he didn't pour a glass of water to my laptop while writing up my PhD thesis, as I have done to him by mistake. He has been always there and helped me whenever I need. Another special thanks go to my mother Gülay, my father Ömer, and my sister Zeynep, for their endless love and support over the years. I am grateful to my mother in law Ayten, she supported me all those years. I want to give a very special thank to my dear son, Selim, for being very patient despite his young age. From now on, I promise to play with him whenever he wants!

TABLE OF CONTENTS

ABSTRACT.....	iv
ÖZET	v
DEDICATION	vi
ACKNOWLEDGEMENTS	vii
TABLE OF CONTENTS.....	viii
LIST OF TABLES.....	xii
LIST OF FIGURES	xiii
ABBREVIATIONS	xv
CHAPTER 1	1
1 INTRODUCTION	1
1.1 Motivation.....	1
1.2 Thesis statement and contributions	2
1.3 Organization of the thesis	5
CHAPTER 2	6
2 BACKGROUND INFORMATION ON BIOLOGICAL & COMPUTATIONAL ASPECTS	6
2.1 Mendelian Disorders.....	6
2.2 Human Complex Diseases	8
2.2.1 Rheumatoid Arthritis (RA)	9
2.2.2 Partial Epilepsy (PE).....	10
2.2.3 Intracranial Aneurysm (IA).....	11
2.2.4 Behçet’s Disease	12
2.3 Biological pathways.....	13
2.3.1 KEGG pathways.....	14
2.3.2 Pathway oriented high-throughput data analysis	14
2.4 Genome wide association studies (GWAS)	16

2.4.1 Overview of the GWAS.....	16
2.4.2 Pathway and network oriented GWAS data analysis.....	18
2.4.3 GWAS on different populations.....	21
CHAPTER 3	23
3 MATERIALS AND METHODS.....	23
3.1 Materials	23
3.1.1 Datasets	23
3.1.1.1 GWAS datasets	23
3.1.1.1.1 Rheumatoid arthritis dataset.....	23
3.1.1.1.2 Partial epilepsy dataset.....	24
3.1.1.1.3 Intracranial aneurysm European population dataset	24
3.1.1.1.4 Intracranial aneurysm Japanese population dataset.....	24
3.1.1.1.5 Behçet’s disease Turkish population dataset.....	25
3.1.1.1.6 Behçet’s disease Japanese population dataset.....	25
3.1.1.2 Protein-protein interaction network	25
3.1.1.3 IA gene expression dataset for Japanese population	25
3.1.2 Computational equipment setup.....	26
3.1.2.1 Java platform.....	26
3.1.2.2 Cytoscape	26
3.1.2.3 SNP functionalization tools.....	27
3.2 Methods	27
3.2.1 Design of Pathway and Network Oriented GWAS Analysis (PANOGA) Tool	27
3.2.1.1 PANOGA Overview	27
3.2.1.2 SNP functionalization	30
3.2.1.3 SNP-wise weighted p-value calculation.....	32
3.2.1.4 SNP to gene assignment.....	32
3.2.1.5 Gene-wise weighted p-value calculation.....	33

3.2.1.6 Active sub-network identification	33
3.2.1.6.1 Overlap threshold parameter	35
3.2.1.7 Functional enrichment, pathway identification	36
3.2.1.8 Integration of the functional enrichments of the generated subnetworks.....	37
3.2.2 Development of a protocol to identify SNP targeted pathways from GWAS.....	37
3.2.2.1 PANOGA input files' formats.....	38
3.2.2.1.1 GWAS dataset file format.....	38
3.2.2.1.2 Protein-protein interaction network file format.....	39
3.2.2.2 Procedure	40
3.2.2.2.1 Install PANOGA	40
3.2.2.2.2 Preprocess GWAS data	40
3.2.2.2.3 Assign SNPs to Genes.....	42
3.2.2.2.4 Install Cytoscape and its plugins	44
3.2.2.2.5 Obtain Functional Information of SNPs.....	45
3.2.2.2.6 Prepare the Gene Attributes data.....	46
3.2.2.2.7 Obtain network data	47
3.2.2.2.8 Load network data.....	47
3.2.2.2.9 Import gene attributes	48
3.2.2.2.10 Identify sub-networks.....	49
3.2.2.2.11 Parse jActiveModules output	49
3.2.2.2.12 Functional enrichment of subnetworks	50
3.2.2.2.13 Combine functional enrichment results.....	51
3.2.2.2.14 Visualize SNP targeted genes in a KEGG pathway map	52
CHAPTER 4	54
4 RESULTS	54
4.1 Anticipated results of PANOGA protocol	54
4.2 Results on rheumatoid arthritis dataset	60

4.2.1 Significant sub-networks for RA.....	61
4.2.2 Functionally important KEGG pathways for RA.....	64
4.2.3 Functionally grouped annotation network of RA.....	69
4.2.4 Comparison with known drug target genes for RA.....	72
4.2.5 Comparison with random networks	73
4.2.6 KEGG pathway map of JAK-STAT signaling, as related to RA.....	73
4.3 Results on partial epilepsy dataset	74
4.4 Results on intracranial aneurysm dataset.....	80
4.5 Results on Behçet’s disease dataset	85
CHAPTER 5	89
5 DISCUSSION	89
5.1 Discussion on rheumatoid arthritis dataset.....	90
5.2 Discussion on partial epilepsy dataset.....	91
5.3 Discussion on intracranial aneurysm dataset	95
5.4 Discussion on Behçet’s disease dataset.....	100
5.5 General Discussion	100
CHAPTER 6	102
6 CONCLUSION	102
REFERENCES.....	105

LIST OF TABLES

Table 2.1 Examples of Mendelian type human disorders, types of inheritance, responsible genes (Chial, 2008)	22
Table 2.2 Comparison of pathway based GWAS data analysis platforms (Yaspan and Veatch, 2011).....	35
Table 3.1 Description of data sources used in our functional score.....	46
Table 4.1 Pathway based representation of PANOGA results, focusing on SNP targeted genes.....	71
Table 4.2 Pathway based representation of PANOGA results, focusing on subnetwork genes.....	72
Table 4.3 Pathway based representation of PANOGA results, focusing on associated SNPs from GWAS and their associated genes (SNP targeted genes).....	73
Table 4.4 Gene list representation of PANOGA for the identified SNP targeted pathways.....	74
Table 4.5 Overrepresented KEGG Pathways found in the highest scoring sub-network for RA.....	81
Table 4.6 Comparison of found KEGG pathways with previous studies in terms of number of genes associated within each KEGG term for RA.....	83
Table 4.7 The top 30 over-represented KEGG pathways identified for PE dataset.	90
Table 4.8 Comparison of the top 30 SNP-targeted KEGG pathways with the pathways of the known genes as associated with PE.....	93
Table 4.9 The top 20 KEGG pathways identified for both populations in IA.....	96
Table 4.10 The top 20 over-represented KEGG pathways for IA, and the SNP targeted genes within these pathways	97
Table 4.11 The top 20 over-represented KEGG pathways identified for gene expression data of IA.....	100
Table 4.12 The top 10 KEGG pathways identified for both populations in Behçet’s disease.....	102
Table 4.13 The top 10 over-represented KEGG pathways for Behçet’s disease, and the SNP targeted genes within these pathways.....	103

LIST OF FIGURES

Figure 2.1 Pathway-level analysis of high-throughput datasets (Kelder, et al., 2010) (Bebek, et al., 2012).....	30
Figure 2.2 Genome-wide association studies (GWAS) (Manolio, 2010).....	33
Figure 3.1 Outline of PANOGA’s assessment process.....	44
Figure 3.2 Summary of PANOGA protocol.....	53
Figure 3.3 Sample gene attributes input file (sample_spot_fsnp_snpnexus.pvals), showing SPOT and F-SNP weighted p-values (Pw-values) for each SNP associated gene.....	61
Figure 4.1 Customized KEGG pathway map for JAK-STAT signaling pathway.....	75
Figure 4.2 (a) The highest scoring sub-network is composed of 275 nodes and 778 edges. Node size is shown as proportional to the degree of a node. (b) Zoomed in view of the highest scoring sub-network. 20 genes known in literature as associated with RA are shown in green. Blue denotes the genes in our highest scoring sub-network that cannot be associated with RA in literature.....	77
Figure 4.3 Highest scoring subnetwork, that is identified by jActiveModule using gene-wise weighted p-values, which combines GWAS p-values with the SNP’s functional score.....	78
Figure 4.4 (a) Node degree distribution of the highest scoring sub-network follows a power-law, showing that our network displays scale-free properties, as expected from a biological network. (b) Node degree distribution of a random network, obtained via randomization of our highest scoring sub-network using Erdos-Renyi algorithm.....	79
Figure 4.5 (a) Functionally grouped annotation network of our highest scoring sub-network. (b) Zoomed in view of the entire functional annotation network.....	85
Figure 4.6 (a) Comparison of KEGG pathway terms with literature verified RA genes/our gene set were shown in green/red, respectively. (b) Zoomed in view of the network. The color gradient showed the gene proportion of each set associated with the term.....	87
Figure 4.7 Functionally grouped annotation network of the identified pathways for epilepsy dataset. The pathways are grouped based on the similarity of their SNP targeted genes.....	94
Figure 5.1 The complement and coagulation cascade (a) Up and down-regulated genes are shown in red and in blue, respectively, as a result of microarray analysis	107

for epilepsy-associated gangliogliomas (Aronica, et al., 2008). (b) The shade of red color in genes indicates the number of GWAS targeted SNPs per base pair of the gene. Red refers to the highest targeted gene, whereas white refers to a gene product, not targeted by the SNPs.....

Figure 5.2 KEGG pathway map for MAPK signaling. The set of genes shown in blue includes genes that are found for EU dataset; yellow includes genes that are found for JP dataset; red includes genes that are found both by EU and JP GWAS of IA..... 111

Figure 5.3 KEGG pathway map for TGF-beta signaling pathway. The shade of red color in genes indicates the number of targeted SNPs in JP population per base pair of the gene. Red refers to the highest targeted gene, whereas white refers to a gene product, not targeted by the SNPs. Blue border indicates that the gene is found to be differentially expressed..... 113

Figure 5.4 KEGG pathway map for calcium signaling pathway. The set of genes shown in blue includes genes that are found for EU dataset; yellow includes genes that are found for JP dataset; red includes genes that are found both by EU and JP GWAS of IA..... 114

ABBREVIATIONS

GWAS	Genome wide association study
GSEA	Gene-set enrichment analysis
IA	Intracranial aneurysm
KEGG	Kyoto Encyclopedia of Genes and Genomes
KEGGDPD	KEGG Disease Pathways Database
LD	Linkage disequilibrium
miRNA	microRNA
NHGRI	National Human Genome Research Institute
OMIM	Online Mendelian Inheritance in Man
PANOGA	Pathway and network oriented GWAS analysis
PE	Partial epilepsy
PPI	Protein-protein interaction
RA	Rheumatoid arthritis
SNP	Single nucleotide polymorphism
TF	Transcription factor
TFBS	Transcription factor binding site

CHAPTER 1

1 INTRODUCTION

1.1 Motivation

Human complex diseases are at the interplay of multiple genetic, life style and environmental factors. As the incidence of human complex diseases increase, researchers attempt to exploit many different experimental techniques to be able to comprehend the complex nature of these diseases. The advances in high-throughput laboratory methods now allow researchers to investigate larger questions in larger populations and to cover the genome in more detail. Thus, the discoveries in the genetics of complex diseases get accelerated. As it becomes easier and cheaper to find out the genotypes of many individuals, now the genetic studies cover a richer set of mutations within individual genes rather than focusing on one or a few coding variants. In parallel, the underlying patterns of coinheritance of markers (linkage disequilibrium, LD) are discovered through the HapMap Project (<http://www.hapmap.org>). Once this information is combined with the chip-based genotyping assays, genome-wide association studies (GWASs) of complex diseases became quite popular.

GWASs aim to identify single-nucleotide polymorphisms (SNPs) that may be associated with a disease under study, via comparing the differences in the frequencies of the SNPs between the cases and the controls. GWASs have been advocated as the most powerful approach to explore polygenic traits for many diseases. Although GWASs are rapidly increasing in number, numerous challenges persist in identifying and explaining the associations between loci and quantitative phenotypes. As observed in many examples of GWASs, few of the many possible variants can contribute to the

explanation of a small percentage of the estimated heritability for complex diseases, and thus it is a major challenge to identify marker SNPs specific to a complex disease or to develop genetic risk prediction tests (Couzin and Kaiser, 2007; Couzin and Kaiser, 2007; Dermitzakis and Clark, 2009; Gibson, 2010; Shriner, et al., 2007; Williams, et al., 2007). Although, there are many success stories that uncover the genetic epidemiology of complex diseases using GWASs, still many of the fundamental questions relating to the mechanisms of complex human disease remain unanswered.

A biological pathway is a sequence of activities between molecules in a cell, which ends up to a particular product or a change in a cell. Most of the times, in complex diseases, several genes and thus several pathways have to be affected for disease development. Multiple factors (e.g. SNPs, miRNAs, metabolic factors) may target different set of genes in the same pathway crippling its function and thus causing the disease development. Therefore, each gene makes a mild contribution to disease risk, which is difficult to detect using existing methodologies. In addition to the significance of the pathways for complex diseases in worldwide, the pathway knowledge can be further exploited to enlighten the underlying disease etiology in different populations. Finally, the knowledge of the genetic determinants of a disease (in the form of variants, genes or pathways) may provide diagnostic tools for identifying individuals at increased risk for that specific disease (McCarthy, et al., 2008).

1.2 Thesis statement and contributions

In this thesis, we hypothesize that *the pathways are more important than individual genes, SNPs and other individual factors to elucidate disease mechanisms*. Hence, to understand the underlying mechanism of complex diseases, rather than focusing on SNP/gene markers, we hypothesize that one should find out affected pathways targeted by different factors. Throughout this thesis, we developed a novel pathway and network oriented GWAS analysis method, PANOGA, that challenges to identify pathway markers by combining nominally significant evidence of genetic association with protein-protein interaction networks, functional information of selected SNPs, and current knowledge of biochemical pathways (Bakir-Gungor and Sezerman,

2011). Our methodology devises functionally important pathways through the identification of SNP targeted genes within these pathways. We have tested our methodology on rheumatoid arthritis (RA), partial epilepsy (PE), intracranial aneurysm (IA) and Behçet's disease datasets and shown that pathway and network oriented analysis of GWASs reveals the underlying mechanisms of complex diseases in more detail, compared to the traditional analyses of GWASs. The main contributions of this thesis can be summarized as following:

- 1) We present PANOGA, pathway and network oriented GWAS analysis, that challenges to identify disease associated Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways by combining nominally significant evidence of genetic association with current knowledge of biochemical pathways, protein-protein interaction networks, and functional information of selected SNPs (Bakir-Gungor and Sezerman, 2011).
- 2) In the rheumatoid arthritis GWAS dataset, we identified both previously known (e.g. Jak-STAT signaling, T cell receptor signaling, leukocyte transendothelial migration, cytokine-cytokine receptor interaction, antigen processing and presentation) and additional KEGG pathways (e.g. pathways in cancer, neurotrophin signaling, chemokine signaling pathways) as associated with RA. The KEGG functional enrichment of the RA specific drug target genes included these additionally found pathway terms. Among the previously known pathways, we identified additional genes as associated with RA (e.g. antigen processing and presentation, tight junction). Importantly, within these pathways, the associations between some of these additionally found genes, such as HLA-C, HLA-G, PRKCQ, PRKCZ, TAP1, TAP2 and RA were verified by either OMIM database or by literature retrieved from the NCBI PubMed module (Bakir-Gungor and Sezerman, 2011). Similarly, we applied our methodology on epilepsy dataset, and showed that PANOGA was able to identify significant pathways, explaining the pathogenesis of the disease. The relation between these pathways and the disease was supported by other studies in literature. 20 out of the top 30 affected pathways were found to be common with at least three different studies, among the seven studies compared (Bakir-Gungor and Sezerman, 2012, submitted).

3) Via applying PANOGA on two aneurysm GWASs, conducted on European and Japanese populations, we have shown that 7 of the top 10 affected pathways are common between these two populations (where, the probability of getting 7 common pathways out of randomly selected 10 pathways from existing 246 human KEGG pathways is $2.44E^{-36}$). These pathways are MAPK signaling, Cell cycle, TGF-beta signaling, Focal adhesion, Adherens junction, Regulation of actin cytoskeleton, and Neurotrophin signaling pathways. The relation between these pathways and the disease is supported by other studies in literature. We have also applied PANOGA on two Behçet's disease GWASs, conducted on Turkish and Japanese populations. Even though there were very few common SNPs and commonly targeted genes, we have shown that 5 of the top 10 pathways are common between these two populations. Hence, we emphasize the importance of pathway-oriented analysis to enlighten disease mechanisms. Although different SNP targeted genes are affected on each population, these genes map to the same pathways among different populations (Bakir-Gungor and Sezerman, 2012, submitted). Accordingly, we introduce pathway marker concept to the literature, which explains universal disease development mechanism. As a potential application, each population may search for disease causing factors targeting the genes within these marker pathways. Rather than the population, the same method can be extended to individuals to identify modifications occurring on the genes within these pathways and thus determine individual reasons for disease development, which can be exploited for drug development and personalized therapeutical applications.

4) Since our method can be easily applied to GWAS datasets of other diseases, it will facilitate the identification of disease specific pathway combinations. In this regard, PANOGA protocol represents a feasible solution for the identification of pathway markers to bridge the gap between GWAS and biological mechanisms of complex diseases (Bakir-Gungor and Sezerman, 2012). PANOGA protocol is designed as a dynamic and modular platform, which can be easily updated with new methodologies and datasets. On the other hand, to present the user a fully automated option, we implemented PANOGA protocol as a web-server, which is almost ready to be published (in preparation).

5) Finally, these research efforts correspond to four journal papers published or submitted, during the course of this thesis (Bakir-Gungor and Sezerman, 2011); (Bakir-Gungor and Sezerman, 2012); (Bakir-Gungor and Sezerman, 2012, submitted); (Bakir-Gungor, et al., 2012, submitted). Two additional manuscripts, describing our results on Behçet's disease and webserver implementation are in preparation.

1.3 Organization of the thesis

We present a brief introduction to the human complex diseases, GWASs, problems in GWAS data analysis, thesis statement and contributions in this chapter. Chapter 2 gives basic background on the biological and computational aspects, and summarizes related literature. Information about biological pathways, Mendelian vs. complex diseases, network and pathway based approaches to GWASs, the significance of conducting GWASs on different populations are also discussed in Chapter 2. Chapter 3 presents the details of the proposed pathway and network oriented GWAS analysis protocol, and the datasets used. The details of the design and the implementation of the PANOGA protocol are also explained in Chapter 3. Chapter 4 provides results of the proposed system on several data sets, i.e. rheumatoid arthritis, partial epilepsies, intracranial aneurysm, Behçet's disease. The results are discussed from both biological and computational perspectives in Chapter 5 for each dataset. In this chapter, the advantages of network, pathway and population based GWAS analysis, over traditional GWASs are discussed in detail. Chapter 6 concludes the thesis and gives some future directions for pathway oriented and integrative GWAS data analysis procedures.

CHAPTER 2

2 BACKGROUND INFORMATION ON BIOLOGICAL & COMPUTATIONAL ASPECTS

2.1 Mendelian Disorders

Mendelian disorders are a type of human diseases that obey Mendelian pattern of inheritance and are caused by the variances in a single gene. Hence, they are also called single-gene or mono-genic diseases. They are relatively uncommon. According to their modes of inheritance, single-gene diseases can fall into one of the following five categories:

1. Autosomal recessive inheritance,
2. Autosomal dominant inheritance,
3. X-linked recessive inheritance,
4. X-linked dominant inheritance,
5. Mitochondrial inheritance.

Depending on where the gene for the trait is located, a single-gene disorder is categorized as autosomal vs. X-linked, or may be mitochondrial. Depending on how many copies of the mutant allele are required to express the phenotype, a single-gene disorder is categorized as recessive vs. dominant. Examples of Mendelian type human disorders from these categories and known associated genes are shown in Table 2.1.

Table 2.1 Examples of Mendelian type human disorders, types of inheritance, responsible genes (Chial, 2008).

Disease	Type of Inheritance	Gene Responsible
Phenylketonuria (PKU)	Autosomal recessive	Phenylalanine hydroxylase (<i>PAH</i>)
Cystic fibrosis	Autosomal recessive	Cystic fibrosis conductance transmembrane regulator (<i>CFTR</i>)
Sickle-cell anemia	Autosomal recessive	Beta hemoglobin (<i>HBB</i>)
Albinism, oculocutaneous, type II	Autosomal recessive	Oculocutaneous albinism II (<i>OCA2</i>)
Huntington's disease	Autosomal dominant	Huntingtin (<i>HTT</i>)
Myotonic dystrophy type 1	Autosomal dominant	Dystrophia myotonica-protein kinase (<i>DMPK</i>)
Hypercholesterolemia, autosomal dominant, type B	Autosomal dominant	Low-density lipoprotein receptor (<i>LDLR</i>); apolipoprotein B (<i>APOB</i>)
Neurofibromatosis, type 1	Autosomal dominant	Neurofibromin 1 (<i>NF1</i>)
Polycystic kidney disease 1 and 2	Autosomal dominant	Polycystic kidney disease 1 (<i>PKD1</i>) and polycystic kidney disease 2 (<i>PKD2</i>), respectively
Hemophilia A	X-linked recessive	Coagulation factor VIII (<i>F8</i>)
Muscular dystrophy, Duchenne type	X-linked recessive	Dystrophin (<i>DMD</i>)
Hypophosphatemic rickets, X-linked dominant	X-linked dominant	Phosphate-regulating endopeptidase homologue, X-linked (<i>PHEX</i>)
Rett's syndrome	X-linked dominant	Methyl-CpG-binding protein 2 (<i>MECP2</i>)
Spermatogenic failure, nonobstructive, Y-linked	Y-linked	Ubiquitin-specific peptidase 9Y, Y-linked (<i>USP9Y</i>)

Online Mendelian Inheritance in Man, OMIM is a comprehensive database that contains information on all known Mendelian disorders, including 5,264 phenotypes and 13,916 genes, as of May 22nd 2012 (Amberger, et al., 2011). To understand the genetic causes of Mendelian diseases, several attempts have been made, and these efforts resulted in major discoveries of gene variations that predispose to such diseases. This happens due to the simplicity of their inheritance patterns, compared to the human complex diseases.

2.2 Human Complex Diseases

In contrast to the Mendelian diseases, in which a single gene defines susceptibility to a disease, human complex diseases arise from the joint effects of multiple genetic, environmental factors and life style (Kiberstis and Roberts, 2002; Lander and Schork, 1994; Weeks and Lathrop, 1995). Hence they are also referred as multifactorial or polygenic diseases. Complex diseases appear commonly in the population and are of major clinical and economic significance. Many human diseases fall into this category, including cardiovascular diseases, cancer, Alzheimer's disease, diabetes mellitus, scleroderma, nicotine and alcohol dependence, asthma, rheumatoid arthritis, Parkinson's disease, epilepsies, multiple sclerosis, aneurysm, osteoporosis, connective tissue diseases, kidney diseases, autoimmune diseases, and many more (Hunter, 2005; Merikangas and Risch, 2003). These diseases are accepted as the major source of disability and death worldwide.

The genes related to complex disease phenotypes are inherited, but these genetic factors only illuminate one side of the coin. Environmental factors, including life style choices, act on the other side of the coin, differently from Mendelian diseases. In this regard, genetic predisposition indicates that a person has a genetic susceptibility to develop a certain disease. But, this does not guarantee that an individual with such a genetic tendency will develop the disease phenotype. At this point, the combined effect of environmental factors makes the final decision on the development of the disease phenotype. For example, researchers show that some type of the skin cancer is associated with mutations in the melanocortin 1 receptor gene (MC1R) in people with fair skin color (Box, et al., 2001). When these individuals are exposed to sunlight, then

the combined action of ultraviolet light B and the variants on the MC1R increases the risk of developing a skin cancer (Hunter, 2005).

Although complex diseases appear more frequently than the Mendelian diseases, little progress has been made in the identification of the genetic causes of these diseases. Even if some individual gene variants have been associated with multifactorial diseases, they typically have small effect sizes or account for only a few percent of disease risk. That said, the combined effects of gene variants within pathways might better explain complex disease development mechanisms (the paradigm of complex genetics).

2.2.1 Rheumatoid Arthritis (RA)

Rheumatoid Arthritis (RA, OMIM 180300) is a systemic inflammatory disease, primarily affecting synovial joints. As reported at the 2008 American College of Rheumatology meeting, about 1% of the world's population is afflicted by RA and women affected three times more often than men. Disease onset is most frequent between the ages of 40 and 50, but people of any age can be affected. While the earlier stages of the disease appear a disabling and painful condition, in the later stages it can lead to substantial loss of functioning and mobility.

Being a complex disease, the etiology of RA depends on a combination of multiple genetic and environmental conditions, involving a yet unknown number of genes. The heritability of this disease is estimated as ~50% based on family studies, including twin studies (Bali, et al., 1999; MacGregor, et al., 2000). In GWASs among RA patients of European ancestry, multiple risk alleles have been identified in the major histocompatibility complex (MHC) region, and 25 RA risk alleles have been confirmed in 23 non-MHC loci (Barton, et al., 2009; Begovich, et al., 2004; Gregersen, et al., 2009; Kurreeman, et al., 2007; Plenge, et al., 2007; Raychaudhuri, et al., 2008; Raychaudhuri, et al., 2009; Remmers, et al., 2007; Suzuki, et al., 2000; Thomson, et al., 2007; Zhernakova, et al., 2007). These variants explain about 23% of the genetic burden of RA (Raychaudhuri, et al., 2008), indicating that additional variations remain to be discovered to explain the polygenic etiology of RA.

2.2.2 Partial Epilepsy (PE)

Epilepsy is a common neurological disorder that affects around 1% of the world's population, including one in 200 children (Cowan, 2002; Pitkanen and Sutula, 2002; Sander, 2003). Even though it has myriad etiologies, it is characterized by recurrent and spontaneous seizures. In roughly 30% of epilepsy cases, it is a result of an insult to the brain, such as trauma, stroke, hypoxia, brain infection, tumour, postnatal insults, and status epilepticus (Hauser, 1994). Despite the heterogeneity in the causes of epilepsies, it is accepted as a highly genetic and heritable disorder in many cases (Gourfinkel-An, et al., 2001; Prasad, et al., 1999; Reid, et al., 2009; Walsh and McCandless, 2001). While the risk of having epilepsy in general population is 0.5 percent, the same risk among first-degree relatives of individuals with idiopathic generalized epilepsy reaches to 8-12 percent (Steinlein, 2004). This statistic also indicates a strong genetic component underlying epilepsy, but which is considered as a complex one in ~99% of the cases, rather than displaying the characteristics of Mendelian inheritance (Kasperaviciute, et al., 2010).

Partial epilepsy (PE) is a subcategory of epilepsy, which is characterized by localized origin of seizures. In other words, seizure affects only one part of the brain in PE. Although cortical dysplasias and low-grade neoplasms are the most frequently detected reasons in children, no identifiable etiology exist in adults (ie, neuroimaging studies are most often normal). Still, epilepsy patients share some biological features including EEG abnormalities, secondary generalization of partial seizures, and the elemental biophysical and neurochemical cellular components of seizures, e.g. action potentials and synaptic transmission processes. These observations indicate that there are some shared mechanisms in individual's predisposition to PE. Different studies report different estimates for PE heritability, even reaching up to 70% (Kjeldsen, et al., 2001). Reviews by Poduri *et al* (Poduri and Lowenstein, 2011) and Pandolfo *et al* (Pandolfo, 2011) summarize the current status in epilepsy genetics. Although the significance of genetic factors is well known for PE, the factors themselves are still ambiguous. Advancing genetic technologies such as genome wide association studies, whole-genome oligonucleotide arrays, whole exome, whole genome sequencing now allow researchers to discover epilepsy genetics from many different perspectives, which is not thought to be possible using traditional methodologies. For example, the identified copy number

variations as associated with idiopathic epilepsy explain higher percent of epilepsies than any single gene discovered so far (de Kovel, et al., 2010; Mefford, et al., 2010; Poduri and Lowenstein, 2011). Although the traditional pedigree studies of epilepsy genetics focus on ion channels and neurotransmitters, newly discovered genes, as identified with the help of advancing technologies reveal the significance of novel pathways involved in epileptogenesis (Kasperaviciute, et al., 2010; Poduri and Lowenstein, 2011). Even if the first GWAS of epilepsy on European population reported that no genome-wide significant association is found, it highlighted two candidate genes (ADCY9 and PRKCB) related to the chemokine signaling pathway, which is also identified through genome level expression analysis in epileptogenesis (Kasperaviciute, et al., 2010; Sharma, 2012). Second GWAS of epilepsy on Chinese population detected two highly correlated SNPs, rs2292096 ($P=1.0 \times 10^{-8}$, OR=0.63) and rs6660197 ($P=9.9 \times 10^{-7}$, OR=0.69). One of these SNPs is located on 1q32.1, in the CAMSAP1L1 gene, which encodes a cytoskeletal protein (Guo, et al., 2012). They showed once again the association of rs9390754 ($P = 1.7 \times 10^{-5}$) with epilepsy, which is found on 6q21 in the GRIK2 gene, that encodes a glutamate receptor. Additionally, they reported several other loci in genes involved in neurotransmission or neuronal networking, which requires further analysis (Guo, et al., 2012). Unfortunately, the GWAS dataset of this study is not publicly available.

2.2.3 Intracranial Aneurysm (IA)

Intracranial aneurysm (IA, OMIM 105800) is a cerebrovascular disease that affects around 1 per 50 people (Rinkel, et al., 1998). IA is thought to be a major public health concern since the rupture of an IA leads to subarachnoid hemorrhage (SAH), which is a destructive subset of stroke. One third of the patients with SAH die within the initial weeks after the bleed and the rest end up with severe physical disabilities (Ruigrok and Rinkel, 2010). Both environmental risk factors such as smoking, hypertension, excessive alcohol intake; and non-modifiable risk factors such as family history of IA, female gender and systemic diseases (e.g. polycystic kidney disease and vascul type of Ehlers Danlos disease) are accepted to have a role in the development of IA and SAH (Feigin, et al., 2005; Gieteling and Rinkel, 2003; Juvela, 2000; Juvela, et al., 2001; Pepin, et al., 2000; Taylor, et al., 1995). Since the subjects with familial preponderance

of IA have a higher risk of being affected by IA, the genetic components are thought to be related with the tendency of developing an IA. To identify these IA related genetic factors, several approaches including DNA linkage, candidate gene studies and genetic association studies have been used (Krischek and Noue, 2006; Nahed, et al., 2007; Ruigrok and Rinkel, 2008). Since these studies included relatively small numbers of patients and controls, results have been conflicting and have not been replicated (Krischek and Inoue, 2006; Nahed, et al., 2007; Ruigrok and Rinkel, 2008). Compared with the candidate gene studies, the hypothesis-free approach of GWAS allows testing for the association of all common variations in the entire genome with disease. Four recent GWAS identified some variants associated with IA (Akiyama, et al., 2010; Bilguvar, et al., 2008; Low, et al., 2012; Yasuno, et al., 2010). In JP population, five SNPs (rs1930095 ($P=1.31 \times 10^{-5}$), rs4628172 ($P=1.32 \times 10^{-5}$), rs7781293 ($P=2.78 \times 10^{-5}$), rs7550260 ($P=4.93 \times 10^{-5}$), rs9864101 ($P=3.63 \times 10^{-5}$)) were associated with IA (Akiyama, et al., 2010; Low, et al., 2012). In EU population, five loci were found to be strongly related with IA on chromosomes 18q11.2 (rs11661542, OR=1.22, $P=1.1 \times 10^{-12}$), 10q24.32 (rs12413409, OR=1.29, $P=1.2 \times 10^{-9}$), 13q13.1 (rs9315204, OR=1.20, $P=2.5 \times 10^{-9}$), 8q11.23-q12.1 (rs10958409, rs9298506, OR=1.28, $P=1.3 \times 10^{-12}$), 9p21.3 (rs1333040, OR=1.31, $P=1.5 \times 10^{-22}$) (25) and a further 14 loci displayed suggestive association (Gaal, et al., 2012). However, these variants explain only a small percentage of the familial risk of IA, which makes genetic risk prediction tests currently unfeasible for IA (Ruigrok and Rinkel, 2010).

2.2.4 Behçet's Disease

Behçet's disease is a chronic systemic disease, characterized by recurrent inflammatory attacks affecting several organs such as orogenital mucosa, eyes and skin. It is firstly described by the Turkish clinician Hulusi Behçet in 1937 as a complex disorder (Behçet, 1937), and its etiology remains poorly characterized. Although Behçet's disease exists worldwide, it is more widespread in countries along the ancient silk route spanning from Japan to the Middle East and the Mediterranean basin. With a prevalence of 4 cases per 1,000 individuals, Behçet's disease is most frequently observed in Turkey among the Middle Eastern countries (Remmers, et al., 2010), (Hatemi and Yazici, 2011). In the Turkish population, the sibling recurrence risk ratio of Behçet's disease is

estimated to be between 11.4 and 52.5, which supports the genetic contributions to the disease (Remmers, et al., 2010). Candidate gene studies and two small GWASs (Fei, et al., 2009; Meguro, et al., 2010) have investigated the genetics of Behçet's disease, but the results have generally been underpowered, making interpretation and replication of the outputs problematic. Recently, two GWASs of Behçet's disease are conducted on Turkish (Remmers, et al., 2010) and Japanese (Mizuki, et al., 2010) populations. In these studies, a variant on HLA-B gene is found as the most strongly associated genetic factor to Behçet's disease, but it accounts for less than 20% of the genetic risk. This result indicates that other genetic factors are waiting to be discovered.

2.3 Biological pathways

One important goal of biology is to comprehend life at the molecular level, more specifically at the DNA, RNA, gene, or protein levels. This knowledge is central to perceive how cells act in concert in an organism and also how they dysfunction to cause a disease. In this regard, biological pathways organize our knowledge with respect to a functional mechanism and describe an order of events at the molecular level that realize this specific mechanism. For instance, the steps followed within the cell to replicate DNA, to control the cell division, or to degrade glucose in order to produce energy may each be represented as a biological pathway (Lamond, 2002). Typically, a pathway defines a group of molecular entities, their cellular locations and their relations, e.g. activates, degrades, inactivates, inhibits, phosphorylates. Most importantly, each such set of molecules are specialized to perform a specific biological function. Over the years, several canonical pathways, which cover many generic biological processes in the cell, have been proposed. One significant advantage of pathway representations is that they aid the comprehension of complex molecular relationships with their carefully designed maps. Pathway maps present an overview of the cascade of events, participating molecules and relations among them in a single diagram, which is easy to perceive. Since these diagrams capture the overall structure of a biological mechanism, they help to analyze potential consequences of perturbations (e.g. when one of the genes is mutated in a disease or when one of the proteins is targeted by a drug). In summary, biological pathways are fundamental to enlighten the functions of individual genes and

proteins in terms of systems and processes that contribute to normal physiology and to disease. Hence, the pathway-level analysis is a powerful approach to understand complex biological systems at multiple levels of biological organization; to create a full picture of a system's behaviour; and to interpret experimental data at a higher level than that of individual biomolecules.

2.3.1 KEGG pathways

Kyoto Encyclopedia of Genes and Genomes (KEGG, <http://www.genome.jp/kegg>) present experimental knowledge on biological systems, systemic functions of the cell in terms of molecular pathway maps. KEGG database is frequently curated by Kaneisha Labs, from published literature. As of May 2012, it holds 249 human pathways.

Soh et al conducted a comparative analysis between three widely used pathway databases (KEGG, Ingenuity and Wikipathways) (Soh, et al., 2010). They defined "Pathway Comprehensive Score" metric as the number of pathways a database hosts, divided by the total number of unique pathways present within that pathway database. According to this metric, KEGG achieves the highest score of 0.59, indicating that KEGG Pathways are the most comprehensive of all databases. Their second metric, "Gene Pair Coverage Score" is computed via dividing the number of gene pairs a database hosts by the total number of unique gene pairs. In terms of Gene Pair Coverage Score, KEGG achieves the highest score of 0.65. KEGG pathways are also widely used for high throughput data analysis. Hence, we will focus our pathway analysis on KEGG pathways.

2.3.2 Pathway oriented high-throughput data analysis

The tremendous boost in the "omics" technologies such as transcriptomics, proteomics and metabolomics makes it possible to generate a global picture of system characteristics, and to look for the interactions and coordinated behavior among different levels of biochemical activity. These experiments measure tens of thousands of entities in parallel, e.g., gene expression (Tarca, et al., 2009), protein abundance (Patterson and Aebersold, 2003) or metabolite concentrations (Ouattara, et al., 2012) in

various biological samples. Additionally, functional data, e.g., PPI (Bonetta, 2010), protein-DNA interactions (Luo, et al., 2009); or miRNA expressions (Duan, et al., 2011), or genetic variations (Knight, 2010; Lam, et al., 2012) can also be measured using high-throughput techniques. Due to the enormous size of these datasets, in practice, it gets impossible to manually curate them and to deduce the underlying mechanisms. At this point, to assist the human mind, bioinformatics approaches are crucial to integrate, summarize and present the high-throughput data in the context of biological knowledge (Gehlenborg, et al., 2010). In this regard, biological pathways rise as an effective strategy. They provide an abstraction of existing knowledge, which is more amenable to computing, rather than purely textual information. Moreover, as mentioned before pathway maps present an approach to integrate biological knowledge with data visualization to facilitate human interpretation of the results. Hence, as shown in Figure 2.1, performing pathway-level analysis for high-throughput datasets helps to identify relevant biological mechanisms and generate hypotheses, which can be further tested with smaller scale, but more sensitive experiments.

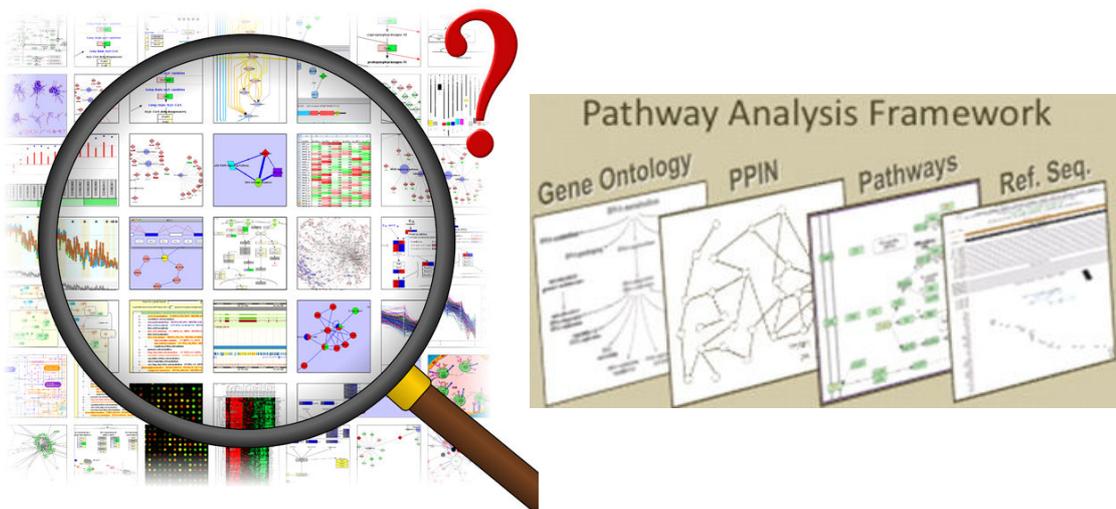


Figure 2.1 Pathway-level analysis of high-throughput datasets (Kelder, et al., 2010) (Bebek, et al., 2012).

There are several studies in the literature trying to analyze high-throughput data in a pathway related context, as reviewed in (Khatri, et al., 2012). A widely used method for conducting pathway-level analysis on single omic data is functional enrichment, which is also referred as over-representation analysis or the first generation approach in pathway analysis (Khatri and Draghici, 2005). In this method, firstly, a set of genes that are observed to be correlated with the phenotype under study, or a set of genes that are

differentially expressed is selected. Secondly, this gene set is compared with a priori defined molecular sets (e.g. genes in established pathways, gene ontologies (GO)). At the end of this comparison, the goal is to identify the established pathways or GO terms that result in higher levels of overlap with the phenotype-associated genes than expected by chance. Finally, the list of significantly overrepresented or ‘enriched’ sets/pathways is used to comment on the biological relevance of the data. Since the development of the original tools (e.g. DAVID (Dennis, et al., 2003), GoMiner (Zeeberg, et al., 2003)), around a hundred of modified implementations of these functional enrichment analysis have been published and most are reviewed in (Huang, et al., 2009). While most of these tools perform functional enrichment in terms of gene ontologies (e.g. Go-Mapper (Smid and Dorssers, 2004), ADGO (Nam, et al., 2006), Ontologizer (Bauer, et al., 2008), topGO (Alexa, et al., 2006)); some other tools conduct pathway based functional enrichment (e.g. Webgestalt (Zhang, et al., 2005), PANTHER (Mi, et al., 2010), KOBAS (Wu, et al., 2006)). There is also a third type of enrichment tool that checks for over-representation of genes both in gene ontologies and established pathways (e.g. ClueGO (Bindea, et al., 2009), DAVID (Huang, et al., 2007)). Following over-representation analysis, functional class scoring approaches are developed as a second generation approach in pathway analysis. While detecting affected pathways, these approaches make use of molecular measurements (e.g., gene expression levels) and take into account the dependence between genes in a pathway in (Khatri, et al., 2012). The third generation approaches, namely pathway topology based approaches incorporate topological features of pathways, instead of treating the pathways as simple lists of genes (Khatri, et al., 2012). Although most of these pathway analysis tools are initially developed to gain insight into the underlying biology of differentially expressed genes; in the meantime they get adapted to the analysis of other types of high-throughput datasets, which is still a very hot research field.

2.4 Genome wide association studies (GWAS)

2.4.1 Overview of the GWAS

Within the human genome, there are millions of sequence variations that vary in their frequencies and in the range of their effects on a particular disease. Single nucleotide

polymorphisms (SNPs) are the most common type among all other variants, which arise due to a single base substitution at a given genetic locus. Differently from point mutations, polymorphism terminology is restricted to the genetic variations with a population frequency of at least 1% (Ku, et al., 2010). During and after the completion of the Human Genome Project, millions of SNPs were detected. In parallel, International HapMap Project have been crucial to validate these SNPs and characterize their correlation or linkage disequilibrium (LD) patterns in populations of European, Asian and African ancestry. This knowledge had a central role in making the study of the genetics of common disease a reality and has been integral to the development of genome-wide association studies.

Genome-Wide Association Studies (GWAS) – in which hundreds of thousands of single nucleotide polymorphisms (SNPs) are tested simultaneously in thousands of cases and controls for association with a human complex disease, as shown in Figure 2.2,- have revolutionized the search for genetic basis of these diseases (Hardy and Singleton, 2009). The success of GWAS can be summarized with the published 600 genomewide association studies covering 150 distinct diseases and traits, explaining 800 SNP-trait associations. These studies not only identified novel common genetic risk factors, but also confirmed the importance of previously identified genetic variants. However, GWASs suffer from multiple-testing problem. To define the true DNA variant, that is associated with disease, a stringent statistical threshold is used (genotypic P value threshold of less than 5×10^{-8} for a SNP). Hence, in a typical GWAS, only a minority of DNA sequence variations that modulate disease susceptibility and their neighboring genes with the strongest evidence of association is explained. Whereas, in this “most-significant SNPs/genes” approach, genetic variants that confer a small disease risk but are of potential biological importance are likely to be missed. Hence, it is recognized that GWAS data is undermined in most cases and concentrating on a few SNPs and/or genes with the strongest evidence of disease association is not enough to exploit underlying physiological processes and disease mechanisms (Elbers, et al., 2009). For instance, PPARG variants are known to be associated with type 2 diabetes (T2D) (Altshuler, et al., 2000). Whereas, this true association is missed by the four out of five GWAS designed to replicate the initial finding, due to its modest effect on disease susceptibility (odds ratio 1.2) (Baranzini, et al., 2009; Frayling, 2007). A similar situation was recently observed regarding the association of IL7R variants with multiple

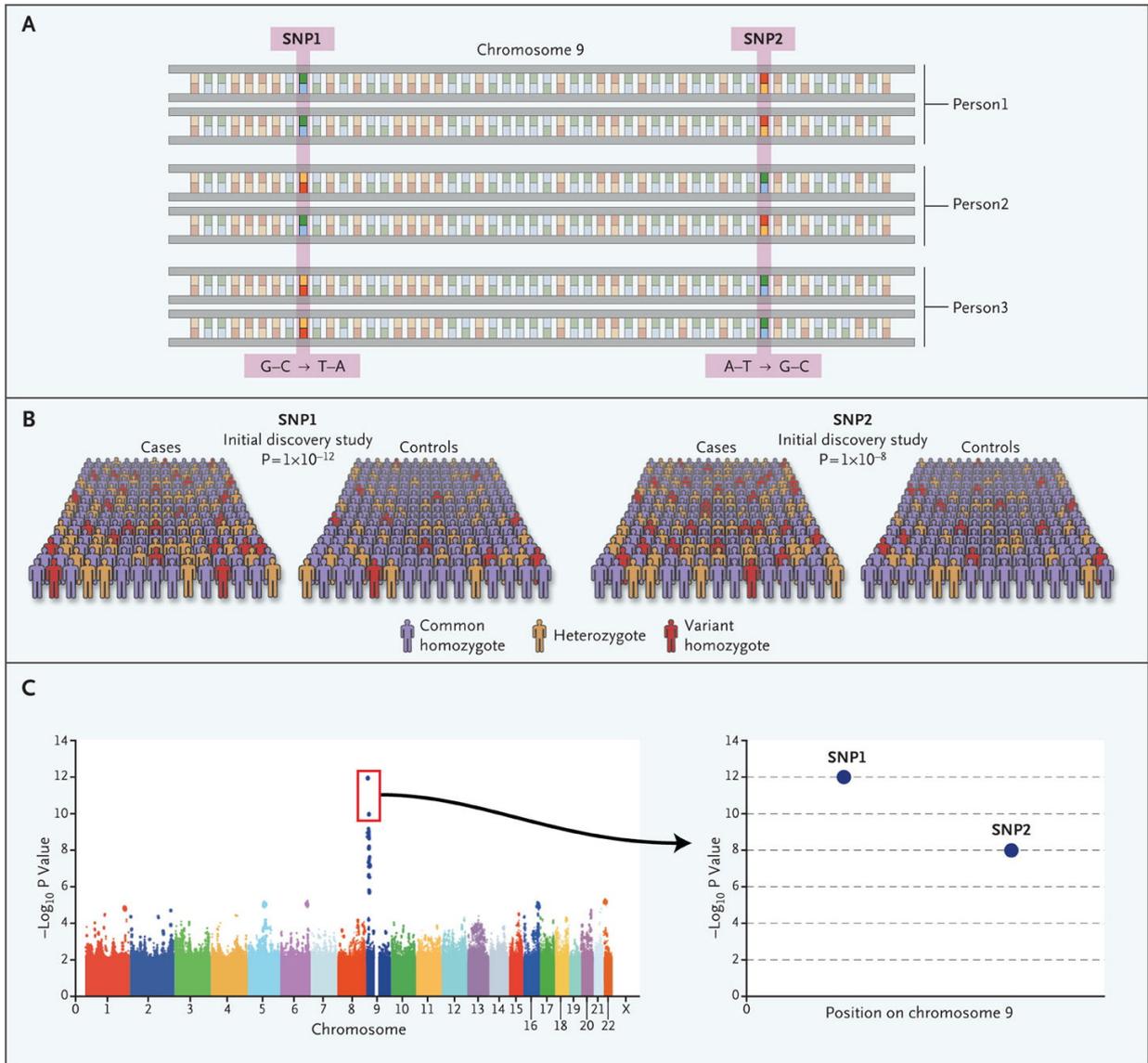


Figure 2.2 Genome-wide association studies (GWAS) (Manolio, 2010).

sclerosis (Baranzini, et al., 2009). Especially in complex diseases, which are intrinsically multifactorial, rather than identifying single genes, the identification of affected pathways would shed light into understanding of disease development mechanism.

2.4.2 Pathway and network oriented GWAS data analysis

Following its successful application on gene expression studies, the pathway analysis for GWAS is originated in the form of gene-set enrichment analysis (GSEA) by Wang et al. (Wang, et al., 2007). Since then, several different implementations of gene set enrichment for genome-wide pathway analysis of SNP-chip datasets have been

published (Askland, et al., 2009; Baranzini, et al., 2009; Chen, et al., 2010; Elbers, et al., 2009; Holmans, et al., 2009; Neibergs, et al., 2010; Peng, et al., 2010; Purcell, et al., 2007; Wang, et al., 2010; Weng, et al., 2011; Zhang, et al., 2011; Zhang, et al., 2010). Comparative evaluation of some of these existing pathway based GWAS data analysis platforms are shown in Table 2.2. The review of these tools and issues related to GWAS pathway analysis can be found in (Cantor, et al., 2010).

Pathway-based approaches are thought to complement the most-significant SNPs/genes approach and provide additional insights into interpretation of GWAS data on complex diseases (Askland, et al., 2009; Baranzini, et al., 2009; Elbers, et al., 2009; Peng, et al., 2010). These pathway-based GWASs are based on the hypothesis that multiple genes in the same biological pathway contribute to disease etiology, whereas common variations in each of these genes make mild contributions to disease risk. The use of prior knowledge in the form of pathway databases is demonstrated in GWAS of diseases such as Parkinson's disease, age-related macular degeneration, bipolar disorder, rheumatoid arthritis, and Crohn's disease (Lesnick, et al., 2007; Pattin and Moore, 2008; Torkamani, et al., 2008; Wang, et al., 2007; Wilke, et al., 2008). While the concept of pathway analysis for GWAS is attractive, it is restricted by our limited knowledge of cellular processes.

Since the analysis of single variants within isolated genes is not informative enough to explain the underlying disease mechanisms, another recent trend to further mine GWAS data is to incorporate network-based analysis (Bakir-Gungor and Sezerman, 2011; Barabasi, et al., 2011; Baranzini, et al., 2009; Barrenas, et al., 2009; Feldman, et al., 2008; Franke, et al., 2006; Lage, et al., 2007; Menon and Farina, 2011; Pattin and Moore, 2008; Tu, et al., 2006). However, some of these studies either do not use actual genetic (genotypic) data or are applied to model organisms. To the best of our knowledge, the only study to date that uses both a protein interaction network and pathway analysis to reveal significant disease related genes and pathways in genetic association studies is conducted by Baranzini et al. (Baranzini, et al., 2009) on Multiple Sclerosis. Since this study is gene centered, it is possible that true associations with markers that lie in large intergenic regions were neglected and the analysis is limited to the known functional properties of genes.

Table 2.2 Comparison of pathway based GWAS data analysis platforms (Yaspan and Veatch, 2011).

Method used	Pathway databases	Requires original dataset	LD correction	Gene size correction	Study designs allowed	Reference	Download link	Notes
ALIGATOR	GO	No	None (assumes LD is equal across pathways)	Yes	Any	Holmans et al. (2009)	http://x004.psych.uwcm.ac.uk/~peter	Controls for gene size and number of SNPs tested through use of permuted gene lists
GeSBAP	GO, KEGG BioCarta	No	None	No	Any	Medina et al. (2009)	http://bioinfo.cipf.es/gesbap/	Web-based interface
GRASS	Any	Yes	eigenSNP (randomizes case and control status)	Yes	Case-control ^a	Chen et al. (2010)	http://linchen.fhcrc.org/grass.html	Most powerful when several genes in the pathway are associated with disease risk
Ingenuity Pathway Analysis	Ingenuity Knowledge Base	No	Done prior to IPA	Done prior to IPA	Any	http://www.ingenuity.com	http://www.ingenuity.com/products/pathways_analysis.html	Commercial product; web-based interface
PARIS	KEGG, GO, Reactome, NetPath, Pfam, DIP, personalized	No	Genomic randomization	Yes	Any	Yaspan et al. (2011)	http://chgr.mc.vanderbilt.edu/ritchielab/subscriptions	Currently only for Caucasian datasets
PLINK	Any	Yes	Case-control status randomization	No	Any	Purcell et al. (2007)	http://pngu.mgh.harvard.edu/~purcell/plink/	Instructions at http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml#set
proxyGeneLD	Any	No	Bonferroni-type based on # of LD blocks and LE SNPs	Yes	Any	Hong et al. (2009)	http://ki.se/ki/jsp/polopoly.jsp?d=26072&l=en	Currently only for Caucasian datasets
SRT	Any	Yes	Case-control status randomization	Yes	Case-control ^a	O'Dushlaine et al. (2009)	https://sourceforge.net/projects/snpratiotest/	Designed to work with PLINK input/output
VSEA	Any	Yes	Gene-score normalization	Yes	Case-control ^a	Yang et al. (2011)	Component of R package	Produces normalized gene score and uses GSEA to calculate enrichment of gene set

Another important piece of information that could improve the analysis of GWAS datasets is the functional effect of a SNP. To better understand the biological processes underlying complex diseases, in this thesis, in addition to the pathway and network based approaches, we considered the functional effect of a typed SNP in GWAS. While the DNA polymorphisms that change protein function can have very significant consequences, such as NOD2 mutations in inflammatory bowel disease (Hugot, et al., 2001) and FLG mutations in eczema (Palmer, et al., 2006), other types of SNPs, such as synonymous SNPs do not have such serious effects in disease development mechanism. Hence, functionally important SNPs, such as those that change amino acids, splicing sites; those that lead to gain or loss of stop codon; those that result in frame shift; those that are found in regulatory region (including known transcription factor binding sites (TFBSs), DNase I hypersensitive sites which marks open chromatin, histone modification sites, CCCTC-binding factor (CTCF) sites which characterize insulator/enhancer elements) are priority targets in disease studies and large-scale genotyping projects (Calabrese, et al., 2009; Flicek, et al., 2010; Zhang, et al., 2011). There are a few existing web-servers that prioritize GWAS results based on the SNP's functional consequences, e.g. SPOT (Saccone, et al., 2010), SNPinfo (Xu and Taylor, 2009), ICSNPathway (Zhang, et al., 2011). Hence, we decided that SNP functional knowledge is valuable information to strengthen our pathway and network oriented GWAS analysis method. As summarized here, in order to mine GWAS results further, there are attempts to combine different sets of knowledge. Yet, to the best of our knowledge, none of these platforms can successfully integrate functional information of typed SNPs in a GWAS with LD analysis and protein protein interaction networks to identify SNP targeted pathways; and make a comparative evaluation between different populations.

2.4.3 GWAS on different populations

The potential of GWAS on disparate populations to uncover the links between genetics and pathogenesis of human complex diseases is discussed in the literature (Rosenberg, et al., 2010). One reason is that the risk variants can vary in their occurrence across populations (Goldstein, 2007; Goldstein and Hirschhorn, 2004). For example, while the

high-risk variant at MYBPC3 gene is observed with a frequency of ~4% in cardiomyopathy patients in Indian populations; this variant is rare or absent in other populations (Dhandapany, et al., 2009). Another reason is the difference in allele frequencies and biological adaptations among populations, which in turn affects the detectability and importance of risk variants. The identification of a variant might be easier in some populations compared to other populations since the particular histories of recombinations, mutations and divergences of genealogical lineages in the various populations affect the mappability of a variant. This situation is observed in the variants of TCF7L2 and KCNQ1 genes in type 2 diabetes (Adeyemo and Rotimi, 2010; Myles, et al., 2008). Also, in a review paper by Stranger *et al.* it has been pointed out that studying additional populations in GWAS may provide valuable insights for current and future research in medical genetics (Stranger, et al., 2011).

CHAPTER 3

3 MATERIALS AND METHODS

3.1 Materials

3.1.1 Datasets

3.1.1.1 GWAS datasets

RA, IA, PE, and Behçet's disease GWAS datasets are used within this thesis. The details of each dataset are explained below:

3.1.1.1.1 Rheumatoid arthritis dataset

We have applied our methodology on Wellcome Trust Case Control Consortium (WTCCC) Rheumatoid Arthritis (RA) dataset, in which 500,475 SNPs were tested on 5003 samples (1999 cases and 3004 controls) using Affymetrix GeneChip Human Mapping 500 K Array Set. SNP data and the genotypic p-values of association for each tested SNP were downloaded from the WTCCC project webpage (www.wtccc.org.uk). In total, 25,027 SNPs were included from WTCCC dataset, showing nominal evidence of association ($P < 0.05$).

3.1.1.1.2 Partial epilepsy dataset

We have used the dataset of Kasperaviciute et al's GWAS, which tested 3445 PE patients and 6935 controls of European ancestry (Kasperaviciute, et al., 2010). In that study, after the population structure analysis, 528,745 SNPs were included using the Human610-QuadV1 genotyping chips (Illumina). SNP data and the genotypic p-values of association for each tested SNP were obtained from <http://www.ion.ucl.ac.uk/departments/epilepsy/themes/genetics/PEvsCTRL>. Cochran–Mantel–Haenszel test results were used as the genotypic p-values of the identified SNPs.

3.1.1.1.3 Intracranial aneurysm European population dataset

The first IA GWAS dataset, that we used in this thesis, is a multicenter collaboration in Finnish, Dutch and Japanese cohorts totaling 5891 cases and 14,181 controls (Yasuno, et al., 2010). This study tested ~832,000 genotyped and imputed SNPs using the Illumina platform. In personal communication with the authors, upon our request, JP population specific data was removed and EU population specific results were obtained, including 2780 cases and 12,515 controls.

3.1.1.1.4 Intracranial aneurysm Japanese population dataset

The second IA GWAS dataset, that is used in this thesis, tested 312,712 SNPs on 1069 Japanese IA patients and 904 Japanese controls using the HumanHap300 or HumanHap300-Duo Genotyping BeadChips (Illumina) (Akiyama, et al., 2010). For both IA datasets, SNP data and the genotypic p-values of association for each tested SNP (calculated via Cochran-Armitage trend test) were obtained from our collaborators.

3.1.1.1.5 Behçet's disease Turkish population dataset

This GWAS is conducted on 1,215 Turkish Behçet's disease cases vs 1,278 unaffected controls (Remmers, et al., 2010). 311,459 autosomal SNPs were typed using the Infinium assay (Illumina), HumanCNV370-Duo v1.0 and HumanCNV370-Quad v3.0 chips.

3.1.1.1.6 Behçet's disease Japanese population dataset

This GWAS tested 500,568 SNPs on 612 Japanese individuals with Behçet's disease (cases) and 740 healthy controls (Mizuki, et al., 2010). DNA samples were typed using the Affymetrix GeneChip Human Mapping 500K Array Set.

For both Behçet's disease datasets, SNP data and the genotypic p-values of association for each tested SNP (calculated via allelic chi-squared test) were obtained from our collaborators.

3.1.1.2 Protein-protein interaction network

PPI network file, used within this thesis, is composed of two high quality systematic yeast two-hybrid experiments and PPIs obtained from literature by manual curation (Rual, et al., 2005; Stelzl, et al., 2005). The integrated set of PPIs contains 61,070 interactions between 10,174 genes. This file is obtained in the SIF format, which offers a straightforward means to import networks into Cytoscape as text.

3.1.1.3 IA gene expression dataset for Japanese population

A list of differentially expressed genes along with their p-values was obtained from the study of Krischek *et al.* (Krischek, et al., 2008). In this study, four unruptured and six ruptured IA specimens, which were collected during 42 months, were used as cases. Four arteriovenous malformation feeders, which were obtained during microsurgical

resection, were used as intracranial control tissue. The average age of the IA patients was 56.4 years, and that of the controls was 60.25 years. All patients and controls were of Japanese ethnicity. All tissue samples were profiled using oligonucleotide microarrays (Agilent Technologies). In the original study, in order to find out the differentially expressed genes between the aneurysmal cases and the controls, the analytical tools in the GeneSpringGX v11 was utilized. The statistical significance of the difference between the gene expression levels was calculated via the Student's t-test (Krischek, et al., 2008). In our study, we used these genes showing significant difference at the false discovery rate of 0.05 according to the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995).

3.1.2 Computational equipment setup

A computer with Windows or Linux operating system and internet access is required to follow the steps of the PANOGA protocol, which is developed within this thesis. We recommend a 1 GHz CPU or higher, a high-end graphics card, 500MB of available hard disk space, at least 1 GB of free physical RAM and a minimum screen resolution of 1,024 x 768.

3.1.2.1 Java platform

We recommend to install the Standard Edition of Java, version 5.0 or higher (Java SE 5 or higher). (<http://java.sun.com/javase/downloads/index.jsp>).

3.1.2.2 Cytoscape

Cytoscape is an open source network data integration, analysis, and visualization platform (Cline, et al., 2007; Shannon, et al., 2003). Subnetwork identification and functional enrichment steps of PANOGA protocol are realized by Cytoscape plugins. Hence, to follow PANOGA protocol, users need to install Cytoscape version 2.6.3 by on a local computer by following the steps in Box 2 of the Cytoscape paper, published in

Nature protocols (Cline, et al., 2007). Although Cytoscape has newer versions, jActiveModules and ClueGO plugins are verified to work in Cytoscape version 2.6.3.

3.1.2.3 SNP functionalization tools

PANOGA protocol utilizes four external web-servers to functionalize SNPs, i.e., SPOT (Saccone, et al., 2010), F-SNP (Lee and Shatkay, 2008), SNPnexus (Chelala, et al., 2009), SNPinfo (Xu and Taylor, 2009); jActiveModules plugin (Ideker, et al., 2002) of Cytoscape (Shannon, et al., 2003) to identify sub-networks; ClueGO plugin (Bindea, et al., 2009) of Cytoscape (Shannon, et al., 2003) for functional enrichment of the identified sub-networks. All of these web-servers, programs and plugins are freely available for academic use.

3.2 Methods

3.2.1 Design of Pathway and Network Oriented GWAS Analysis (PANOGA) Tool

3.2.1.1 PANOGA Overview

Starting with a list of SNPs, found to be associated with disease in GWAS, in this thesis, we propose a novel methodology to determine disease related pathways through the identification of SNP targeted genes within these pathways. PANOGA is the first algorithm that integrates functional information of typed SNPs in a GWAS with LD analysis and protein protein interaction networks to identify SNP targeted pathways. With its multifactorial basis, PANOGA has a good potential to decipher the combination of biological processes underlying disease and to aid the development of novel therapies at molecular level. PANOGA has been tested on several complex diseases including rheumatoid arthritis (Bakir-Gungor and Sezerman, 2011), epilepsy (Bakir-Gungor, et al., 2012, submitted), intracranial aneurysm (Bakir-Gungor and Sezerman, 2012, submitted), and Behçet's disease; and proved to be useful. Throughout the time, the method has evolved with our efforts, and here, we present the latest version of PANOGA. In this methodology, GWAS results are used in the form of SNP

rs ids vs. p-values, where the p-values refer to the genotypic p-values of association for each tested SNP. We only focused on SNPs with nominal evidence of association ($P < 0.05$) in a GWAS, following the study in (Baranzini, et al., 2009).

PANOGA proceeds in nine steps, as outlined in Figure 3.1. Briefly, step (i) of PANOGA utilizes SPOT (Saccone, et al., 2010) and F-SNP (Lee and Shatkay, 2008) web-servers to obtain functional information of a SNP. In step (ii), PANOGA combines the functional scores obtained from SPOT and F-SNP web-servers with GWAS p-values; and it calculates a weighted p-value, P_w , for each score (15). In step (iii), SNPs are assigned to genes using SPOT's SNP to gene assignment module (Saccone, et al., 2010). In step (iv), SPOT (Saccone, et al., 2010) and F-SNP (Lee and Shatkay, 2008) P_w -values are assigned to each gene as two separate attributes. If more than one SNP is assigned to the same gene in step (iii), SPOT and F-SNP P_w values of all these SNPs are taken into account and lowest SPOT and F-SNP P_w values are assigned to the gene. In step (v), a possible overlap of the input SNPs with known Transcription Factor Binding Site (TFBSs) at TRANSFAC (Wingender, et al., 2000) is also checked. If this transcription factor (TF) is not already found in step (iii), this TF is added to our list by transferring its SPOT and F-SNP P_w -values from its associated SNP. In step (vi), genes with two separate weighted P-values (P_w values) are mapped to a human protein protein interaction network. By using the P_w values of the genes and network topology, step (vii) aims to find out active sub-networks in the human PPI network using jActive Modules algorithm (Ideker, et al., 2002). Although this algorithm was originally developed for microarray gene expression data, steps (i)-(v) of PANOGA successfully adapts GWAS data to be used with this algorithm. In terms of GWAS data, jActive Modules algorithm integrates the network topology with the calculated P_w -values of each gene to extract potentially meaningful active sub-networks. Following the identification of sub-networks, we evaluated whether these sub-networks were biologically meaningful in step (viii) of PANOGA, using functional enrichment techniques. In step (ix), we integrate the functional enrichments of the generated sub-networks. KEGG pathways that might have a role in disease mechanism are identified via including the pathway, if it is found as significant for at least one of the identified sub-networks. For each identified KEGG pathway in our final list, PANOGA counts the

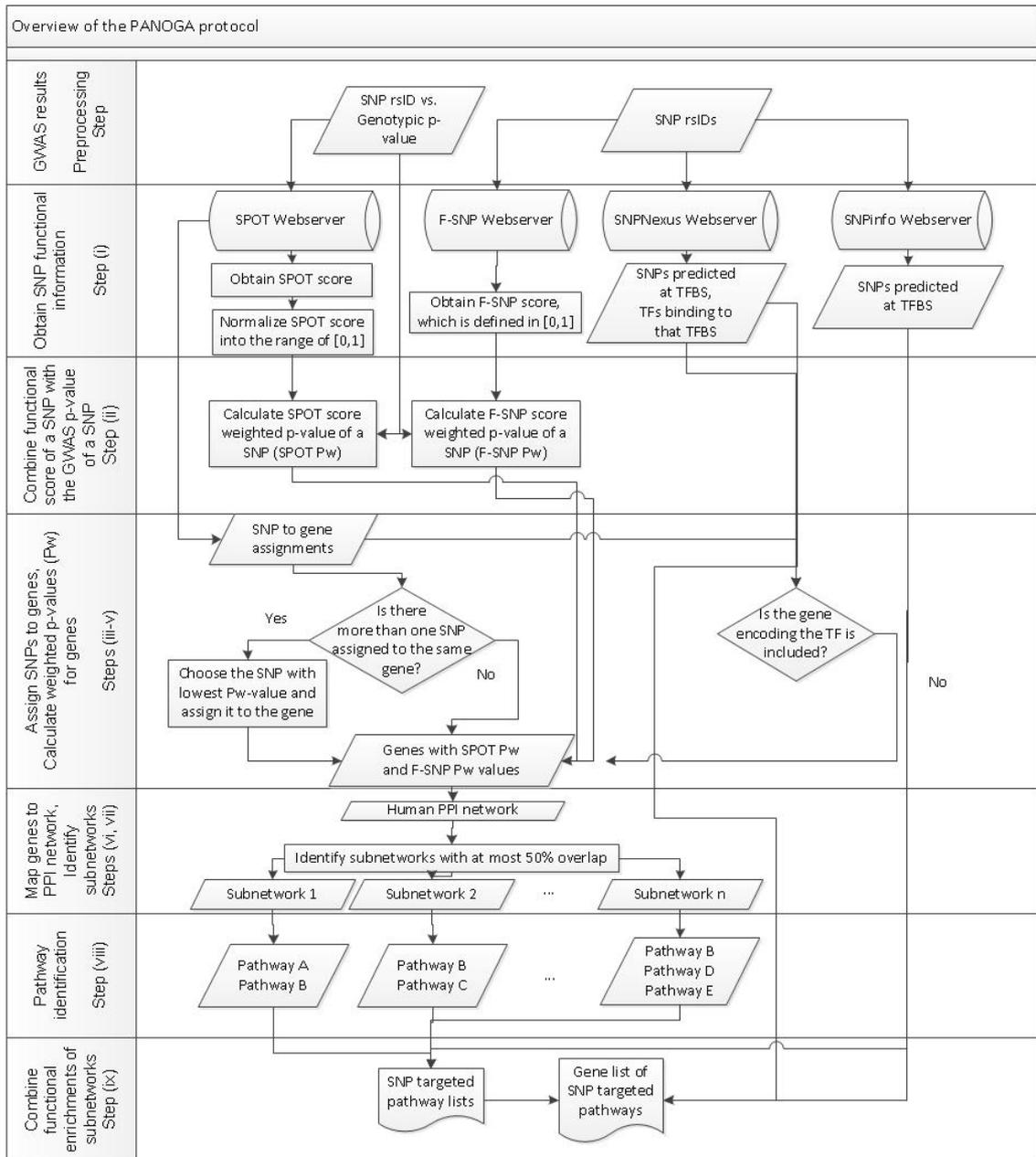


Figure 3.1 Outline of PANOGA’s assessment process. In steps (i) to (v), a gene-wise Pw-value for association with disease was computed by integrating functional information. In step (vi), Pw-values were loaded as two separate attributes of the genes in a PPI network. In step (vii), active sub-networks of interacting gene products that were also associated with the disease, are identified. In step (viii), genes in an identified active sub-network were tested whether they are part of functionally important KEGG pathways. Lastly, step (ix) integrates the functional enrichments of the generated sub-networks.

number of associated SNPs from GWAS, the number of regulatory SNPs (SNPs located on TFBSs or miRNAs) among those disease predisposing SNPs, the number of SNP-targeted genes, the number of sub-networks that this pathway is found to be statistically

significant. Due to its modular design pattern, PANOGA protocol gives flexibility to the user and it is very easy to adapt novel datasets or more precise tools, e.g. better, more comprehensive SNP functionalization tools, higher quality, higher coverage PPI networks. We further describe each module below.

3.2.1.2 SNP functionalization

SNPs might have different functional impacts such as: an effect on transcriptional regulation by changing TFBS's activity; premature termination of amino-acid sequence (generate a stop codon); alteration in the splicing pattern or efficiency by disturbing splice site, exonic splicing enhancers (ESE) or silencers (ESS); a change in protein structures or properties by altering single amino acids or changing the frame of the protein-coding region; regulation of protein translation by affecting microRNA (miRNA) binding sites activity. To predict such functional properties of SNPs, many different web tools are developed, and a comprehensive comparison of these tools can be found in (Karchin, 2009). Among these tools, we have decided to combine the scores of SPOT (Saccone, et al., 2010) and F-SNP (Lee and Shatkay, 2008) servers as following. SPOT score (Saccone, et al., 2010) takes into account SNP/gene transcript functional properties (including nonsense, frameshift, missense and 5' and 3'-UTR designations), impact of an amino acid substitution on the properties of the protein product from PolyPhen server (Adzhubei, et al., 2010; Ramensky, et al., 2002), evolutionary conserved regions from ECRbase (Loots and Ovcharenko, 2007), and all possible LD proxies - SNPs with r^2 over a predefined threshold in a specific HapMap sample (Frazer, et al., 2007). Hence, in the SNP functionalization step, PANOGA captures the functional consequences of other candidate SNPs that are in the same LD based on the HapMap data. On the other hand, F-SNP score (FS score) reflects the deleterious effect of a SNP, where the functional consequence of a SNP is obtained from multiple independent tools at four major categories (i.e. splicing, transcriptional, translational and post- translational levels) (Lee and Shatkay, 2009). FS scores of known disease-related SNPs, which are collected from OMIM, are previously shown to be significantly different from FS scores neutral SNPs (Lee and Shatkay, 2009). The details of the data sources used in our functional score can be found in Table 3.1. In the preprocessing step of PANOGA, SNPs are submitted to the above mentioned SNP

Table 3.1 Description of data sources used in our functional score.

Functional Category	Tool	Description	Meta-tool
Protein Coding	LS-SNP, SNPs3D, SIFT, SNPeffect	SNP annotation tool, Impact of nsSNPs on protein function, Prediction of amino acid substitution effects, SNP annotation with human disease	F-SNP
Protein Coding	PolyPhen	Prediction of amino acid substitution effects	SPOT, F-SNP
Protein Coding, Splicing Regulation, Transcriptional Regulation	Ensembl	Extensive genomic database including SNPs and gene transcripts	F-SNP
Splicing Regulation	ESEfinder, ESRSearch, PESX, RescueESE	Exonic splice sites, Exonic-splicing regulatory (ESR) sequences, Exon splicing enhancers/silencers, Exonic splice sites	F-SNP
Transcriptional Regulation	Consite TFSearch	Conserved transcription factor binding sites, Transcription factor binding sites	F-SNP
Transcriptional Regulation	SNPnexus	Conserved transcription factor binding sites	SNPnexus
Transcriptional Regulation, Conserved Region	GoldenPath	MicroRNA, cpgIslands, evolutionary conserved regions	F-SNP
Conserved Region	ECRBase	Evolutionary conserved regions	SPOT
Post-translation	KinasePhos, OGPET, Sulfinator	Phosphorylation sites, Prediction of O-glycosylation sites in proteins, Tyrosine sulfination sites	F-SNP
Genomic Coordinates	dbSNP	General SNP/gene transcript properties	SPOT
Genomic Coordinates	UCSC	Extensive genomic database including SNPs and gene transcripts	F-SNP
LD estimation	HapMap, Haploview	Dense genotyping on multiple populations, useful for LD estimates Estimation of r^2 LD coefficients for each population	SPOT

functionalization web-servers. Step (i) of PANOGA obtains results from these web-servers and normalizes if needed. FS Score is defined in the range of [0,1], where 0 means the functional consequence of a SNP on the gene product is negligible and 1 means the functional consequence of the SNP on the gene product is serious (Lee and Shatkay, 2009). SPOT scores are not limited to a range of [0,1] and hence we normalized SPOT scores to this range in step (i).

If a SNP lies within the transcription factor binding site (TFBS) of a gene, it may disrupt the level or timing of gene expression. We used two other SNP functionalization web-servers within PANOGA in step (i), i.e. SNPnexus (Chelala, et al., 2009), SNPinfo (Xu and Taylor, 2009) to evaluate whether the GWAS SNPs interfere transcriptional

regulation by affecting TFBS's activity. SNPnexus (Chelala, et al., 2009) checks for a possible overlap of a SNP with conserved TFBSs from TRANSFAC Matrix Database, v.7.0, (Wingender, et al., 2000) and returns the related TF name. On the other hand, SNPinfo determines whether the alternative alleles of a SNP, which is located in the TFBS, have a different activity than usual and returns the rsIDs of such SNPs (Xu and Taylor, 2009). As shown in the step (v) of Figure 3.1, we used the SNPnexus results as part of the assigning SNPs to genes step, as described in detail below. We incorporated SNPinfo results in the last step, step (ix), as shown in Figure 3.1.

3.2.1.3 SNP-wise weighted p-value calculation

To combine biological information with evidence for genetic association, the following scoring scheme is proposed in (Saccone, et al., 2008). In (Saccone, et al., 2008), firstly, a non-negative prioritization score (PS) was specified for each SNP and then, the weighted P-value P_w is defined by $P_w = P/10^{PS}$ (Roeder, et al., 2006; Saccone, et al., 2008), where P denotes GWAS P-value for a particular SNP. In this scheme, smaller values of P_w indicate higher priority. Following this convention, in step (ii), for each SNP, we have calculated SPOT P_w -value using SPOT prioritization score and F-SNP P_w -value using F-SNP prioritization score.

3.2.1.4 SNP to gene assignment

It is hypothesized that meaningful combination of genes harboring markers with only modest evidence of association can be identified if they belong to the same biological pathway or mechanism (Baranzini, et al., 2009). Therefore, the gene and pathway-based association analysis allows us to gain insight into the functional basis of the association and facilitates to unravel the mechanisms of complex diseases. However, a SNP may be associated with many genes, i.e. it can be located in a gene with several known transcripts due to alternative splicing, or in one gene and very close to another gene, or at the intersection of different genes on different strands and hence a SNP may have different functional consequences on each transcript. In step (iii), SNPs are assigned to genes using SPOT's SNP to gene assignment module (Saccone, et al., 2010). SPOT

considers all known SNP/gene transcript associations and assigns the SNP to the gene with the highest priority (Saccone, et al., 2010). To generate those SNP/gene transcript associations, SPOT program utilizes information from the PolyPhen method of predicting the effect of an amino acid substitution on the properties of the protein product (Adzhubei, et al., 2010; Ramensky, et al., 2002). Those effects can be directly detected from DNA and RNA sequences, like nonsense and missense amino acid substitutions, untranslated regions, coding regions, and frameshifts. Hence, by prioritizing all known SNP/gene transcript consequences, propitious association signals found in GWAS, are not lost at the SNP to gene transition step. At this stage, PANOGA creates a gene list including the gene symbols which are associated with GWAS SNPs.

3.2.1.5 Gene-wise weighted p-value calculation

Since SNPs are associated with genes in step (iii) of our method, these two weighted p-values (Pw-values) can be automatically transferred into the SNP's associated gene as two separate attributes. Hence, in step (iv), each gene has a SPOT Pw-value and a F-SNP Pw-value, indicating the association with the disease (gene-wise Pw-values). If more than one SNP is assigned to the same gene in step (iii), SPOT and F-SNP Pw values of all these SNPs are taken into account and lowest SPOT and F-SNP Pw values are assigned to the gene. In other words, the SPOT Pw-value of a gene is calculated as the lowest SPOT Pw-value of the SNP that is assigned to that particular gene among all the SPOT Pw-values of the SNPs assigned to the same gene. Same is true for F-SNP Pw-value. A possible overlap of the input SNPs with known TFBSs is already checked in the SNP functionalization step, step (i). If the related TF is not already found in Step (iii), this TF is added to our list by transferring its SPOT and F-SNP Pw-values from its associated SNP in step (v). At the end of this step, step (v), PANOGA returns a list of genes with SPOT and F-SNP Pw-values.

3.2.1.6 Active sub-network identification

By using weighted p-values of the genes, as calculated at the end of the step (v), this step aims to find out active sub-networks in the human PPI network. Here, an active

sub-network refers to a connected subgraph of the interactome that has high total significance of genotypic p-values of the disease-predisposing SNPs with respect to the controls. It should be noted that in jActive Modules algorithm, an identified sub-network with a high score is not necessarily the sub-network that includes the genes with very significant genotypic p-values. Instead, the identified sub-network can be composed of many genes with moderately significant genotypic p-values. Hence, jActive Modules algorithm helps to discover groups of genes that display seemingly negligible association with disease when evaluated individually, but display strong association when considered as a group.

In step (vi), PANOGA maps the genes with two separate Pw-values (SPOT and F-SNP Pw-values) into a human PPI network. In step (vii), active sub-networks of interacting gene products, that were possibly associated with the disease, are identified using jActive Modules (Ideker, et al., 2002). Basically, jActive Modules (Ideker, et al., 2002) is a Cytoscape plugin that identifies active sub-networks via incorporating both the topological properties of a PPI network and the attributes of the nodes (proteins). In this approach, firstly the attributes (SPOT and F-SNP Pw-values) are mapped into biological networks, secondly a statistical measure (as explained below) is used to score sub-networks based on the attributes, and finally a search algorithm is used to find active sub-networks with high score.

Biologically speaking, an active sub-network (statistically significant module) is a sub-network in our PPI network that the protein products of this set of genes – probably associated to the disease- also physically interact, thus raises the possibility that they belong to the same pathway or biological process. To rate the biological activity of a particular sub-network, jActive Modules starts by assessing the significance of differential association with disease for each gene (by comparing the gene-wise Pw-values of association with the disease). In this procedure, jActive Module samples p-values from the distribution of p-values loaded into Cytoscape, and not from a normal uniform distribution. Then, a network is generated from each node by systematically adding one neighbor at a time. The aggregate z-score (S) of an entire sub-network, consisting of k genes is calculated via summing the scores of all genes z_i in the sub-network and then dividing by the square-root of k. To extend the z-score over multiple

conditions (attributes), jActive Module sorts z-scores for each attribute, adjusts for rank, maximum score is corrected using the background score distribution (Ideker, et al., 2002). The scoring system of jActive Modules ensures that the expected mean and variance of the subgraph scores are independent of subgraph size (Ideker, et al., 2002). jActive Modules plugin also corrects for the fact that a bigger sub-network is more likely to contain nodes with significant p-values by random chance (Ideker, et al., 2002). When S stops to increase, the sub-network stops growing and is reported as a module. Next, the test statistic (S) is compared with an appropriate background distribution to properly capture the connection between network topology and association with disease. As a background distribution, we used the scores of sub-networks randomly selected from the entire human PPI network, as provided by jActive Modules. In order to make the background distribution independent of the module size, jActive Modules creates a background distribution by scoring 10,000 random sub-networks of each size in a Monte Carlo procedure. In our study, modules with $S > 3$ were reported as significant (active sub-network), as stated in the original publication (Ideker, et al., 2002).

3.2.1.6.1 Overlap threshold parameter

At the initial version of PANOGA, we focused only on the highest scoring sub-network. But later we noticed that the scores of the identified sub-networks were very close to each other. We also realized that the highest scoring sub-network does not cover the initial PPI network and thus, we lose information. That is why in the improved PANOGA, we decided to combine the pathway enrichment results of the identified sub-networks. At this stage, due to the nature of the search algorithm, several of these sub-networks overlap extensively in their component genes. While we wanted to cover the whole PPI network with the identified sub-networks as much as possible, we did not want to include the same genes over and over in our sub-networks. To this end, starting with 0, we experimented PANOGA with 10% increments of the overlap threshold values, where this parameter defines the max level of identity between the constituent genes of any two identified sub-networks. The coverage of the PPI network was 0.02, 0.22, 0.376, 0.462, 0.391, 0.321 for overlap threshold values 0, 0.2, 0.4, 0.5, 0.6, 0.8, respectively. Hence, we set this parameter as 50% to balance between disabling

repetitive sets of identical genes (obtained via high overlap threshold parameter) and enabling moves to the different parts of the network (low overlap threshold parameter). So, in the current version of PANOGA, rather than focusing on the highest scoring sub-network, we find all significant sub-networks (with $S > 3$) that overlap less than 50% with each other.

3.2.1.7 Functional enrichment, pathway identification

Next step following the identification of sub-networks is to evaluate whether these sub-networks were biologically meaningful. For each sub-network, in step (viii), we compute the proportion of the genes in an identified sub-network that are also found in a specific human biochemical pathway, compared to the overall proportion of genes described for that pathway. For this purpose, ClueGO plugin (Bindea, et al., 2009) of Cytoscape (Shannon, et al., 2003) is utilized in this step. ClueGO is an open-source Java tool that extracts the non-redundant biological information for groups of genes using GO, KEGG and BioCarta ontologies (Bindea, et al., 2009). Unlike other functional enrichment analysis tools (Boyle, et al., 2004; Huang, et al., 2007; Maere, et al., 2005; Ramos, et al., 2008; Zeeberg, et al., 2003) that present their results as long lists or complex hierarchical trees; ClueGO facilitates the biological interpretation via visualizing functionally grouped terms in the form of networks and charts (Bindea, et al., 2009). To link the terms in the network, ClueGO uses kappa statistics, in a similar way as described in (Huang, et al., 2007). Among different ontologies, since KEGG database primarily categorizes genes into bona-fide biological pathways; and since biological interpretation of pathways is more straightforward compared to GO terms, we report only our functional enrichment results using KEGG pathways. We used two-sided (Enrichment/Depletion) test based on the hypergeometric distribution to examine the association between the genes targeted by disease predisposing SNPs and the genes in each KEGG pathway. To correct the P-values for multiple testing, Bonferroni correction procedure is applied (Bindea, et al., 2009). Since PANOGA identifies hundreds of active sub-networks with $S > 3$, in step (viii) we used the command-line version of ClueGO_v1.4 (Bindea, et al., 2009).

3.2.1.8 Integration of the functional enrichments of the generated subnetworks

While an identified sub-network represents only one part of the whole interaction network, the identified pathways for this sub-network represents one aspect of the disease. Since the human complex diseases are multifactorial, via discovering the pathways from different sub-networks, we aimed to enlighten different aspects of the disease. To this end, step (ix) integrates the functional enrichments of the generated sub-networks. If a KEGG pathway is found to be statistically significant for at least one of the active sub-networks with S score >3 , PANOGA adds this pathway into our final list of significant KEGG pathways as associated with disease. At this step, PANOGA calculates the significance of a pathway in relation to disease as the minimum p-value of the enrichment test, among all p-values calculated for this pathway during the enrichment of each identified sub-network. The pathways are ranked according to the significance scores and are referred as SNP targeted pathways.

3.2.2 Development of a protocol to identify SNP targeted pathways from GWAS

Following the design of PANOGA, as explained in Section 3.2.1, we implemented a protocol to devise functionally important pathways through the identification of genes within these pathways, where these genes are targeted by SNPs obtained from the GWAS analysis. The protocol, developed within this thesis, is publicly available at: http://akademik.bahcesehir.edu.tr/~bbgungor/panoga_protocol.zip

PANOGA protocol is composed of 43 steps, as summarized in Figure 3.2. Briefly, the preprocessing step of PANOGA is realized by a java script (createpanogainput.jar); SNP functionalization steps of PANOGA are realized via sending the input files into four different web-servers (SPOT (Saccone, et al., 2010), F-SNP (Lee and Shatkay, 2008), SNPnexus (Chelala, et al., 2009), SNPinfo (Xu and Taylor, 2009)); the subnetwork identification step of PANOGA is realized by the jActiveModules plugin (Ideker, et al., 2002) of Cytoscape (Shannon, et al., 2003); the remaining steps are performed via running java executable programs. The detailed instructions for each step are presented below.

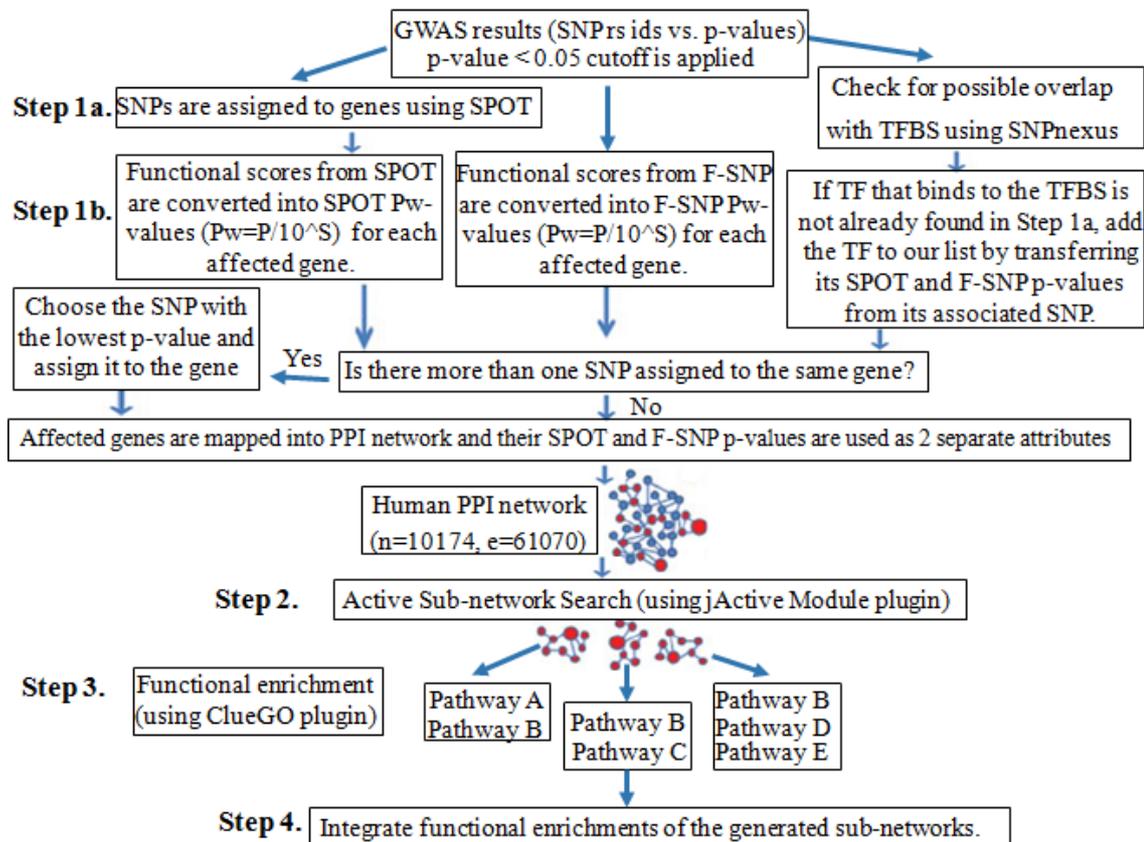


Figure 3.2 Summary of PANOGA protocol. In Step 1, a gene-wise Pw-value for association with disease was computed by integrating functional information. In Step 2, significant Pw-values were loaded as two separate attributes of the genes in a PPI network and visualized using Cytoscape [20]. At this step, active sub-networks of interacting gene products that were also associated with the disease are identified using jActive Modules plugin plugin (Ideker, et al., 2002). In Step 3, genes in an identified active sub-network were tested whether they are part of functionally important KEGG pathways. Step 4 integrates the functional enrichments of the generated sub-networks.

3.2.2.1 PANOGA input files' formats

3.2.2.1.1 GWAS dataset file format

As an input file, PANOGA protocol requires GWAS result of a disease saved in a tab delimited text file (.txt) or excel file (.xls) including “SNP rs id” and “p-value” information. Here, the p-value refers to the genotypic p-value of association for each tested SNP. In this input file, the user should include only the SNPs with nominal evidence of association ($P < 0.05$, or a user defined threshold) in a GWAS. It is important to note that PANOGA protocol does not require individual genotypes, odds

ratio (OR), minor allele frequency (MAF), or confidence intervals (CI) computed in a GWAS, which can have ethical considerations. A sample input file might look like this:

```
rs1320565    0.0354782368664204
rs2887286    0.0485440172506189
rs12736358   1.85031556014792e-05
rs10102164   3.40287797939709e-11
```

In this GWAS result input file, SNP rs ids are unique and the p-values are listed using dot after first digit and with e- or E- notation for exponentials.

CRITICAL STEP A different notation of the p-values other than the above mentioned format may block the PANOGA procedure.

Because of its basic format, PANOGA input file can be easily created either manually by a user (e.g., in Excel) using GWAS results or programmatically by a text-processing script. A sample PANOGA input file, `sample_panoga_input.txt` is provided under `PANOGA_protocol/data/sample/`.

3.2.2.1.2 Protein-protein interaction network file format

Cytoscape program (Shannon, et al., 2003) realizes the network oriented steps of PANOGA protocol, and it accepts a variety of file formats for importing networks, e.g., .sif, .gml, .xgmml, .xls, SBML, BioPAX, PSI-MI. A brief description of these file formats are presented in (Cline, et al., 2007) and the details of these file formats can be found at: http://wiki.cytoscape.org/Cytoscape_User_Manual#Supported_Network_File_Formats

Although Cytoscape (Shannon, et al., 2003) allows the usage of various file formats, PANOGA users are encouraged to use Simple Interaction File (SIF or .sif) file format, due to its simplicity to create either manually by a user (e.g., in Excel) or programmatically by a text-processing script. As a network input file, a sample human protein-protein interaction file is provided at: `PANOGA_procedure/data/humanPPI.sif`. This .sif file looks like as following:

```
geneSymbolA pp geneSymbolB
geneSymbolA pp geneSymbolC
```

geneSymbolC pp geneSymbolD

The first line of this file indicates that proteinA that is produced by geneSymbolA interacts with proteinB that is produced by geneSymbolB. Here “pp” refers to ‘protein-protein’ interaction type. In a typical sif file, the interaction type might be one of the following relationships: ‘protein-protein’, ‘degrades’ or ‘phosphorylates’.

CRITICAL STEP For best results, use ‘pp’ interaction type in the sif formatted file, because PANOGA protocol uses undirected network.

CRITICAL STEP Use standard HUGO gene symbols (Seal, et al., 2011) as node identifiers in the sif formatted network file. Because the node attributes file that PANOGA protocol generates uses official HGNC gene symbols as node identifiers and Cytoscape does not allow to import node attributes if the identifier types used in the network file and in the attributes file do not match.

3.2.2.2 Procedure

3.2.2.2.1 Install PANOGA

- 1) Set up necessary environment to run PANOGA (as detailed in EQUIPMENT SETUP).
- 2) Download the PANOGA files at: http://akademik.bahcesehir.edu.tr/~bbgungor/panoga_protocol.zip. Unzip the downloaded PANOGA_protocol.zip file and extract it. The executable jar files of PANOGA are found at: PANOGA_protocol/.

3.2.2.2.2 Preprocess GWAS data

- 3) Pick a disease name for your project, which can be any disease name (e.g.,

diabetes), not necessarily a standard OMIM disease name. In the following steps of PANOGA procedure, we will refer to this disease name as \$DISEASE_NAME.

CRITICAL STEP Do not use space in the \$DISEASE_NAME since it will corrupt the further steps of PANOGA procedure.

- 4) Create a folder with your disease name under PANOGA_protocol/data/ and under PANOGA_protocol/out/ via typing the following commands:

```
>cd PANOGA_protocol/data
```

```
>mkdir $DISEASE_NAME
```

```
>cd ../out
```

```
>mkdir $DISEASE_NAME
```

```
>cd ..
```

Replace \$DISEASE_NAME above with the disease name that you specified in Step 3.

- 5) Format GWAS results input file following the instructions in Box1, and save this file under PANOGA_protocol/data/\$DISEASE_NAME/ using any input file name. e.g., PANOGA_protocol/data/diabetes/diabetes_panoga_input.txt or bipolar_gwas_result.xls. sample_panoga_input.txt file is also provided under: PANOGA_protocol/data/sample/.

- 6) Run the java script “createpanogainput.jar” to create four separate input files that will be used in SNP to gene assignment and SNP functionalization steps of PANOGA:

Replace \$INPUT_FILE_NAME with your input file name, e.g. (diabetes_panoga_input.txt), \$DISEASE_NAME with your disease name and \$PVALUE_THRESHOLD with genotypic p-value threshold that you would like to use to restrict your SNPs based on their significance for disease. The default \$PVALUE_THRESHOLD is 0.05.

```
>java -jar createpanogainput.jar $INPUT_FILE_NAME $DISEASE_NAME $PVALUE_THRESHOLD
```

e.g. `java -jar createpanogainput.jar sample_panoga_input.txt sample 0.05`

This run generates `$DISEASE_NAME_spot_input.txt`, `$DISEASE_NAME_fsnp_input.txt`, `$DISEASE_NAME_snpnexus_input.txt`, `$DISEASE_NAME_snpinfo_input.txt` files under `PANOGA_protocol/data/$DISEASE_NAME`.

CRITICAL STEP Using an input filename with an extension other than `.txt` or `.xls` interferes this step.

3.2.2.2.3 Assign SNPs to Genes

- 7) PANOGA procedure uses SPOT webserver (Saccone, et al., 2010) to assign SNPs to genes. Go to the SPOT webserver at: <https://spot.cgsmc.isi.edu/submit.php>.
- 8) Click into “Upload SNP File” button; select SPOT input file, i.e. `$DISEASE_NAME_spot_input.txt`.
- 9) Change “Maximum SNPs to output:” parameter to 50,000 in SPOT webserver.
- 10) If your `$PVALUE_THRESHOLD` (from Step 6) is different than 0.05, change it in the “p-value threshold:” parameter of SPOT webserver.
- 11) Under “Linkage Disequilibrium (LD) options” select the appropriate HAPMAP sample among the available options in SPOT webserver.
- 12) Click into “Run” button and download the result under “Primary Results” section. Save the SPOT output as Tab-delimited file under `PANOGA_protocol/data/$DISEASE_NAME/$DISEASE_NAME_spot_output.txt`.
- 13) At this step, the users need to choose one of the following two options: option A to proceed with the full PANOGA procedure, including network oriented stages and functional information of SNPs; option B to proceed with only pathway oriented steps of PANOGA procedure. We highly recommend the users to follow the full PANOGA procedure (option A).

(A) Proceed with the full PANOGA procedure

Continue with Step 14.

(B) Proceed with only pathway oriented steps of PANOGA procedure

(i) Run the java script “parsespotoutput.jar” to get a list of gene symbols assigned into typed SNPs.

```
>java -jar parsespotoutput.jar $DISEASE_NAME
```

This run will create the gene symbol file (\$DISEASE_NAME_partial_panoga_gene_symbols.txt) under PANOGA_procedure/ClueGO/data/ and \$DISEASE_NAME_partial_panoga_gene2snp.txt file under PANOGA_procedure/data/\$DISEASE_NAME/.

(ii) Type the following command to perform functional enrichment of identified gene symbols:

```
>cd ClueGO
```

Replace \$DISEASE_NAME below with the disease name that you specified in Step 3.

```
>java -jar ClueGO_v1.4.command-line.jar -props clueGO.props -file1 data\$DISEASE_NAME_partial_panoga_gene_symbols.txt -analysis-name $DISEASE_NAME_partial_panoga -out out
```

At the end of this step, enrichment results of the gene symbols are saved under PANOGA_procedure/ClueGO/out/

(iii) Run the java script “analyzecluegooutput.jar” to create SNP targeted pathway lists and gene list for identified SNP targeted pathways.

```
>cd ..
```

```
>java -jar analyzecluegooutput.jar $DISEASE_NAME
```

At the end of this step pathway based lists and gene list are created as explained in the “Anticipated Results” section and “PANOGA’s

Application to Human Complex Diseases” subsection of Introduction section.

3.2.2.2.4 Install Cytoscape and its plugins

14) Install Cytoscape version 2.6.3 by following its installation guide (Cline, et al., 2007). Follow Cytoscape installation instructions to get the executable file.

CRITICAL STEP Although Cytoscape has newer versions, jActiveModules and ClueGO plugins are verified to work in Cytoscape version 2.6.3.

15) Install jActiveModules and ClueGO version 1.4 plugins of Cytoscape. These plugins should be installed into Cytoscape_v2.6.3/plugins/ using the following options: option A to install jActiveModules plugin; option B to install ClueGO version 1.4 plugin:

(A) Installing jActiveModules plugin

jActiveModules plugin is used to identify active sub-networks. Copy jActiveModules plugin from:

PANOGA_protocol/EXTERNAL_TOOLS/jActiveModules.jar

and save under Cytoscape_v2.6.3/plugins/.

(B) Installing ClueGO version 1.4 plugin

(i) ClueGO plugin is used in the functional enrichment step of PANOGA. Copy .cluegoplugin, provided under PANOGA_protocol/ClueGO/ into the home directory of the user.

(ii) Obtain ClueGO licence from its website (<http://www.ici.upmc.fr/cluego/cluegoLicense.shtml>) and save. If file under home/.cluegoplugin/License/.l/ and .lcf file under home/.cluegoplugin/License/.lc/.

CRITICAL STEP Before running PANOGA, ensure that Cytoscape, jActiveModules and ClueGO plugins are working properly.

3.2.2.2.5 Obtain Functional Information of SNPs

16) PANOGA procedure utilizes SPOT (Saccone, et al., 2010), F-SNP (Lee and Shatkay, 2008), SNPnexus (Chelala, et al., 2009) and SNPinfo (Xu and Taylor, 2009) webservers to functionalize SNPs. SNP functional information through SPOT web-server (Saccone, et al., 2010) is already obtained in the previous step while assigning SNPs to genes. Run “runfsnp.jar” to obtain SNP functional information from F-SNP webserver (Lee and Shatkay, 2008):

Replace \$DISEASE_NAME with the disease name that you specified in Step 3.

```
>java -jar runfsnp.jar $DISEASE_NAME
```

This step will save the F-SNP output into PANOGA_procedure/data/
\$DISEASE_NAME/ \$DISEASE_NAME_fsnp_output.txt.

17) Go to the SNPnexus webserver at: <http://www.snp-nexus.org/>. Under “Batch Query” option, Browse SNPnexus input file, i.e. \$DISEASE_NAME_snpnexus_input.txt.

18) Under “Annotation Categories”-> “Regulatory Elements”, select “Conserved Transcription Factor Binding Sites (TFBS)” option and click “Run” button.

19) Download the result under “Regulatory Elements” section via clicking into TXT icon. Save the SNPnexus output as text file under PANOGA_procedure/data/\$DISEASE_NAME/\$DISEASE_NAME_snpnexus_output.txt.

20) Go to the SNPinfo webserver at: <http://snpinfo.niehs.nih.gov/snpfunc.htm>. Browse and upload SNPinfo input file, i.e. \$DISEASE_NAME_snpinfo_input.txt.

21) Click “Submit” button and download the results via clicking into “Export To Excel” button under “SNP Function Prediction Results”. Save the SNPInfo output as csv file under PANOGA_procedure/data/\$DISEASE_NAME/\$DISEASE_NAME_snpinfo_output.csv.

3.2.2.2.6 Prepare the Gene Attributes data

22) PANOGA needs the attributes file (in .pvals format) to identify the sub-networks (using jActive Modules plugin (Ideker, et al., 2002)). This file has two weighted P-values (SPOT Pw and F-SNP Pw values) for each gene, where the weighted P-value combines the genotypic p-value of a SNP with the functional information of a SNP that is associated with the gene. The following steps of the PANOGA procedure will create an attributes file similar to the provided sample_panoga_spot_fsnp.pvals file at PANOGA_procedure/. Run “combinespotfsnp.jar” to combine SPOT and F-SNP output files:

Replace \$DISEASE_NAME with the disease name that you specified in Step 3.

```
>java -jar combinespotfsnp.jar $DISEASE_NAME
```

23) Run “incorporatesnpnexus.jar” to incorporate functional information from SNPnexus. Replace \$DISEASE_NAME with the disease name that you specified in Step 3.

```
>java -jar incorporatesnpnexus.jar $DISEASE_NAME
```

This run will create the gene attributes file (\$DISEASE_NAME_spot_fsnp_snpnexus.pvals) under PANOGA_procedure/data/\$DISEASE_NAME/. A sample .pvals file is shown in Figure 3.3.

GeneSymbol	SPOTPvalue	FSScorePvalue
PIK3C2A	0.002829698	0.02247709
SPATA18	2.54002e-5	4.97005e-4
DNAI2	0.001442935	0.014429346
EPB41L4A	0.00230284	0.018250034468831174

Figure 3.3 Sample gene attributes input file (sample_spot_fsnp_snpnexus.pvals), showing SPOT and F-SNP weighted p-values (Pw-values) for each SNP associated gene. Each of the two Pw values combines functional information of a SNP and the genotypic p-value of a SNP, that is found to be significant in GWAS.

3.2.2.2.7 Obtain network data

24) Decide which human protein-protein interaction (PPI) dataset you would like to use as your initial network—follow option A to use the default human PPI network or option B to use a customized PPI network.

(A) Using the default human PPI network

A user can work with the default human PPI network supplied in the PANOGA installation package. The default human PPI network is available in sif format in: PANOGA_protocol/data/humanPPI.sif.

(B) Using another PPI network

A user can work with their own human PPI network. Since Cytoscape (Shannon, et al., 2003) accepts networks in many different file formats (e.g., .sif, .gml, .xgmml, .xls, SBML, BioPAX, PSI-MI.), the user has the option to choose the network that they want to work with.

3.2.2.2.8 Load network data

25) Start Cytoscape via following option A for Windows users, option B for Linux users.

(A) Windows Users

Run Cytoscape.exe.

(B) Linux Users

Run ./cytoscape.sh.

26) Decide how you would like to load network data into Cytoscape. Cytoscape allows to import networks from a local or remote computer, or from Web Services—follow option A to import a network file from a local computer, option B from a remote computer or option C to use Web Services. We recommend PANOGA users to follow option A to load network data.

(A) Loading the default human PPI network from a local computer

(i) Assemble your network data into a SIF file, as described in Box 1.

(ii) Import human PPI network using File->Import->Network commands of Cytoscape. The user is free to load any human PPI network, as long as the official HUGO gene symbols are used as node identifiers. A sample human PPI network is also provided at: PANOGA_procedure/data/humanPPI.sif.

(B) Loading a PPI network from a remote computer

Follow the procedure described at:

http://wiki.cytoscape.org/Cytoscape_User_Manual/#Cytoscape_User_Manual.2BAC8Creating_Networks.Load_Networks_from_a_Remote_Computer_.28URL_import.29

(C) Loading a PPI network using Web Services

Follow the procedure described at:

http://wiki.cytoscape.org/Cytoscape_User_Manual/ImportingNetworksFromWebServices

3.2.2.2.9 Import gene attributes

27) Assign values (two attributes for each identified gene) to nodes (genes) using File->Import->Attribute/Expression Matrix commands of Cytoscape and selecting the gene attributes file (\$DISEASE_NAME_spot_fsnp_snpnexus.pvals) that is created in Step 23. A sample gene attributes file (sample_spot_fsnp_snpnexus.pvals) is also provided at PANOGA_procedure/data/sample/.

3.2.2.2.10 Identify sub-networks

- 28) Start jActiveModules plugin from Cytoscape->Plugins->jActiveModules.
- 29) Select SPOTPvaluesig and FSScorePvaluesig from Expression Attributes panel.
- 30) In the General Parameters panel, set “Number of Modules” parameter as 1000. “Overlap Threshold” parameter defines max percent of overlap between any two identified subnetworks. The default value used in PANOGA_protocol is 0.5.
- 31) Click “Find Modules” to identify active sub-networks.
- 32) Save the result as text file into PANOGA_procedure/data/\$DISEASE_NAME/\$DISEASE_NAME_jactivemodules_output.txt via clicking into “Save All Results” button on “Results Panel”. Replace \$DISEASE_NAME with the disease name that you specified in Step 3.

3.2.2.2.11 Parse jActiveModules output

- 33) Create a folder with your disease name under PANOGA_protocol/ClueGO/data/ and under PANOGA_protocol/ClueGO/out/ via typing the following commands:

```
>cd ClueGO/data
```

```
>mkdir $DISEASE_NAME
```

```
>cd ../out
```

```
>mkdir $DISEASE_NAME
```

```
>cd ../../
```

Replace \$DISEASE_NAME above with the disease name that you specified in Step 3.

- 34) Run “parsejactivemodulesoutput.jar” to create individual files containing gene symbols for each identified sub-network:

Replace \$DISEASE_NAME with the disease name that you specified in Step 3.

```
>java -jar parsejactivemodulesoutput.jar $DISEASE_NAME
```

At the end of this step, for the sub-networks with scores higher than 3, individual files containing gene symbols are saved under `PANOGA_procedure/ClueGO/data/ $DISEASE_NAME/` and the number of subnetworks created is printed on the screen.

3.2.2.2.12 Functional enrichment of subnetworks

35) Decide which pathway resource you would like to use for the functional enrichment of the identified subnetworks. ClueGO (Bindea, et al., 2009) assigns a set of genes into KEGG (Kanehisa, et al., 2012) or BioCarta pathways—follow option A to assign genes into KEGG pathways, option B to assign genes into Biocarta pathways.

(A) Identifying KEGG pathways

Use the `clueGO.props` file provided under `PANOGA_procedure/ClueGO/`. In order to identify KEGG pathways, make sure that under the “Select Ontologies” title “SelectedOntologySources=KEGG_14.03.2012” in the ClueGO properties file (`clueGO.props`).

(B) Identifying BioCarta pathways

In order to identify BioCarta pathways, under the “Select Ontologies” title change “SelectedOntologySources = REACTOME_BioCarta_07.04.2011” in the ClueGO properties file (`PANOGA_procedure/ClueGO/clueGO.props`).

36) At this step, the users need to choose one of the following two options, depending on their operating systems: Windows Users, follow option A; Linux Users, follow option B. For both options, replace `$DISEASE_NAME` with the disease name that you specified in Step 3, `$NUMBER_OF_SUBNETWORKS` with the number of subnetworks, as created in Step 34. Type the following command to perform functional enrichment for each of the identified subnetworks using the `clueGO.props` file created in Step 35:

(A) Windows Users

```
>java -jar functionalenrichment.jar $DISEASE_NAME  
$NUMBER_OF_SUBNETWORKS
```

(B) Linux Users

```
>./functionalenrichment.sh $DISEASE_NAME  
$NUMBER_OF_SUBNETWORKS  
($OPTIONAL_JAVA_PATH)
```

If java is installed as root user, skip \$OPTIONAL_JAVA_PATH and run as:

e.g. ./functionalenrichment.sh diabetes 508

If java is already installed on a different path, specify \$OPTIONAL_JAVA_PATH and run as:

e.g. ./functionalenrichment.sh diabetes 508 ../../jre1.7.0_04/bin

At the end of this step, enrichment results of each of the identified sub-networks are saved under
PANOGA_procedure/ClueGO/out/\$DISEASE_NAME/ for both options.

3.2.2.2.13 Combine functional enrichment results

37) Run the java script “combinesubnetworkpathways.jar” to create SNP targeted pathway lists and gene list for identified SNP targeted pathways. Replace \$DISEASE_NAME with the disease name that you specified in Step 3, \$NUMBER_OF_SUBNETWORKS with the number of subnetworks, as created in Step 34.

```
>java -jar combinesubnetworkpathways.jar $DISEASE_NAME  
$NUMBER_OF_SUBNETWORKS
```

At the end of this step pathway based lists and gene list are created as explained in the “Anticipated Results” section and “PANOGA’s Application to Human Complex Diseases” subsection of Introduction section.

3.2.2.2.14 Visualize SNP targeted genes in a KEGG pathway map

38) Create a directory under PANOGA_protocol/out/ to store gene attribute files for each pathway, via typing the following command:

```
>cd out/KeggPathwayMapGeneAttributeFiles
```

```
>mkdir $DISEASE_NAME
```

```
>cd ../../
```

Replace \$DISEASE_NAME above with the disease name that you specified in Step 3.

39) Run the java script “createattributesforpathwaymap.jar” to create a gene attributes file for each identified pathway, which will be used in the next step to customize KEGG pathway maps. Each pathway specific file contain identified gene symbols and color specifications depending on the number of SNP targeted genes per base pair. Replace \$DISEASE_NAME with the disease name that you specified in Step 3.

```
>java -jar createattributesforpathwaymap.jar $DISEASE_NAME
```

At the end of this step, gene attribute file for each of the identified sub-networks are saved under:

```
PANOGA_protocol/out/KeggPathwayMapGeneAttributeFiles/$DISEASE_NAME .
```

40) Color SNP targeted genes for the pathway of interest using the KEGG Mapper – Color Pathway tool available at: http://www.genome.jp/kegg/tool/map_pathway3.html.

41) Type “hsa” followed by the KEGG Term ID for the pathway of interest to the “Select KEGG pathway map:” field. KEGG Term IDs of the pathways can be

obtained from the first column of the \$DISEASE_NAME_pathways_subnetwork_genes.csv file under PANOGA_procedure/out/\$DISEASE_NAME/.

- 42) Browse gene attribute file created in Step 39 for the pathway of interest.
- 43) Hit “Execute” button. KEGG Mapper – Color Pathway tool (Kanehisa, et al., 2012) generates a customized pathway map, where the SNP targeted genes are colored based on the number of SNPs per base pair.

CHAPTER 4

4 RESULTS

4.1 Anticipated results of PANOGA protocol

Using PANOGA protocol, a GWAS can be further mined to identify SNP targeted pathways as associated with a specific human complex disease. These pathways can also be used as markers of a disease, which would have higher explanatory power than SNP or gene markers. The strength of our methodology stems from its multidimensional perspective, where we combine evidence from the following 5 resources: i) Genetic association information obtained through GWAS, ii) SNP functional information, iii) protein-protein interaction data, iv) LD, v) biochemical pathways. At the end of PANOGA protocol, pathway and gene tables with several features in .csv format (comma separated values) are generated, as shown in Tables 4.1-4.3 and 4.4, respectively. The files can be opened by Microsoft Excel or Open Office and displayed as spreadsheets. Each row of the pathway spreadsheet corresponds to the features of the identified pathway, i.e., KEGG term, KEGG term ID, p-value, rank, number of times found significant for different subnetworks, number of SNP targeted genes, number of typed SNPs in GWAS that are associated with the genes as part of the pathway under study, number of regulatory SNPs which are also found significant in GWAS, SNP targeted genes and their SNP counts.

Gene table file, as shown in Table 4.4, includes different features of the genes that are found as part of the identified pathways. While some of these genes are SNP targeted

genes, some others are identified as the neighbours of SNP targeted genes within the generated sub-networks. Each row of the gene spreadsheet correspond to a gene symbol, entrez gene ID, number of times found in subnetwork, number of associated pathways, list of associated pathways, number of typed SNPs in GWAS, functional information of the typed SNPs in GWAS, SNP regulatory potential, number of regulatory SNPs. If the number of typed SNPs in GWAS is zero, this means this gene is identified through neighbour effect. Tables 4.1-4.4 are described in more detail below, within the Results on rheumatoid arthritis dataset section.

Table 4.1 Pathway based representation of PANOGA results, focusing on SNP targeted genes. The top 5 SNP targeted KEGG pathways are shown along with their KEGG term IDs, ranks and p-values in the 1st, 3rd and 4th columns, respectively. SNP targeted genes that are identified in PANOGA protocol are shown in the 5th column; along with the number of typed SNPs, shown in paranthesis. For each identified SNP targeted pathway, number of SNP targeted genes, number of associated SNPs in GWAS, number of regulatory GWAS SNPs and how many times this pathway is identified are shown in columns 6 to 9, respectively.

KEGG ID	KEGG Term	Rank	Term Pvalue Corrected Bonferroni	SNP Targeted Genes (typed SNP counts)	# of SNP Targeted Genes	# of Associated SNPs in GWAS	# of Regulatory SNPs	Times Found
KEGG:04512	ECM-receptor interaction	1	8,76E-21	COL4A2(1); COL4A1(5); ITGA2(1); ITGB3(2); ITGA4(5); ITGB1(10); SDC2(1); COL5A1(2); SDC3(1); VWF(4); LAMA3(9); ITGA6(4); CD44(3); LAMA5(1); ITGB8(1); TNFR(3); ITGB6(8); FN1(3);	18	64	2	23
KEGG:04630	Jak-STAT signaling pathway	2	1,01E-19	PIK3CG (2); IL2RB (1); OSMR (1); STAM2 (2); SOCS1 (1); CBL (1); LIFR (4); STAT1 (5); STAT3 (4); IFNAR1 (1); IFNAR2 (2); CBLB (2); CSF3R (2); CSF2RB (12); JAK2 (1); IL5RA (1);	16	42	0	10
KEGG:04610	Complement and coagulation cascades	3	2,42E-19	PLAT (3); KNG1 (1); F11 (1); MBL2 (2); C3 (2); F13A1 (7); VWF (4); KLKB1 (1); SERPINC1 (4); PROS1 (1);	10	26	0	18
KEGG:04510	Focal adhesion	4	5,46E-19	BCAR1 (1); ITGB5 (5); ITGB3 (2); ITGB1 (10); ITGB8 (1); PAK4 (3); ITGB6 (8); PAK1 (4); FN1 (3); PRKCA (6); EGFR (8); FLT4 (1); ITGA2 (1); ITGA4 (5); PPP1CB (4); FLNB (1); VWF (4); VEGFC (3); ITGA9 (6); ITGA6 (4); FYN (10); GSK3B (1);	22	91	2	31
KEGG:04144	Endocytosis	5	1,14E-18	STAM2 (2); KIT (1); CLTC (7); IGF1R (1); CDC42 (1); AP2B1 (1); SH3GLB1 (3); WWP1 (2); NEDD4L (2); ITCH (3); SH3GL3 (5); EGFR (8); RET (1); FLT1 (1); CBL (1); HLA-C (8); CBLB (2); NTRK1 (1); SH3GL2 (4); EPN2 (1);	20	55	1	3

Table 4.2 Pathway based representation of PANOGA results, focusing on subnetwork genes. The top 5 SNP targeted KEGG pathways are shown along with their KEGG term IDs, ranks and p-values in the 1st, 3rd and 4th columns, respectively. Pathway associated genes that are found in the subnetworks are shown in the 5th column. While the genes without ‘*’ symbol are SNP targeted genes (e.g. JAK2 gene in the Jak-STAT signaling pathway), the genes with ‘*’ symbol are identified in the subnetwork due to the neighbour effect (e.g. JAK1 gene in the Jak-STAT signaling pathway). The genes with neighbour effect (not targeted by SNPs) are incorporated using PPI network in the subnetwork identification step of PANOGA and they help to identify SNP targeted pathways. Column 6 displays other members (genes) of the identified SNP targeted pathway, that are not found in PANOGA subnetworks.

KEGG ID	KEGG Term	Rank	Term Pvalue Corrected Bonferroni	Pathway Associated Genes Found in Subnetworks	Pathway Associated Genes Not Found in Subnetworks
KEGG:04512	ECM-receptor interaction	1	8,76E-21	COL4A2; COL4A1; ITGA2; ITGB3; ITGA4; ITGB1; SDC2; COL5A1; SDC3; VWF; LAMA3; CD36*; ITGA6; CD44; LAMA5; ITGB8; TNR; ITGB6; COL1A1*; FN1;	GP9; HMMR; HSPG2; IBSP; ITGA1; ITGA10; ITGA11; ITGA2B; ITGA3; ITGA5; ITGA7; ITGA8; ITGA9; ITGAV; ITGB4; ITGB5; ITGB7; LAMA1; LAMA2; LAMA4; LAMB1; LAMB2; LAMB3; LAMB4; ...
KEGG:04630	Jak-STAT signaling pathway	2	1,01E-19	PIK3CG; IL6*; IL2RB; OSMR; IL6ST*; STAM2; SOCS1; CBL; LIFR; IL6R*; STAT1; STAT3; IL11*; IFNAR1; TYK2*; OSM*; IFNAR2; CBLB; JAK1*; CSF3R; CSF2RB; JAK2; IL5RA;	CREBBP; CRLF2; CSF2; CSF2RA; CSF3; CSH1; CTF1; EP300; EPO; EPOR; GH1; GH2; GHR; GRB2; IFNA1; IFNA10; IFNA13; IFNA14; IFNA16; IFNA17; IFNA2; IFNA21; IFNA4; IFNA5; IFNA6; IFNA7; IFNA8; TPO; TSLP;
KEGG:04610	Complement and coagulation cascades	3	2,42E-19	PLAT; KNG1; F11; MBL2; F12*; F10*; C3; F13A1; PLG*; PROC*; VWF; FGG*; FGA*; FGB*; KLKB1; F2*; SERPINC1; PROS1; PLAU*;	A2M; BDKRB1; BDKRB2; C1QA; C1QB; C1QC; C1R; C1S; C2; C3AR1; C4A; C4B; C4BPA; C4BPB; C5; CD59; CFB; CFD; CFH; CFI; CPB2; CR1; CR2; F13B; SERPINA1; SERPINA5; SERPIND1; SERPINE1;
KEGG:04510	Focal adhesion	4	5,46E-19	TNC*; ERBB2*; BCAR1; ITGB5; ITGB3; ITGB1; PTK2*; ITGB8; PAK4; ITGAV*; ITGB6; PAK1; SPP1*; FN1; PRKCA; EGFR; FLT4; ITGA2; ITGA4; PPP1CB; FLNB; VWF; VEGFC; ITGA9; ITGA6; ITGA5*; FYN; GSK3B;	PIP5K1C; PPP1CA; PPP1CC; PPP1R12A; PRKCB; PRKCG; PTEN; PXN; RAC1; RAC2; RAC3; RAF1; RHOA; ROCK1; ROCK2; SHC1; SHC2; SHC3; SHC4; SOS1; SOS2; SRC; THBS1; THBS2; THBS3; THBS4; TLN1; TLN2; TNN; TNR; TNXB; VASP; VAV1; VAV2; VAV3; VCL; VEGFA; VEGFB; VTN; XIAP; ZYX;
KEGG:04144	Endocytosis	5	1,14E-18	STAM2; KIT; CLTC; IGF1R; CDC42; AP2B1; SH3GLB1; WWP1; NEDD4L; ITCH; SH3GL3; EGFR; RET; FLT1; CBL; HLA-C; EPS15*; CBLB; AP2A2*; NEDD4*; NTRK1; SH3GL2; EPN1*; DNMT1*; EPN2;	SMAD2; SMAD3; SMAD6; SMAD7; SMAP1; SMAP2; SMURF1; SMURF2; SNF8; SRC; STAM; STAMPB; TFRC; TGFB1; TGFB2; TGFB3; TGFB1; TGFB2; VPS45; VPS4A; VPS4B; VTA1; ZFYVE16; ZFYVE20; ZFYVE9;

Table 4.3. Pathway based representation of PANOGA results, focusing on associated SNPs from GWAS and their associated genes (SNP targeted genes). The top 5 SNP targeted KEGG pathways are shown along with their KEGG term IDs, ranks and p-values in the 1st, 3rd and 4th columns respectively.

KEGG ID	KEGG Term	Rank	Term Pvalue Corrected Bonferroni	Pathway Associated Genes Found in Subnetworks [Associated SNPs & Functional Properties]	# of Associated SNPs from GWAS
KEGG:04512	ECM-receptor interaction	1	8,76E-21	ITGB3 [rs2292700 ITGB3/intron; rs4629025 ITGB3/cds-synon]; ITGB1 [rs4587680 ITGB1/intron; rs9417094 ITGB1/UTR-3; rs11009021 ITGB1/UTR-3; rs16933501 ITGB1/intron; rs7914799 ITGB1/intron; rs7910994 ITGB1/UTR-3; rs2490486 ITGB1/UTR-3; rs11008969 ITGB1/UTR-3; rs10827054 ITGB1/UTR-3;ITGB1/intron; rs2230395 ITGB1/cds-synon]; SDC2 [rs10100191 SDC2/intron];.....	64
KEGG:04630	Jak-STAT signaling pathway	2	1,01E-19	IL2RB [rs3218253 IL2RB/intron]; JAK1 ; JAK2 [rs10491652 JAK2/intron]; IL5RA [rs2290611 IL5RA/intron]; IL2RA [rs2104286 IL2RA/intron; rs942200 IL2RA/intron; rs10795791 IL2RA/intron; rs11596355 IL2RA/intron; rs10905668 IL2RA/intron; rs10905669 IL2RA/intron; rs942201 IL2RA/intron; rs12722527 IL2RA/intron; rs12722489 IL2RA/intron; rs11256448 IL2RA/intron]; STAT1 [rs11687659 STAT1/intron; rs3024912 STAT1/nearGene-3; rs6718902 STAT1/nearGene-5;STAT1/intron; rs16833177 STAT1/nearGene-3; rs1914408 STAT1/nearGene-5;STAT1/intron]; STAT3 [rs3785898 STAT3/intron; rs744166 STAT3/intron; rs8069645 STAT3/intron; rs16967738 STAT3/intron]; IL11 ;.....	42
KEGG:04610	Complement and coagulation cascades	3	2,42E-19	KNG1 [rs698078 KNG1/missense;KNG1/intron]; F11 [rs4253417 F11/intron]; MBL2 [rs11003123 MBL2/nearGene-3; rs7095891 MBL2/nearGene-3]; FGG ; KLKB1 [rs925453 KLKB1/cds-synon];.....	26
KEGG:04510	Focal adhesion	4	5,46E-19	TNC ; ERBB2 ; BCAR1 [rs4887810 BCAR1/cds-synon]; ITGB5 [rs6438856 ITGB5/intron; rs4678169 ITGB5/intron; rs4678168 ITGB5/intron; rs4422355 ITGB5/intron; rs614664 ITGB5/intron]; ITGB3 [rs2292700 ITGB3/intron; rs4629025 ITGB3/cds-synon];	91
KEGG:04144	Endocytosis	5	1,14E-18	HLA-C [rs2524051 HLA-C/intron; rs4394275 HLA-C/nearGene-5; rs2524115 HLA-C/intron; rs3873385 HLA-C/UTR-3; rs10456057 HLA-C/nearGene-5; rs396038 HLA-C/intron; rs2853934 HLA-C/intron; rs2844615 HLA-C/intron]; EPS15 ;	55

Table 4.4 Gene list representation of PANOGA for the identified SNP targeted pathways. For each SNP targeted gene, number of associated SNPs in GWAS, number of regulatory GWAS SNPs, how many times this pathway is identified, number of associated SNP targeted pathways, list of these pathways, associated SNPs from GWAS, functional information regarding associated SNPs from GWAS, SNP regulatory potential and number of regulatory SNPs are shown in columns 2 to 9, respectively.

Gene Symbol	# of Associated SNPs from GWAS	Times Found in Subnetwork	# of Associated Pathways	Associated Pathways	Associated SNPs from GWAS	Functional info regarding Associated SNPs from GWAS	SNP Regulatory Potential	# of Regulatory SNPs
IL2RA	10	68	4	[HTLV-I infection; Endocytosis; Jak-STAT signaling pathway ; Measles]	[rs11256448; rs12722527; rs11596355; rs12722489; rs10795791; rs10905669; rs942200; rs942201; rs10905668; rs2104286]	[rs2104286 IL2RA/intron; rs942200 IL2RA/intron; rs10795791 IL2RA/intron; rs11596355 IL2RA/intron; rs10905668 IL2RA/intron; rs10905669 IL2RA/intron; rs942201 IL2RA/intron; rs12722527 IL2RA/intron; rs12722489 IL2RA/intron; rs11256448 IL2RA/intron]	rs11256448 0.243147; rs12722527 0.17259; rs11596355 0.115745; rs12722489 0.0; rs10795791 0.0; rs10905669 0.0; rs942200 0.12037; rs942201 NA; rs10905668 0.0; rs2104286 0.0;	0
IL2RB	1	141	4	[HTLV-I infection; Endocytosis; Jak-STAT signaling pathway ; Measles]	[rs3218253]	[rs3218253 IL2RB/intron]	rs3218253 0.113255;	0
JAK2	1	150	7	[Leishmaniasis; Cholinergic synapse; Measles; Jak-STAT signaling pathway ; Adipocytokine signaling pathway; Toxoplasmosis; Chemokine signaling pathway]	[rs10491652]	[rs10491652 JAK2/intron]	rs10491652 0.0;	0
STAT1	5	235	11	[Leishmaniasis; Osteoclast differentiation; Jak-STAT signaling pathway ; Measles; Influenza A; Pathways in cancer; Toll-like receptor signaling pathway; Pancreatic cancer; Hepatitis C; Toxoplasmosis; Chemokine signaling pathway]	[rs16833177; rs1914408; rs3024912; rs6718902; rs11687659]	[rs11687659 STAT1/intron; rs3024912 STAT1/nearGene-3; rs6718902 STAT1/nearGene-5; STAT1/intron; rs16833177 STAT1/nearGene-3; rs1914408 STAT1/nearGene-5; STAT1/intron]	rs16833177 0.0; rs1914408 0.176371; rs3024912 0.0; rs6718902 0.0; rs11687659 0.0;	0

In addition to the SNP targeted pathway list, and gene list representations of PANOGA, customized KEGG pathway maps, as shown in Figure 4.1, enrich the utility of PANOGA results. These pathway maps help the users to visualize affected genes along different routes within the pathway map. In these maps, the shade of red color in genes indicates the number of targeted SNPs (typed in the GWAS), per base pair of the gene. Figure 4.1 is described in more detail below, within the Results on rheumatoid arthritis dataset section.

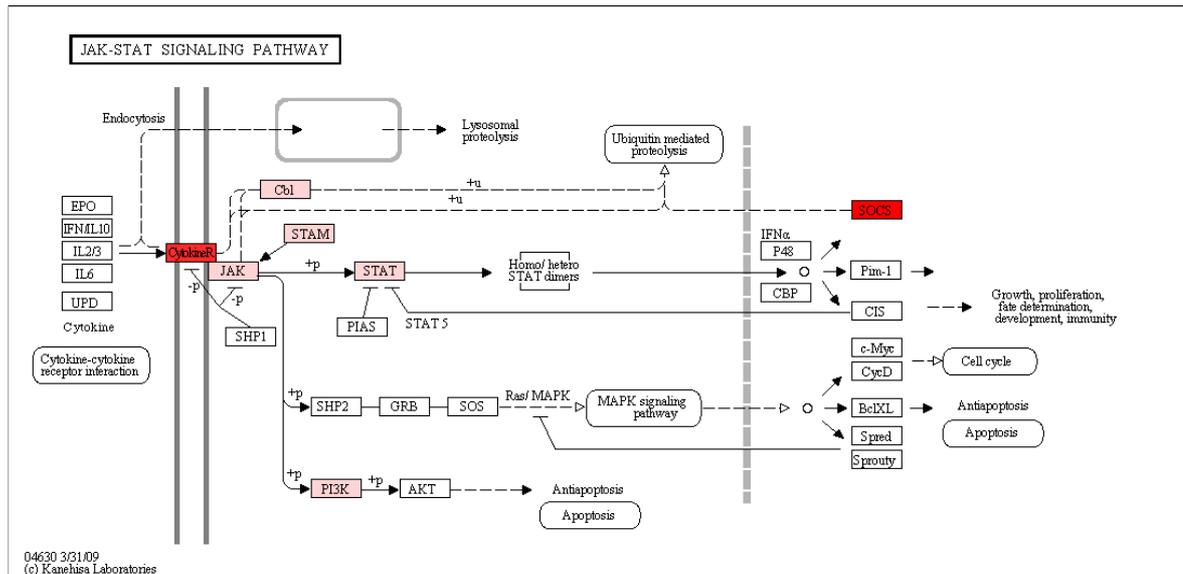


Figure 4.1 Customized KEGG pathway map for JAK-STAT signaling pathway. The shade of red color in genes indicates the number of targeted SNPs (typed in the GWAS of RA), per base pair of the gene. Red refers to the highest targeted gene, whereas white refers to a gene product, not targeted by the SNPs.

In the following sections, we present our findings on RA, PE, IA and Behçet’s disease datasets, using PANOGA protocol.

4.2 Results on rheumatoid arthritis dataset

Starting with 25,176 SNPs, that are found to be significant in a GWAS (WTCCC RA dataset), PANOGA was performed to identify RA related genes and functionally important KEGG pathways. These SNPs were assigned into 4,029 genes using SPOT

webservice (Saccone, et al., 2010) by considering all known SNP/gene transcript associations. As the possible overlap of a SNP with conserved TFBSs was considered, by using SNPnexus program (Chelala, et al., 2009), we incorporated 65 more proteins (TFs) that bind to the TFBS, that an RA associated SNP resides in. In order to incorporate functional information (functional score) to these genes, SPOT and F-SNP Pw-values were calculated as mentioned in the methods section. Following these calculations, network oriented steps of the PANOGA were realized using Cytoscape (Shannon, et al., 2003). SPOT and F-SNP Pw-values were used as attributes of the nodes (4094 genes) in the PPI network. We next searched for active sub-networks using the Cytoscape plugin jActive Modules (Ideker, et al., 2002). Once again, this plugin combines the network topology with attributes (SPOT and F-SNP Pw-values in our case) of each gene to extract potentially meaningful sub-networks. The higher the assigned aggregate z-score of a sub-network is, biologically more active the sub-network is. As in the original publication of jActive Modules (Ideker, et al., 2002), sub-networks with a score $S > 3$ (3 SD above the mean of randomized scores) were considered significant. Hence, our results with scores around 17.5 showed that this sub-network is statistically significant. But the involvement of the genes in this network with RA is further investigated through comparison with existing RA related information in databases.

4.2.1 Significant sub-networks for RA

Using both GWAS p-values and functional score, we identified 5 significant sub-networks on the basis of their aggregate degree of genetic association with RA. Due to the nature of the search algorithm, several of these sub-networks overlap extensively in their component genes. Thus, to describe a sub-network representative of association with RA, we selected the one with the highest score. This selected active sub-network is composed of 275 genes (our gene set) and 778 edges, as shown in Figure 4.2.a. Associations between 20 genes from this sub-network (XCL1, VCAM1, TRPV1, TRPC1, SPP1, RUNX1, RAC1, PRKCZ, NR3C1, NFKB1, MAP2K4, JUN, ITGB1, ITGAV, HMGB1, HLA-DMB, HLA-C, ERBB2, EPAS1, CCL21) and RA were verified by literature retrieved from the NCBI PubMed module and OMIM, as shown in Figure 4.2.b.

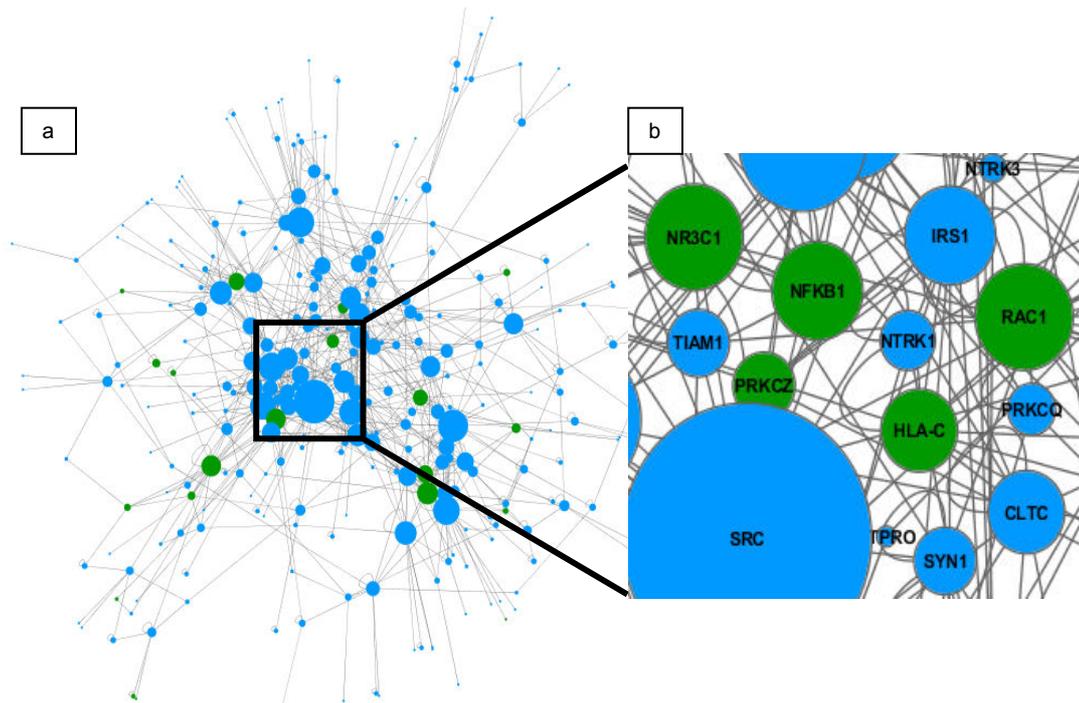


Figure 4.2 (a) The highest scoring sub-network is composed of 275 nodes and 778 edges (as found in Step 2 of PANOGA). Node size is shown as proportional to the degree of a node. **(b)** Zoomed in view of the highest scoring sub-network. 20 genes known in literature as associated with RA are shown in green. Blue denotes the genes in our highest scoring sub-network that cannot be associated with RA in literature.

In this highest scoring subnetwork, many of the nodes have modest p-values, and would not be seen as significant in a conventional GWAS analysis, as shown in Figure 4.3. For example, the Pw-value of HLA-DRA gene is $8.29E-48$, but its interacting partners MBP gene has Pw-value 0.016, thus, it is included in the subnetwork.

Next, we checked the topological parameters of this network. The distribution of the number of links per node (degree distribution, $P(k)$) is an important measure for a network to decide if it is a random, scale-free or hierarchical network. As shown in Figure 4.4.a, the degree distribution of our highest scoring sub-network follows a power-law distribution ($P(k)=ax^{-\gamma}$, $a=120.03$, $\gamma=1.353$, $R^2=0.773$, Correlation= 0.891 in log log scale) and hence it is scale-free, as expected from a biological network

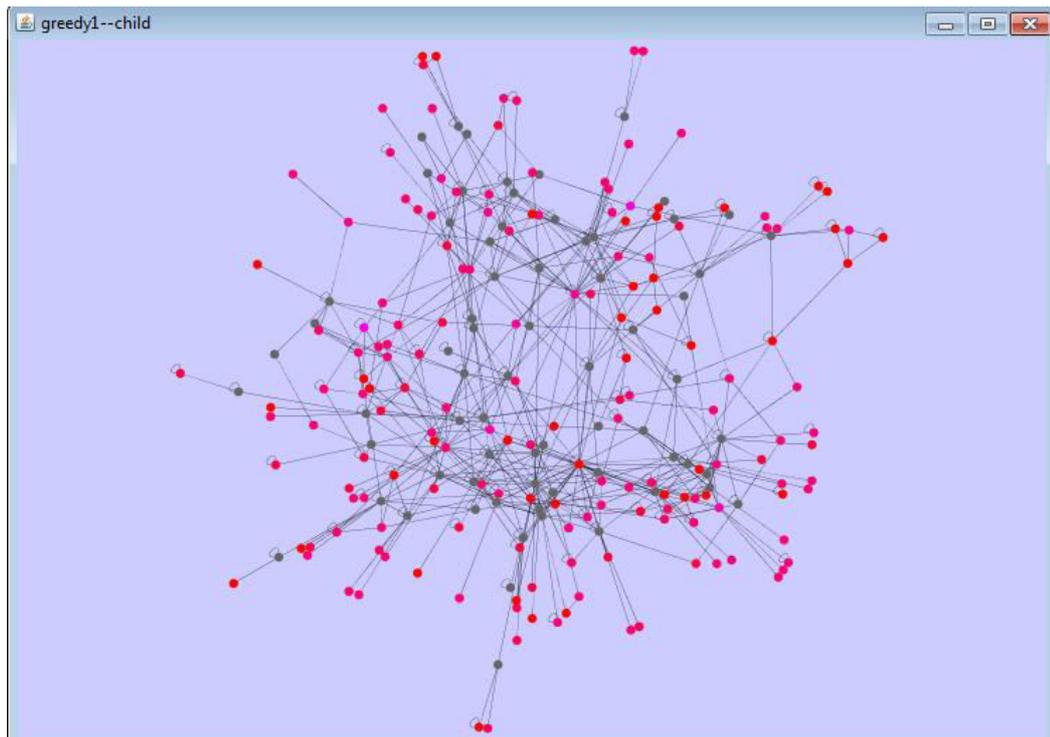


Figure 4.3 Highest scoring subnetwork, which is identified by jActiveModule using gene-wise weighted p-values, which combines GWAS p-values with the SNP's functional score. Node's color gradient from red to blue represents the Pw-values associated with each gene (from E-61 to 0.05), and the nodes with grey color indicate that these genes are not associated with a SNP, which is found to be significant in GWAS.

(Albert, 2005; Barabasi, 2009; Jeong, et al., 2000; Vallabhajosyula, et al., 2009). The unusual properties of scale-free networks are valid only for $\gamma < 3$ and the smaller the value of γ , the more important the role of the hubs is in the network (Barabasi and Oltvai, 2004). Similar to the degree distribution of the main PPI network ($\gamma=1.617$), the degree distribution of other top 5-scored sub-networks follows a power-law distribution ($\gamma=1.418, 1.365, 1.406, 1.330$). We also randomized our highest scoring sub-network using Erdos-Renyi algorithm and observed that its node degree distribution follows a Poisson distribution as expected from a random network (Figure 4.4.b).

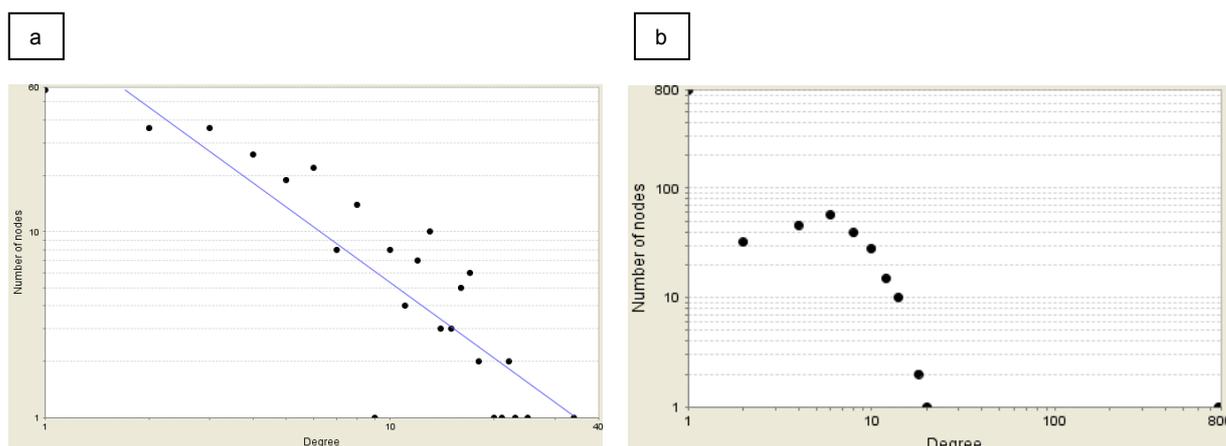


Figure 4.4 (a) Node degree distribution of the highest scoring sub-network follows a power-law ($P(k)=ax-\gamma$, $a= 120.03$, $\gamma=1.353$, $R^2=0.773$, Correlation= 0.891 in log log scale), showing that our network displays scale-free properties, as expected from a biological network. **(b)** Node degree distribution of a random network, obtained via randomization of our highest scoring sub-network using Erdos-Renyi algorithm.

4.2.2 Functionally important KEGG pathways for RA

As a result of the functional enrichment step (Step 3) of our methodology, we identified 87 KEGG pathway terms. In Table 4.5, we represent 20 most significant pathways (determined by their p-values), which are mostly related to immunity and inflammation, cell adhesion and cancers. Most of these pathways (Chemokine signaling, Neurotrophin signaling, Pathways in cancer, Leukocyte transendothelial migration, T cell receptor signaling, Toll-like receptor signaling, Allograft rejection, MAPK signaling, Apoptosis, Jak-STAT signaling) have been previously found to be associated with RA experimentally. In Table 4.5, we color coded the pathways and genes in blue, green and red, respectively, if they are computationally found only, experimentally found only, or found both experimentally and computationally. For example, Toll-like receptor (TLR) signaling pathway term was colored in red since other computational methods identified this term and it is also experimentally known to play an important role in the development and progress of RA. Among the most significant pathways identified by our methodology are Focal Adhesion and Cell Adhesion Molecules (CAM) pathways.

These pathways are experimentally shown to play a critical role in cellular processes such as osteoclast pathology and angiogenesis, which are known to be important for RA (Shahrara, et al., 2007).

We compared our findings with previously found RA related KEGG pathways and with the genes found from those pathways. Wu et al. (Wu, et al., 2010) created a comprehensive molecular interaction map for RA by combining the molecules and pathways found to be associated with RA based on merging all available papers related to high throughput experiments on RA. Following a procedure as in (Calzone, et al., 2008), they have decomposed their network into 11 modules using the Cytoscape plugin BiNoM (Zinovyev, et al., 2008). DAVID (Huang, et al., 2007) pathway analysis on their largest module with 292 nodes for 104 proteins and 334 edges returned 26 different KEGG pathways. In summary, this module contains 43 proteins from the MAPK signaling pathway, 36 proteins from focal adhesion, 23 proteins from the ErbB signaling pathway, and some cancer associated pathways such as leukemia, prostate cancer and colorectal cancer.

In another study (Martin, et al., 2010), the genomic regions showing low-significance associations in previous GWAS of RA (WTCCC and NARAC datasets) were further explored. Using Prioritizer software (Franke, et al., 2006), they have prioritised genes from similar pathways but located in different regions. This tool searches for those genes belonging to the same biological pathways or related biological pathways, based on the assumption that true disease-causing genes are functionally related. Prioritizer software uses a Bayesian approach to reconstruct a functional gene network based on known functional interactions from several databases such as the KEGG. Martin et al., 2010 reported 18 overrepresented KEGG pathways; in which Jak-STAT signaling pathway, Glioma, Calcium signaling pathway, Long-term potentiation, Apoptosis had the top 5 scores.

Table 4.5 Overrepresented KEGG Pathways found in the highest scoring sub-network for RA. Green denotes experimental, blue denotes computational, red denotes both experimental and computational verification regarding susceptibility to RA.

KEGG Term	Num. of Genes Found	Associated Genes (%)	Term Pvalue Corr. w/ Bonfer.	Associated Genes Found
Focal adhesion	30	14,9	9,33E-11	ACTB, ACTG1, AKT1, COL4A4, CRKL, CTNNB1, EGF, EGFR, FLNA, FLNB, FLT4, FYN, GRLF1, ITGA5, ITGB1 , ITGB3 , ITGB5 , MAP2K1, PAK4, PIK3R2, PTK2, RAC1 , RHOA, SHC3 , SRC, VASP, VAV1, VAV3, VTN, ZYX
ErbB signaling pathway	20	22,9	2,13E-10	AKT1, CAMK2D , CAMK2G , CBL, CRKL, EGF, EGFR, ERBB3, ERBB4, HBEGF, KRAS, MAP2K1, NCK2, NRG1, PAK4, PIK3R2, PTK2, SHC3 , SRC, STAT5A
Tight junction	22	16,4	1,80E-08	ACTG1, ACTN2, CASK, CTNNB1, EPB41L1, EPB41L2, EPB41L3, GNAI1, INADL, KRAS, LLGL1, MAGI1, MAGI3, PARD3, PRKCE, PRKCI, PRKCQ , PRKCZ , RHOA, SPTAN1, SRC, TJP1
Chemokine signaling pathway	26	13,7	2,31E-08	ADCY2, ADCY5, ADCY8 , AKT1, CHUK , CRKL, DOCK2, ELMO1, FGR, GNG2, IKBKB, KRAS, MAP2K1, NCF1, PARD3, PIK3R2, PRKCZ, PTK2, PTK2B, RAC1, RHOA, SHC3, STAT3, TIAM1, VAV1, VAV3
Adherens junction	17	22,6	1,16E-07	ACTB, BAIAP2, CREBBP, CTNNB1, EP300 , FYN, PARD3, PTPRF, PTPRM , RHOA, SMAD2, SMAD4, SORBS1, SRC, TCF7L2, TGFB1, TJP1
Bacterial invasion of epith. cells	15	20,5	1,57E-07	ACTB, ACTG1, CBL, CLTC, CTNNB1, CTTN, DNM3, ELMO1, ITGB1 , PIK3R1 , PTK2, RAC1 , RHOA, SRC, WASL
Neurotrophin signaling pathway	20	15,8	2,36E-07	ARHGDI1, CALM1, CALM3, CAMK2D , IKBKB, IRS1, JUN, KRAS, MAPK10, MAPK3, NFKB1, NTRK1, NTRK3, PLCG1, RAC1, RHOA, RPS6KA1, TP73, YWHAE, YWHAH
Long-term potentiation	15	21,4	3,67E-07	ADCY8 , CALM1, CALM3, CAMK2D , EP300 , GRIA1, GRIN1, GRIN2B, GRM5, ITPR1 , ITPR3 , KRAS, MAPK3, PPP1CB, RPS6KA1
Pathways in cancer	32	9,7	1,12E-06	CASP8 , CBL, CHUK , COL4A4, CTNNB1, EP300 , EPAS1 , ERBB2, FOXO1, FZD4, IKBKB, ITGAV, ITGB1 , JUN, KIT, KRAS, MAPK10, MAPK3, NFKB1, NTRK1, PIAS1, PIAS2, PLCG1, PTK2, RAC1, RHOA, RUNX1, SMAD4, STAT1, STAT5A, TPM3
Chronic myeloid leukemia	14	19,1	1,44E-06	CBL, CRK, CRKL, HRAS , IKBKB, MAPK3 , NFKB1 , PIK3R2, SHC1, SMAD3, SMAD4, SOS1, STAT5B, TGFB1
Cell adhesion molecules (CAMs)	18	13,2	1,42E-05	CD226, CD28 , CD4 , CDH2, HLA-B, HLA-C , HLA-DMB , HLA-DPA1 , HLA-DQA2 , HLA-DRA , ITGB1 , L1CAM, NCAM1, NLGN1 , PTPRC , PTPRF, PTPRM , SDC3
Leukocyte transendothelial migration	17	11	1,72E-05	ACTG1, ACTN2, CTNNB1, EZR, GNAI1, GRLF1, ITGB1 , NCF1, PLCG1, PTK2, PTK2B, RAC1 , RHOA, TXK , VAV1, VAV3, VCAM1
T cell receptor signaling pathway	16	14,8	2,70E-05	CBL, CD247 , CD28 , CD4 , CHUK , FYN, HRAS , IKBKB, LCK, MAP2K1, NCK2, PLCG1, PRKCQ , PTPRC , RHOA, VAV3
Toll-like receptor signaling pathway	13	12,7	1,97E-03	CASP8 , CHUK , IFNAR1 , IFNAR2, IKBKB, JUN, MAP2K4, MAPK10, MAPK3, NFKB1 , RAC1 , SPP1, STAT1
Antigen processing and presentation	11	13,9	2,08E-03	CALR, CANX, HLA-B, HLA-C , HLA-DMB , HLA-DRA , HLA-F, HLA-G , HSPA1L, TAP1 , TAP2
Allograft rejection	8	20	2,16E-03	CD28 , HLA-B, HLA-C , HLA-DMB , HLA-DPA1 , HLA-DQA2 , HLA-DRA , IL12A
MAPK signaling pathway	20	7,4	6,13E-03	CACNA1A, CHUK , CRKL, DAXX, EGF, FLNA, FLNB, FOS, HRAS , HSPA1L, JUN, MAPK10, MAPK3 , MAPK8, NF1, RAC1 , RPS6KA1, RRAS2, SOS1, TGFB1
Type I diabetes mellitus	8	17,3	6,24E-03	CD28 , HLA-B, HLA-C , HLA-DMB , HLA-DPA1 , HLA-DQA2 , HLA-DRA , IL12A
Apoptosis	11	12,5	6,84E-03	CAPN1, CASP10 , CASP8 , CHUK , CSF2RB , FADD, IKBKB, IRAK1, IRAK4 , PRKAR2A, PRKAR2B
Jak-STAT signaling pathway	15	9,6	7,41E-03	CBL, CREBBP, CSF2RB , EP300 , IFNAR1 , IFNAR2, IL12A, IL2RA , IL2RB , JAK1, LIFR, SOCS5, STAT1, STAT3, STAT5A

Baranzini et al., 2009 conducted a pathway-oriented analysis on WTCCC GWAS data for RA and another GWAS data by Plenge and collaborators. 9 KEGG pathways were identified in this study including Cell adhesion molecules (CAMs), Antigen processing and presentation, Type I diabetes mellitus. Lastly, the screening approach developed in (Zhang, et al., 2010) to further analyze GWAS data considers all SNPs with nominal evidence of Bayesian association, structural and functional similarities of corresponding genes. Responsible pathways identified in their study include Jak-STAT signaling pathways, cell adhesion molecules, and MAPK signaling pathways.

Comparative results with these four studies are shown in Table 4.6 in terms of number of genes found in commonly identified KEGG pathways. While most of these associations are computational predictions only, the functional relations of five of these pathways (Jak-STAT signalling, apoptosis, T cell receptor signalling, leukocyte transendothelial migration and cytokine-cytokine receptor interaction) with RA pathogenesis are known (Plenge, et al., 2007; Raychaudhuri, et al., 2009). Also, the effect of Toll-like receptor (TLR) signaling pathway and MAPK signaling pathway on RA is known. Here it is important to note that these associations are obtained by different methods on different datasets. For example, while Wu et al. utilizes text-mining (Wu, et al., 2010), Martin et al. mines GWAS data from WTCCC and NARAC studies (including variations on more cases and controls) (Martin, et al., 2010), and Zhang et al. applies their methodology on GAW16 (Genetic Analysis Workshop) data (Zhang, et al., 2010). PANOGA identifies previously found KEGG pathway terms with high statistical significance (terms shown in blue for former computational identification, in red for both computational and experimental identification).

From those previously identified pathways, we identified additional genes associated with RA within some of these pathways (e.g. Antigen processing and presentation, Tight junction). Importantly, within these pathways, the associations between some of these additionally found genes, such as HLA-C, HLA-G, PRKCQ, PRKCZ, TAP1, TAP2 (colored in green in Table 4.5) and RA were also verified by either OMIM database or by literature retrieved from the NCBI PubMed module.

Table 4.6 Comparison of found KEGG pathways with previous studies in terms of number of genes associated within each KEGG term for RA. Blue denotes computationally found pathways, green denotes experimentally verified RA associated pathways, and red denotes both experimental and computational verification.

KEGG Term	Number of Genes Found						Term Pvalue Corrected Bonfer-roni
	Baran zini et.al.	Martin et.al.	Wu et.al.	Zhang et.al.	PANOGA (only GWAS p-values)	PANOGA (w/2 attributes SPOT Pw and F-SNP Pw)	
Focal adhesion	0	0	36	32	22	30	9,33E-11
ErbB signaling pathway	0	0	23	0	18	20	2,13E-10
Tight junction	0	0	0	5	20	22	1,80E-008
Chemokine signaling pathway	0	0	0	0	24	26	2,31E-08
Adherens junction	0	0	0	18	16	17	1,16E-07
Bacterial invasion of epithelial cells	0	0	0	0	15	16	1,57E-007
Neurotrophin signaling pathway	0	0	0	0	20	20	2,36E-07
Long-term potentiation	0	22	0	7	14	15	3,67E-07
Pathways in cancer	0	0	0	0	29	32	1,12E-06
Chronic myeloid leukemia	4	0	21	18	10	14	1,44E-06
Cell adhesion molecules (CAMs)	8	26	0	10	12	18	1,42E-05
Leukocyte transendothelial migration	0	24	14	0	17	17	1,72E-05
T cell receptor signaling pathway	4	21	16	16	13	16	2,70E-05
Toll-like receptor signaling pathway	0	0	22	6	7	13	1,97E-03
Antigen processing and presentation	6	0	0	3	11	11	2,08E-03
Allograft rejection	0	0	0	0	8	8	2,16E-03
MAPK signaling pathway	0	0	43	34	16	20	6,13E-03
Type I diabetes mellitus	5	0	0	1	8	8	6,24E-03
Apoptosis	0	18	12	11	6	11	6,84E-03
Jak-STAT signaling pathway	0	25	0	16	13	15	7,41E-03
Prostate cancer	0	0	22	0	10	11	5,04E-02
Calcium signaling pathway	0	35	0	4	15	16	1,63E-01
VEGF signaling pathway	3	0	15	13	8	9	2,71E-01
Total	30	171	224	194	332	385	

Different from previous studies, we also identified Chemokine signaling, Neurotrophin signaling, Pathways in Cancer, Allograft rejection pathways as significant for RA. While the significance of these pathways in relation to RA were not thoroughly discussed in literature, the KEGG functional enrichment of RA-specific drug target genes, included these terms. List of drug target genes for RA, is downloaded from

Pharmacogenomics Knowledge Base website. In this database, 83 genes are associated with drugs that are used to treat RA. Furthermore, within these pathways, the associations between some of the genes, such as EPAS1, CD28, HLA-C (colored in green in Table 4.5) and RA were verified by either OMIM database or by literature retrieved from the NCBI PubMed module.

In order to assess the contribution of the found pathways and associated genes to disease mechanism, we also searched all identified genes from all found pathways in the Pharmacogenomics Knowledge Base website. When we filtered SNPs based on their significance in GWAS (p-value < 0.05 cutoff is applied) and assigned into genes, 14 out of 85 drug target genes were found. Whereas, via considering all the genes in the found KEGG pathways, we identified 25 out of 85 drug target genes, which are associated with RA. Hence, we showed that incorporating pathway knowledge on top of GWASs provides additional insights into the pathogenesis of RA.

To emphasize the effect of the functional score in PANOGA, we have applied our analysis on 4,094 genes firstly by using only GWAS p-values, secondly by using both SPOT and F-SNP Pw-values as attributes. As can be seen in Table 4.6, (PANOGA (w/ functional scores) column vs. PANOGA (only GWAS pvalues) column), incorporating functional information of a SNP increases the number of genes identified as associated with RA; and hence increases the significance of the identified KEGG pathway term.

4.2.3 Functionally grouped annotation network of RA

The diversity and complexity of the identified KEGG pathways involved in one sub-network confirms that RA is a complex systemic disease. Since a gene can be present in multiple pathways, we would like to show the pathway relationship, based on whether the pathways are sharing same genes. Hence, we generated a functional annotation network from the found KEGG pathways using ClueGO plugin (Bindea, et al., 2009). While the nodes in a functionally grouped network in Figure 4.5 denoted the found KEGG terms associated to RA, the edges were drawn based on the existence of shared genes using kappa statistics, in a similar way as described in (Huang, et al., 2007). 87

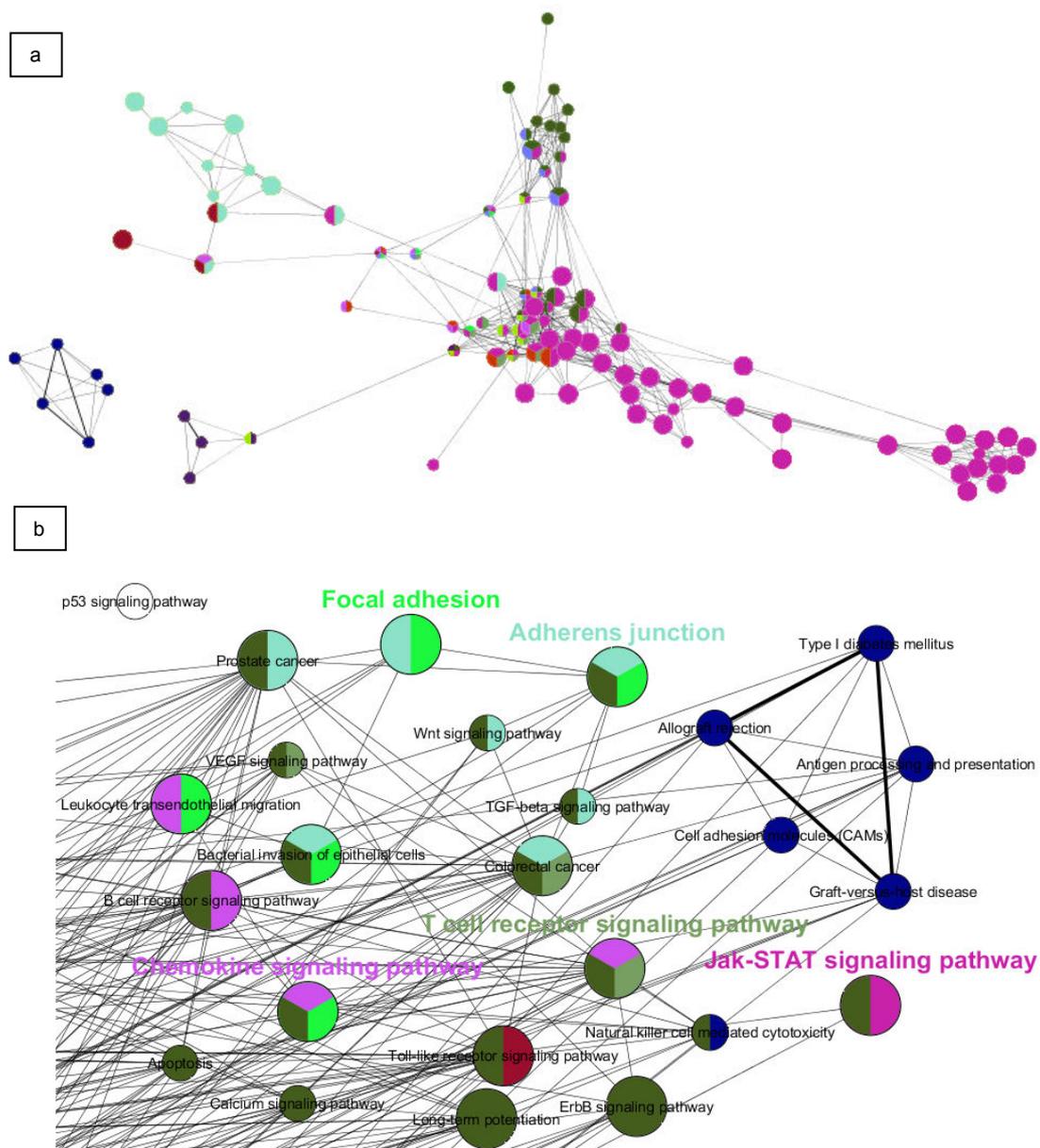


Figure 4.5 (a) Functionally grouped annotation network of our highest scoring sub-network. The relationships between the KEGG terms (nodes) were based on the similarity of their associated genes. The size of the nodes reflected the statistical significance of the terms (term p-values corrected with Bonferroni). Edges represent the existence of shared genes. The thickness of the edges is proportional to the number of genes shared and calculated using kappa statistics, in a similar way as described in (Huang, et al., 2007). The grouped terms (according to their kappa scores) were shown in same color. **(b)** Zoomed in view of the entire functional annotation network. The most significant pathway term of the group with the lowest term p-value (the group leading term) was shown in bold using the group specific color.

pathway terms that were found to be RA associated in our analysis were clustered into 9 groups, as can be seen in Figure 4.5.a (according to their kappa scores) and the pathways in the same group were shown in same color. ClueGO also assigns the most significant pathway terms with the lowest term p-value (corrected with Bonferroni) as group leading terms. For our functional annotation network, Focal adhesion, Adherens junction, Chemokine signaling pathways, T cell receptor signaling, Jak-STAT signaling were selected as group leading terms, as shown in Figure 4.5.b. Indeed, these group leading terms were either experimentally or computationally found to be related with RA, as can be seen in Table 4.5. This experiment generated the interconnections between the pathways that were found to be related with RA in our analysis.

To further check for the biological significance of our results, we compared the functional enrichments of the genes found in the highest scoring active sub-network with the functional enrichments of previously determined 331 genes verified by either OMIM database or by literature retrieved from the NCBI PubMed module to be associated with RA (Wu, et al., 2010). While our highest scoring sub-network with 275 genes enriched for 87 KEGG pathways, these 331 genes mapped to 88 pathways. Among those, 37 pathways were found in common, showing significant overlap between pathways coming from our study and the literature. In Figure 4.6.a, the different proportion of the genes found in KEGG pathways from two sets was represented with a color gradient from green for literature verified RA genes, to red for our gene set. White denoted the pathways found in both sets with equal number of genes. As shown in Figure 4.6.b (the zoomed in view), Pathways in cancer, T cell receptor signaling pathway, MAPK signaling pathway were found in both sets with the contribution of equal number of genes (shown in white). Whereas, the light green color in Neurotrophin signaling pathway term indicated that although most of the RA associated genes in this pathway comes from literature verified set, some of the genes in our gene set were assigned to this pathway.

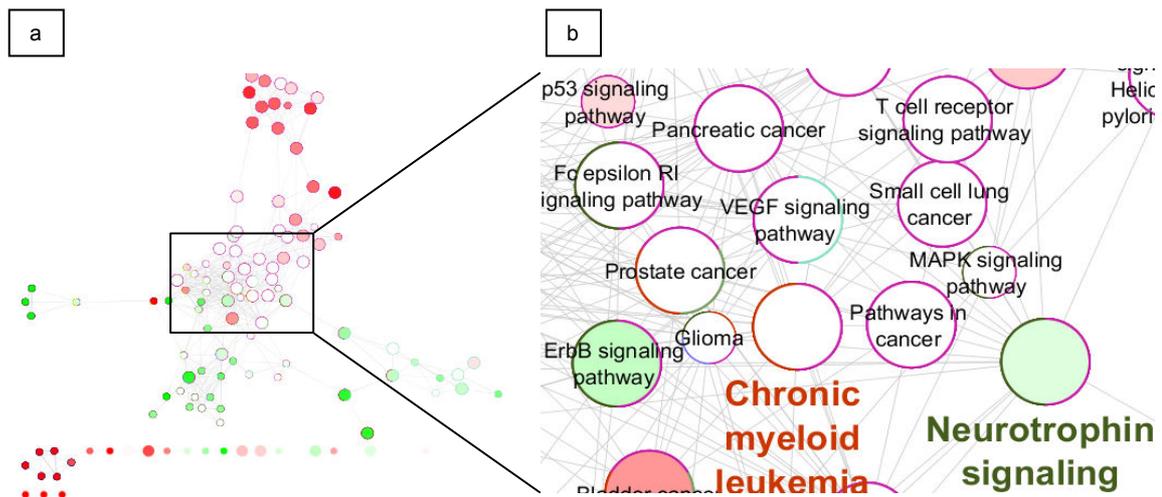


Figure 4.6 (a) Comparison of KEGG pathway terms with literature verified RA genes/our gene set were shown in green/red, respectively. Nodes represent the identified pathway terms from any one of the two sets. **(b)** Zoomed in view of the network. The color gradient showed the gene proportion of each set associated with the term. White color represented equal proportions from the two comparison sets. The size of the nodes reflected the statistical significance of the terms (term p-values corrected with Bonferroni). Following the convention in Figure 4.5.a, edges represented the existence of the shared genes between the pathway terms and node border colors mapped to the group colors.

4.2.4 Comparison with known drug target genes for RA

Since only a couple of KEGG pathways are known to be associated with RA in literature, for verification purposes we also compared the genes as part of these pathways with the drug target genes of RA in Pharmaccogenomics Knowledge Base. To this end, we tried to find out whether taking the genes in pathway context would enhance the results of GWA study by identifying additional target genes. As result of assigning SNPs coming from GWAS to genes we identified 4094 genes. Only 14 of them were mapped to 83 RA specific drug target genes. Following the application of our method, we identified KEGG pathways that are affected by the SNPs, and these pathways contained 25 out of 83 RA specific drug target genes. This provided an added value to GWAS analysis showing that not only the genes affected by the SNPs may be the drug targets but also other genes in these affected pathways may also be the drug

targets, as shown by 11 extra genes identified. The analysis of SNP affected genes in a pathway context provides added value in identification of potential drug targets.

4.2.5 Comparison with random networks

To test whether the identified KEGG pathways could be obtained by chance, we tested the enrichment in KEGG pathways for 100 randomly generated networks of size 275. The enrichment of these 100 random networks returned 68 different KEGG pathways. Among these 68 pathways, only two KEGG pathways (Type I diabetes mellitus and Allograft rejection) overlap with the identified pathways, as shown in Table 4.5. However, the statistical significance of these pathways (out of the random network) were low (term p-values=0.013 and 0.007 respectively). These two pathways are found only for one random network out of 100 randomly generated networks and both pathways are found due to the existence of the following 5 random genes in this network, i.e. PRF1, HLA-B, FAS, HLA-DQA1, IL2. Whereas in our pathway analysis (as shown in Table 4.5), more genes are identified as part of Type I diabetes mellitus and Allograft rejection pathways (i.e. CD28, HLA-B, HLA-C, HLA-DMB, HLA-DPA1, HLA-DQA2, HLA-DRA, IL12A). Hence, our gene list includes different genes compared to the ones found in random network with higher significance (term p-values=6.24E-03 and 2.16E-03 respectively).

4.2.6 KEGG pathway map of JAK-STAT signaling, as related to RA

Results explained so far on RA dataset focuses on the highest scoring sub-network. We have also applied the full PANOGA protocol on the RA dataset. Table 4.1-4.4 and Figure 4.1 summarize our results on the RA dataset, once the full PANOGA protocol is applied. As shown in Table 4.1-4.3, the JAK-STAT signaling pathway is identified by PANOGA on WTCCC rheumatoid arthritis (RA) GWAS dataset in 2nd ranking with $p=1.007E-19$. The effect of JAK-STAT signaling pathways on RA is reviewed in (Plenge, et al., 2007; Raychaudhuri, et al., 2008; Raychaudhuri, et al., 2009). SNP targeted pathway list of PANOGA (Table 4.1) indicates that 16 SNP targeted genes and 42 genotyped SNPs are found in this pathway. For each SNP targeted pathway, PANOGA protocol displays not only SNP targeted genes but also neighbour genes that

are found in subnetworks. For instance, while JAK2 gene is a SNP targeted gene, JAK1 is a neighbour gene (shown with * in Table 4.2) and found in the identified subnetworks. These neighbour genes help to identify SNP targeted pathways, which can not be picked up using SNP targeted genes only. Different from Tables 4.1 and 4.2, Table 4.3 displays associated SNPs from GWAS, their associated genes (SNP targeted genes) and the functional effect of the SNP on the gene. Additionally, customized KEGG pathway map representation of PANOGA (Figure 4.1) demonstrates that CytokineR, which includes IL2RA and IL2RB genes (shown in red) creates a complex with JAK (shown in pink). The dramatic effects of JAK inhibitors on RA in clinical trials are recently discussed in (Migita, et al., 2011). As shown in Figure 3, JAK can phosphorylate (i) STAT (shown in pink) and may lead to immunity related responses; (ii) SHP2 and as a downstream effect, MAPK signaling pathway, which is known to cause chronic synovitis during RA (Schett, et al., 2000), is influenced. MAPK signaling pathway is also identified by PANOGA protocol with $p=4.6E-17$, as shown in Table 4.5. On the other hand, Table 4.4, the gene list representation of PANOGA protocol shows that rs6718902 and rs1914408 are found to be associated with RA ($p<0.05$) according to WTCCC GWAS on RA. PANOGA protocol identifies these SNPs on the 5' end of STAT1 gene (shown in pink in Figure 4.1), which is part of the JAK-STAT signaling pathway.

4.3 Results on partial epilepsy dataset

Among the 528,745 SNPs, which were tested in the original GWAS, 28,450 SNPs showed nominal evidence of association ($P < 0.05$). These affected SNPs mapped to 4347 genes. At the end of the subnetwork identification step, we identified 545 significant sub-networks. Following the identification of sub-networks, we evaluated whether these sub-networks were biologically meaningful. In another words, for each sub-network, we searched for the over-represented pathways. 47 pathways were identified with p-values less than $E-10$. The top 30 over-represented pathways were shown in Table 4.7. Among the 545 subnetworks, these pathways were identified at least for 2 subnetworks and at most for 241 subnetworks. The higher the number of times that the pathway is found significant for a sub-network, the more support that the pathway gets from different parts of the PPI network and hence, such pathways could be more important for disease development mechanisms. Complement and coagulation

Table 4.7 The top 30 over-represented KEGG pathways identified for PE dataset.

KEGG Term	Term Pvalues Corr Bonf	Times Found	# of Associated SNPs in GWAS	# of Regulatory SNPs	# of SNP Targeted Genes
Complement and coagulation cascades	2,16E-025	24	34	1	12
Cell cycle	1,03E-024	27	24	0	14
Focal adhesion	7,10E-023	71	97	3	20
ECM-receptor interaction	1,62E-022	71	62	2	14
Jak-STAT signaling pathway	1,16E-021	3	24	1	16
MAPK signaling pathway	2,32E-019	10	73	2	23
Proteasome	1,15E-018	7	11	1	4
Ribosome	1,57E-018	6	2	0	2
Calcium signaling pathway	5,73E-018	26	154	2	22
Regulation of actin cytoskeleton	9,23E-018	10	88	4	19
Adherens junction	1,01E-017	241	79	4	13
Pathways in cancer	3,94E-017	11	112	5	22
Gap junction	6,32E-017	135	147	3	18
Apoptosis	3,72E-016	21	37	0	13
Long-term depression	2,90E-015	182	151	2	15
Axon guidance	4,01E-015	32	59	1	12
Fc gamma R-mediated phagocytosis	2,22E-014	105	66	3	12
Tight junction	2,82E-014	16	82	2	13
ErbB signaling pathway	4,04E-014	158	86	1	12
Wnt signaling pathway	6,28E-014	17	44	1	13
Chemokine signaling pathway	9,60E-014	18	68	0	19
GnRH signaling pathway	1,22E-013	67	65	2	15
Pentose phosphate pathway	1,29E-013	17	20	1	7
Long-term potentiation	2,28E-013	140	94	1	13
Neurotrophin signaling pathway	3,24E-013	27	19	0	9
Glycolysis/Gluconeogenesis	4,29E-013	3	21	1	8
Notch signaling pathway	9,33E-013	33	5	0	5
Dilated cardiomyopathy	1,40E-012	32	109	1	11
TGF-beta signaling pathway	2,32E-012	23	15	0	7
Endocytosis	3,61E-012	2	72	0	12

cascade pathway, which was identified as the most important pathway in our analysis ($p=2,16E-25$), was discussed previously as associated with epileptogenesis, in a study that conducted transcriptome analysis of the hippocampal cells from rats subjected to the pilocarpine model of epilepsy (Okamoto, et al., 2010). They showed that seven genes from this pathway were commonly up-regulated throughout epileptogenesis, from the early events post-status epilepticus to the onset of recurrent spontaneous seizures (Okamoto, et al., 2010). In our analysis, as part of this pathway, we identified 12 genes,

which were targeted by 34 genotyped SNPs, and one of these SNPs had a regulatory role. In the two previous GWAS studies of epilepsy (Guo, et al., 2012; Kasperaviciute, et al., 2010), this pathway is not pronounced as important. This result shows one more time that the GWAS are undermined in most cases and more comprehensive analytical approaches, as presented here, are needed.

In order to compare our findings more systematically with previous studies, we compiled seven lists of epilepsy associated genes, checked their pathway enrichments using DAVID tool (Huang, et al., 2007), and compared the identified pathways with our top 30 SNP-targeted pathway list. First gene list was obtained through Wang et al's study, in which they comparatively evaluated the status of gene expression profiling in epileptogenesis, as of 2010 (Wang, et al., 2010). We got the list of 53 genes, which was reported to be differentially expressed in more than one study among 18 studies. Second gene list, including 185 genes, was obtained through OMIM, as a result of our search using "partial epilepsy" keyword. Third gene list included 6 genes, shown as the closest genes to the 11 SNPs, which were found to be significant in a GWAS on PE (Kasperaviciute, et al., 2010). Fourth gene list was obtained through a copy number variation (CNV) study in idiopathic generalized and focal epilepsies, in which they identified CNVs in 26 genes (Mefford, et al., 2010). Fifth gene list included 81 epilepsy related genes that was provided by our collaborators. Sixth gene list assembled genes listed in the Epilepsy Genetic Association Database (epiGAD), which summarizes the results of both published and unpublished research efforts relating to genetic association studies in the field of epilepsy. Seventh, the last gene list was collected from Rogic et al's study (Rogic and Pavlidis, 2009). Via reviewing four articles (Lukasiuk, et al., 2006; Lukasiuk and Pitkanen, 2004; Wang, et al., 2010; Zagulska-Szymczak, et al., 2001), they compiled a list of genes that have been previously linked to epileptogenesis or excitotoxic-brain injury by studies of gene expression (Rogic and Pavlidis, 2009). This list of 182 differentially expressed genes during epileptogenesis was used as our last gene list. Even though we collected candidate gene lists of epilepsy, that are obtained in different studies via different whole-genome studies, e.g. gene expression, GWAS and whole-genome oligonucleotide array comparative genomic hybridization, we suspect that the affected pathways might show commonalities. As shown in Table 4.8, all pathways in our list, except proteasome, pentose phosphate and notch signaling pathways were also found for at least one of the seven candidate gene lists. Wang et al

showed that only 53 out of 2000 differentially expressed genes were found in more than one study (Wang, et al., 2010). Whereas in our study, pathways targeted by candidate epilepsy genes, which are coming from different studies show more conservation. 20 out of the top 30 pathways were found to be common for at least three different studies, as shown in Table 4.8. This result also indicated the relevance of our pathway oriented approach. Compared to the individual genes/SNPs, we showed that the pathways were more conserved among different epilepsy studies.

The diversity of the identified KEGG pathways confirmed that PE was a complex disease. Since a gene can be present in multiple pathways, we investigated the relationships among the pathways, based on the shared genes. To this end, we grouped the top 30 SNP targeted pathways of PE with the help of ClueGO plugin (Bindea, et al., 2009), as shown in Figure 4.7. While the nodes in this figure denoted the SNP targeted pathways, the edges were drawn based on the existence of shared genes using kappa statistics, in a similar way as described in (Huang, et al., 2007). The pathway terms, which were shown in color, indicated that this term was a group leading term (GLT). GLT was chosen as the pathway with the smallest p-value among all the pathways in the same group. While 8 pathways did not belong to any groups of pathways, other pathways were grouped into 7 major groups. As expected, signaling pathways were interlinked with synapse pathways. Jimenez-Mateos et al performed a microarray analysis in mice, 24 h after status epilepticus, where the mice had received previously either seizure preconditioning (tolerance) or sham-preconditioning (injury) (Jimenez-Mateos, et al., 2008). Analysis of the genes differentially suppressed in tolerance identified calcium signaling and synapse pathways as over-represented (Jimenez-Mateos, et al., 2008). As shown in Figure 4.7, the functional grouping of our top 30 pathways clustered calcium signaling with Cholinergic, GABAergic, Glutamatergic Serotonergic, Dopaminergic synapse pathways. Our functional grouping placed long term potentiation and long term depression pathways among signaling pathways. Since these pathways shared genes with signaling pathways, a similar situation was observed in Jimenez-Mateos's study, such that long term potentiation pathway comprised 8 downregulated genes, including signaling intermediates (calcium-dependent phosphatase Ppp3r1) and nuclear/transcription-associated genes (calcium/calmodulin-dependent kinase 4). As Figure 4.7 illustrates, GnRH signaling pathway was found as the group leading term of the signaling pathways that were identified in our top 30

Table 4.8 Comparison of the top 30 SNP-targeted KEGG pathways with the pathways of the known genes as associated with PE.

KEGG Term	Term Pvalue Corr Bonf	Wang et al. Study	OMIM	GWAS on PE	CNV Study on Epilepsy	Candidate Gene List	EpiGAD	Rogic et al. Study
Complement and coagulation cascades	2,16E-025	-	Y	-	-	-	-	Y
Cell cycle	1,03E-024	-	Y	-	-	-	-	Y
Focal adhesion	7,10E-023	Y	Y	Y	-	-	-	Y
ECM-receptor interaction	1,62E-022	Y	Y	-	-	-	-	Y
Jak-STAT signaling pathway	1,16E-021	Y	Y	-	-	-	-	Y
MAPK signaling pathway	2,32E-019	Y	Y	Y	-	Y	Y	Y
Proteasome	1,15E-018	-	-	-	-	-	-	-
Ribosome	1,57E-018	-	-	-	-	-	-	Y
Calcium signaling pathway	5,73E-018	Y	Y	Y	Y	Y	Y	Y
Regulation of actin cytoskeleton	9,23E-018	Y	Y	-	Y	-	-	Y
Adherens junction	1,01E-017	-	-	Y	-	-	-	Y
Pathways in cancer	3,94E-017	Y	Y	Y	-	-	-	Y
Gap junction	6,32E-017	Y	Y	Y	-	-	-	Y
Apoptosis	3,72E-016	Y	Y	-	-	-	-	Y
Long-term depression	2,90E-015	Y	Y	Y	Y	Y	Y	Y
Axon guidance	4,01E-015	-	-	-	-	-	-	Y
Fc gamma R-mediated phagocytosis	2,22E-014	Y	Y	Y	Y	-	-	Y
Tight junction	2,82E-014	Y	Y	Y	-	-	-	Y
ErbB signaling pathway	4,04E-014	Y	Y	Y	-	-	-	Y
Wnt signaling pathway	6,28E-014	Y	Y	Y	-	Y	-	Y
Chemokine signaling pathway	9,60E-014	Y	-	Y	Y	-	-	Y
GnRH signaling pathway	1,22E-013	Y	Y	Y	-	-	-	Y
Pentose phosphate pathway	1,29E-013	-	-	-	-	-	-	-
Long-term potentiation	2,28E-013	Y	Y	Y	-	Y	-	Y
Neurotrophin signaling pathway	3,24E-013	Y	Y	-	-	-	-	Y
Glycolysis / Gluconeogenesis	4,29E-013	Y	Y	-	-	-	-	Y
Notch signaling pathway	9,33E-013	-	-	-	-	-	-	-
Dilated cardiomyopathy	1,40E-012	-	Y	Y	-	Y	-	Y
TGF-beta signaling pathway	2,32E-012	-	-	-	-	-	-	Y

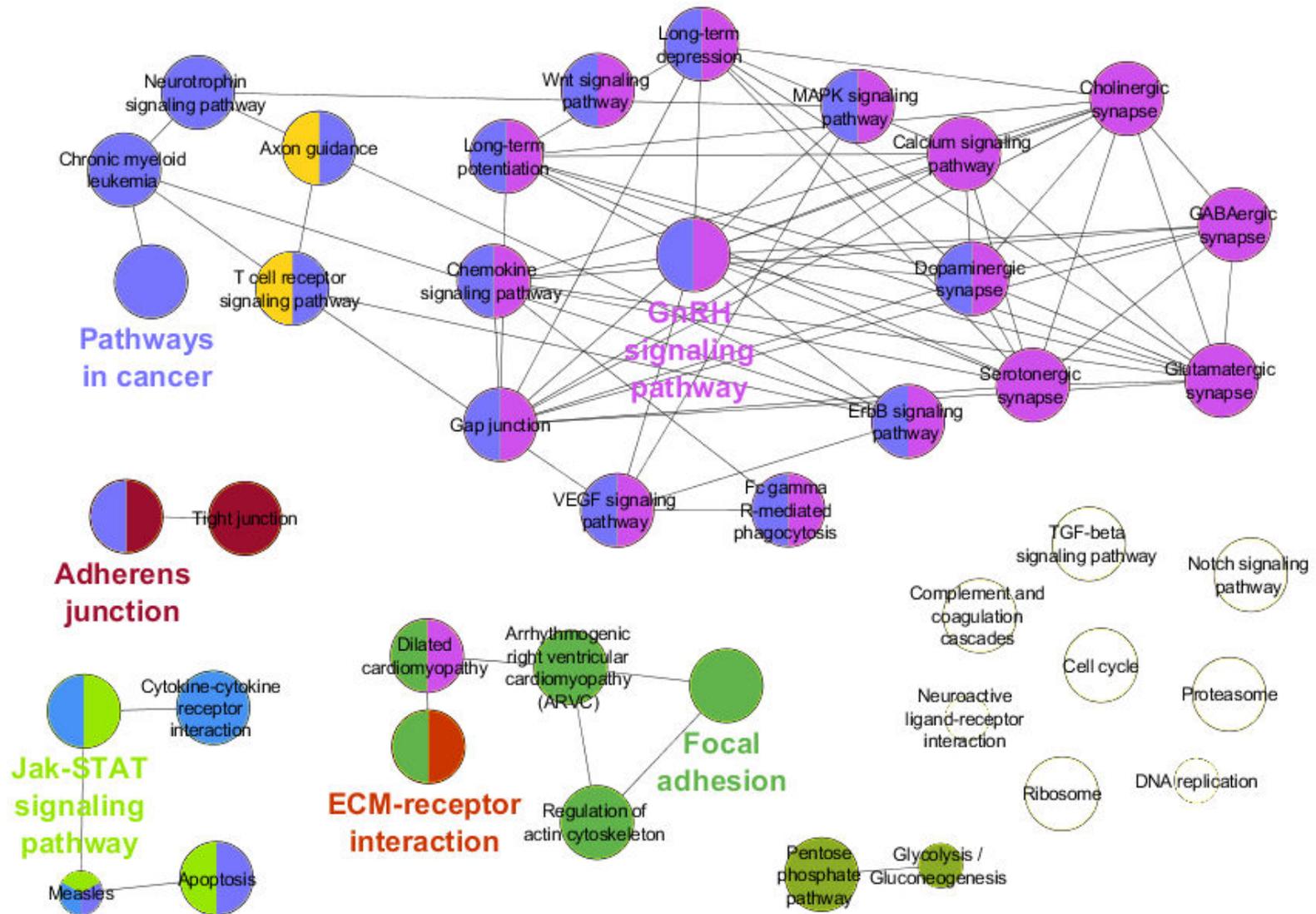


Figure 4.7 Functionally grouped annotation network of the identified pathways for epilepsy dataset. The pathways are grouped based on the similarity of their SNP targeted genes.

pathway list. This pathway was also reported by Lauren et al., in their transcriptome analysis of the hippocampal CA1 subregion after kainic acid-induced status epilepticus in 21-day-old rats, which were developmentally comparable to juvenile children (Lauren, et al., 2010).

4.4 Results on intracranial aneurysm dataset

We applied PANOGA on two IA GWAS separately: i) Finnish, Dutch (European, EU) population of 1701 cases and 7409 control cohorts (Bilguvar, et al., 2008; Yasuno, et al., 2010), ii) Japanese (JP) population of 1069 cases and 904 controls (Akiyama, et al., 2010). In our analysis, we have included 44,351 SNPs from EU population specific dataset, and 14,034 SNPs from the JP population specific dataset with p-values < 0.05, where the genotypic p-value of a SNP is calculated via Cochran-Armitage trend test. Only 576 of these SNPs were common between two populations. While the affected SNPs from EU population map to 3327 genes, the affected SNPs from JP population map to 2804 genes. As the possible overlap of a SNP with conserved TFBSs was considered, by using SNPnexus program (Chelala, et al., 2009), we incorporated 169 (EU dataset) and 126 (JP dataset) more proteins (TFs) that bind to the TFBS, that an IA associated SNP resides in. 1125 of these SNP targeted genes are commonly found in both EU and JP datasets. To identify the biological pathways with the genes responsible for IA susceptibility, we applied the affected SNP functionalization, SNP to gene mapping, gene-wise weighted p-value calculation, sub-network identification and functional enrichment steps of PANOGA separately for each dataset. The details of these steps are explained in the methods section.

We calculated the rankings of each identified pathway in each population and found that the correlation between the two studies was significant (Spearman's $r_2=0.71$, $P<10^{-6}$). Pairwise correlation of pathway statistics between two studies (which were carried out on independent populations with different ethnicities) should indicate common genetic variation associated with IA. As shown in Table 4.9, 12 of the top 20 ($P=4.09E^{-60}$) and 7 of the top 10 ($P=2.44E^{-36}$) affected pathways were found to be commonly identified in both EU and JP populations. In these 12 commonly identified pathways, while 95 and 81 genes are uniquely targeted by disease predisposing SNPs in EU and JP populations

Table 4.9 The top 20 KEGG pathways identified for both populations in IA. 7 out of the top 10 and 12 out of top 20 pathways are shown in bold and in italic, respectively.

KEGG Term	P-values		Rank		# of Associated SNPs in GWAS		# of Common SNPs in GWAS	# of SNP Targeted Genes (STGs)		# of Common STGs	% Common Genes in Both Populations		Common SNPs in GWAS
	EU	JP	EU	JP	EU	JP		EU	JP		EU	JP	
MAPK signaling pathway	3.53E-27	2.70E-18	1	8	133	43	1	14	18	2	14.29	11.11	rs791062
Cell cycle	2.35E-25	2.81E-19	2	4	76	18	1	11	10	2	18.18	20	rs744910
TGF-beta signaling pathway	6.26E-24	2.41E-17	3	9	126	20	3	15	9	5	33.33	55.56	rs2053423. rs1440375. rs744910
<i>ErbB signaling pathway</i>	9.52E-22	2.47E-15	4	16	50	15	0	6	4	0	0	0	
Focal adhesion	9.55E-22	5.60E-21	5	2	117	45	1	21	14	5	23.81	35.71	rs4678167
Proteasome	2.36E-21	4.55E-11	6	35	32	1	0	6	1	0	0	0	
Adherens junction	4.91E-19	2.58E-21	7	1	85	34	1	13	11	2	15.38	18.18	rs1561798
Notch signaling pathway	2.14E-18	4.74E-12	8	31	26	13	0	8	4	1	12.5	25	
Regulation of actin cytoskeleton	2.28E-18	4.05E-17	9	10	102	36	1	18	14	1	5.556	7.143	rs4678167
Neurotrophin signaling pathway	2.49E-18	1.93E-18	10	7	68	14	0	7	7	1	14.29	14.29	
Chronic myeloid leukemia	2.62E-18	8.13E-11	11	36	54	12	1	4	4	1	25	25	rs744910
Apoptosis	7.37E-18	1.71E-8	12	58	17	10	0	8	6	1	12.5	16.67	
<i>Pathways in cancer</i>	1.16E-17	9.38E-19	13	6	147	48	1	16	19	2	12.5	10.53	rs744910
Tight junction	1.84E-17	4.68E-14	14	21	98	37	4	14	11	6	42.86	54.55	rs4578183. rs4654383. rs955749. rs2276266
Long-term potentiation	2.25E-17	2.21E-13	15	24	140	23	0	13	10	3	23.08	30	
Measles	1.06E-16	3.42E-7	16	72	94	16	0	8	6	0	0	0	
<i>T cell receptor signaling pathway</i>	1.62E-16	1.97E-15	17	15	63	28	0	7	13	0	0	0	
<i>Nucleotide excision repair</i>	3.66E-16	8.84E-15	18	18	70	12	0	6	7	1	16.67	14.29	
Chemokine signaling pathway	1.15E-15	8.17E-12	19	32	193	26	0	13	13	2	15.38	15.38	
<i>Calcium signaling pathway</i>	3.27E-15	1.37E-15	20	12	123	42	1	21	16	8	38.1	50	rs7298821

Table 4.10 The top 20 over-represented KEGG pathways for IA, and the SNP targeted genes within these pathways. Seven out of the top ten affected pathways in both EU and JP populations are shown in *italics*. SNP Targeted Genes that are identified in both EU and JP populations are shown in the last column; along with the number of commonly typed SNPs in both populations, only in EU population and only in JP populations are shown in paranthesis.

KEGG Term	Any Common SNPs in Common Genes?	Commonly Associated. SNP Targeted Genes and their SNP Counts: (Common) (EU GWAS) (JP GWAS)
<i>MAPK signaling pathway</i>	Y	MAP3K7 (1)(28)(2), NFATC2 (0)(1)(2),
<i>Cell cycle</i>	Y	SMAD3 (1)(14)(4), SMAD2 (0)(28)(1),
<i>TGF-beta signaling pathway</i>	Y	SMAD6 (2)(7)(4), SMAD3 (1)(14)(4), SMAD2 (0)(28)(1), SMURF1 (0)(4)(3), TGFB2 (0)(6)(1),
ErbB signaling pathway	N	
<i>Focal adhesion</i>	Y	IGF1R (0)(2)(5), LAMA1 (0)(1)(6), ITGB6 (0)(4)(4), ITGA1 (0)(5)(2), ITGB5 (1)(5)(3),
Proteasome	N	
<i>Adherens junction</i>	Y	PTPRB (1)(2)(1), PTPRM (0)(10)(6),
Notch signaling pathway	N	NCOR2 (0)(2)(1),
<i>Regulation of actin cytoskeleton</i>	Y	ITGB5 (1)(5)(3),
<i>Neurotrophin signaling pathway</i>	N	RPS6KA2 (0)(6)(1),
Chronic myeloid leukemia	Y	SMAD3 (1)(14)(4),
Apoptosis	N	BID (0)(1)(1),
Pathways in cancer	Y	SMAD3 (1)(14)(4), CSF1R (0)(3)(1),
Tight junction	Y	MAGI2 (0)(12)(11), EPB41 (2)(8)(5), MPDZ (0)(4)(1), PRKCE (0)(4)(7), JAM3 (1)(1)(1), CTNNA2 (1)(12)(3),
Long-term potentiation	N	GRIN2A (0)(7)(1), PLCB1 (0)(44)(1), GRM1 (0)(1)(3),
Measles	N	
T cell receptor signaling pathway	N	
Nucleotide excision repair	N	RPA3 (0)(1)(2),
Chemokine signaling pathway	N	VAV3 (0)(8)(1), PLCB1 (0)(44)(1),
Calcium signaling pathway	Y	GNA14 (0)(5)(1), NOS1 (0)(1)(1), ADCY2 (0)(8)(5), CHRM2 (0)(1)(1), ADRA1A (0)(2)(4), PLCB1 (0)(44)(1), CACNA1C (1)(3)(5), GRM1 (0)(1)(3),

respectively, only 25 genes (as shown in Table 4.10) are targeted by SNPs in both populations. In the 7 commonly identified pathways, while 15 of the SNP targeted genes (STGs, shown in Table 4.10) are common between populations. 62 and 51 of the STGs are unique to EU and JP populations, respectively. In these 7 commonly found pathways, there were 724 and 195 SNPs unique to EU and JP populations, respectively, and 6 SNPs were common. There were very few commonly affected SNPs/genes and many distinct sets of SNPs/genes targeting the same pathways for each population, which strongly supports our hypothesis. Hence, if one follows a gene or SNP oriented approach, crucial information for disease development mechanism might be missed. Instead, here we emphasize the importance of a pathway oriented approach to investigate the etiology of IA. The 7 pathways in top 10 are MAPK signaling, Cell cycle, TGF-beta signaling, Focal adhesion, Adherens junction, Regulation of actin cytoskeleton, and Neurotrophin signaling pathways, as shown in Table 4.9. In these commonly found pathways, we checked the number of STGs, and the number of typed SNPs separately for EU and JP populations and the commonality of these entities between the two populations. For example in MAPK signaling pathway, there were 14 and 18 STGs in EU and JP populations, respectively. Among these genes, only 2 of them (MAP3K7, NFATC2, as shown in Table 4.10) were common, indicating that the same pathways can be targeted via independent genes in diverse populations. There were 133 and 43 typed SNPs in EU and JP populations, respectively and among these SNPs only 1 of them (rs791062) was common. In addition to these typed SNPs that were commonly identified in both populations, the commonly identified SNP targeted genes harbour other disease predisposing SNPs in different populations. For example, MAP3K7 gene is associated with 28 other typed SNPs in EU population that is not found in JP population. These observations were true for all the 7 commonly found pathways and the genes within them. These results show the relevance of our pathway oriented approach and indicate that if there is a problem in these seven pathways, the disease is more likely to happen.

We also searched for known IA related pathways in KEGG Disease Pathways Database (KEGGDPD) using “aneurysm” as a keyword, which resulted in three hits (H00801, H00800, and H00579). Seven pathways from our top twenty pathway list (MAPK signaling, TGF-beta signaling, Calcium signaling, Focal adhesion, Adherens junction, Tight junction, Regulation of actin cytoskeleton) were amongst the twelve pathways

found to be associated with aneurysm related diseases in KEGGDPD.

Next, we searched for the affected pathways using the gene expression data, which is obtained from ruptured and unruptured IA patients with Japanese ethnicity as cases; and from arteriovenous malformation feeders with Japanese ethnicity as intracranial controls (Krischek, et al., 2008). Even though gene expression and GWAS data are not coming from the same samples, the enriched pathways might show commonalities. Therefore, we mapped the differentially expressed genes to PPI and proceeded with the following steps of PANOGA to detect affected pathways. The top 20 over-represented KEGG pathways identified for gene expression data are shown in Table 4.11. As expected, there is no strong correlation between the rankings of the affected pathways, obtained from GWAS and expression data. Because, the transcriptomics data only includes genes with significant changes in expression levels, whereas, GWAS data includes genes, affected by several factors. Still, compared to the top ten pathways identified GWAS in EU and JP populations, Ribosome pathway is also found by GWAS data on Japanese population (with 5th ranking); ErbB signaling pathway and Proteasome pathways are also found by GWAS data on European population (with 4th and 6th rankings, respectively); Adherens Junction (AJ), Focal Adhesion (FA) and Neurotrophin Signaling (NS) pathways are also found by GWAS data on both Japanese and European populations (with 1st and 7th (AJ), 5th and 2nd (FA), 7th and 10th rankings (NS), respectively). In these 6 pathways (Adherens junction, Focal adhesion, ErbB signaling, Neurotrophin signaling, Ribosome, Proteasome pathways), 25 out of 379 genes were commonly identified with GWAS results. Among these genes, PTPRB gene, as part of the Adherens Junction pathway, is known to have a crucial role in blood vessel remodeling and angiogenesis. Even though this gene is not found to be differentially expressed in Japanese population, rs1561798 variant that this gene contains, is found to be significant in the GWAS of both European and Japanese populations. Interestingly, another gene expression study on IA found this gene to be differentially expressed in Polish population (Pera, et al., 2010). Although PTPRB gene is not found to be differentially expressed in JP population, using GWAS data in EU and JP populations, PANOGA was able to identify this gene as part of an important pathway for IA development mechanism.

Table 4.11 The top 20 over-represented KEGG pathways identified for gene expression data of IA.

KEGG Term	KEGG Term P-values Corrected with Bonferroni			Rankings		
	Gene Expression	GWAS EU	GWAS JP	Gene Expression	GWAS EU	GWAS JP
Ribosome	7.91E-23	1.40E-08	5.93E-19	1	73	5
Spliceosome	7.40E-17	2.05E-13	4.72E-13	2	33	27
RNA transport	3.97E-14	6.26E-09	-	3	69	-
Complement and coagulation cascades	6.05E-13	7.00E-14	1.06E-09	4	31	48
T cell receptor signaling pathway	7.86E-12	1.62E-16	1.97E-15	5	17	15
ErbB signaling pathway	5.70E-09	9.52E-22	2.47E-15	6	4	16
Chronic myeloid leukemia	6.70E-09	2.62E-18	8.13E-11	7	11	36
Natural killer cell mediated cytotoxicity	9.96E-09	2.56E-07	1.29E-09	8	81	50
RNA degradation	1.44E-08	3.44E-11	1.66E-07	9	44	67
Osteoclast differentiation	1.45E-08	8.12E-15	4.97E-10	10	26	43
Neurotrophin signaling pathway	6.68E-08	2.49E-18	1.92E-18	11	10	7
Adherens junction	1.74E-07	4.91E-19	2.58E-21	12	7	1
mRNA surveillance pathway	3.59E-07	-	-	13	-	-
Pyruvate metabolism	1.87E-06	-	5.82E-05	14	-	92
Toll-like receptor signaling pathway	3.26E-06	9.18E-13	1.50E-10	15	35	38
Small cell lung cancer	3.55E-06	-	1.01E-08	16	-	55
Proteasome	4.19E-06	2.35E-21	4.54E-11	17	6	35
Focal adhesion	8.57E-06	9.55E-22	5.60E-21	18	5	2
Fc gamma R-mediated phagocytosis	1.47E-05	4.00E-09	1.32E-13	19	66	22
Toxoplasmosis	2.68E-05	1.06E-08	-	20	72	-

4.5 Results on Behçet's disease dataset

We applied PANOGA separately on two Behçet's disease GWASs: i) Turkish (TR) population of 1,215 cases and 1,278 control cohorts, ii) Japanese (JP) population of 612 cases and 740 controls. In our analysis, we have included 18,479 SNPs from TR population specific dataset, and 20,594 SNPs from the JP population specific dataset with p-values < 0.05 , where the genotypic p-value of a SNP is calculated via allelic chi-squared test.

To identify the biological pathways with the genes responsible for Behçet's disease susceptibility, we applied the affected SNP functionalization, SNP to gene mapping,

gene-wise weighted p-value calculation, sub-network identification and functional enrichment steps of PANOGA separately for each dataset. The details of these steps are explained in the methods section. After the gene mapping step, while the affected SNPs from TR population map to 3,869 genes, the affected SNPs from JP population map to 4,076 genes. As shown in Table 4.12, five of the top ten affected pathways were found to be commonly identified in both TR and JP populations. These pathways are Notch signaling pathway, Focal adhesion, Jak-STAT signaling pathway, Long-term potentiation and Pathways in cancer. In these five pathways, as shown in Table 4.13, in the five commonly identified pathways, 36 of the SNP targeted genes are common between populations. Only 9 of the SNPs that are targeting these 36 genes were common between TR and JP populations. Similar to our results on intracranial aneurysm dataset, the identified pathways between two populations show more commonality than individual genes or SNPs.

We also searched for known Behçet's disease related pathways in KEGGDPD using “Behçet” as a keyword, which resulted in one hit (H00106), including one pathway (Complement and coagulation cascades). This pathway is identified in 5th and 33rd rankings with $P=2.47E-20$, $P=2.6E-12$ in TR and JP populations, respectively.

Table 4.12 The top 10 KEGG pathways identified for both populations in Behçet’s disease. 5 out of the top 10 pathways are shown in bold.

KEGG Term	P-values		Rank		# of Associated SNPs in GWAS		# of SNP Targeted Genes (STGs)		# of Common STGs	% Common Genes in Both Populations		Is Common Genes more than 50% in any population?
	TR	JP	TR	JP	TR	JP	TR	JP		TR	JP	
Notch signaling pathway	1,53E-25	4,66E-17	1	10	37	11	9	6	5	55.55	83.33	Y
Ribosome	7,62E-24	1,28E-15	2	14	5	4	4	2	0	0.0	0.0	N
Focal adhesion	1,15E-20	2,20E-18	3	4	65	80	25	20	7	27.99	34.99	N
Jak-STAT signaling pathway	1,28E-20	2,26E-18	4	5	32	44	16	16	3	18.74	18.74	N
Complement and coagulation cascades	2,48E-20	2,60E-12	5	33	25	27	13	8	3	23.07	37.49	N
Long-term potentiation	4,86E-20	2,69E-18	6	6	59	88	14	16	8	57.14	49.99	Y
Long-term depression	3,30E-19	1,22E-14	7	18	59	73	15	10	9	59.99	89.99	Y
Pathways in cancer	4,30E-19	1,18E-17	8	9	79	98	25	26	4	15.99	15.38	N
Proteasome	1,55E-18	2,65E-16	9	12	2	3	1	3	0	0.0	0.0	N
ECM-receptor interaction	1,27E-17	4,56E-12	10	38	19	41	10	10	4	39.99	39.99	N

Table 4.13 The top 10 over-represented KEGG pathways for Behçet’s disease, and the SNP targeted genes within these pathways. Five out of the top ten affected pathways in both TR and JP populations are shown in bold. SNP Targeted Genes that are identified in both TR and JP populations are shown in the last column; along with the number of commonly typed SNPs in both populations, only in TR population and only in JP populations are shown in paranthesis.

KEGG Term	Any Common SNPs in Common Genes?	Commonly Associated, SNP Targeted Genes and their SNP Counts: (Common) (TR GWAS) (JP GWAS)
Notch signaling pathway	N	CTBP1 (0)(1)(1), KAT2B (0)(2)(1), MAML2 (0)(15)(4), MAML3 (0)(2)(3), NCOR2 (0)(2)(1),
Ribosome	N	
Focal adhesion	Y	PRKCA (0)(4)(5), COL4A2 (1)(3)(12), VAV3 (0)(6)(7), LAMA5 (0)(2)(1), CAPN2 (0)(3)(2), LAMB1 (1)(1)(1), FLNB (0)(2)(6),
Jak-STAT signaling pathway	N	IL2RB (0)(2)(1), OSMR (0)(1)(1), JAK2 (0)(3)(14),
Complement and coagulation cascades	N	PLAT (0)(1)(1), F5 (0)(3)(2), F13A1 (0)(4)(7),
Long-term potentiation	N	GRM5 (0)(3)(10), PRKCA (0)(4)(5), GRIA1 (0)(2)(15), ADCY8 (0)(3)(1), GRIN2A (0)(16)(10), CACNA1C (0)(5)(2), PLCB1 (0)(3)(10), ITPR1 (0)(5)(4),
Long-term depression	Y	GRM5 (0)(3)(10), PRKCA (0)(4)(5), LYN (0)(1)(2), GRIA1 (0)(2)(15), PLCB1 (0)(3)(10), ITPR3 (0)(9)(6), PRKG1 (1)(11)(8), GRM1 (0)(3)(1), ITPR1 (0)(5)(4),
Pathways in cancer	N	PRKCA (0)(4)(5), CTBP1 (0)(1)(1), ETS1 (0)(5)(3), MITF (0)(1)(1),
Proteasome	N	
ECM-receptor interaction	N	LAMA1 (0)(3)(1), CD44 (0)(1)(1), LAMA5 (0)(2)(1), SDC2 (0)(1)(1),

CHAPTER 5

5 DISCUSSION

Many reports of the genome wide association studies emerging in the literature, and the online GWAS catalog, including 273 published GWAS so far by National Human Genome Research Institute (NHGRI), are the clear evidences of the success of GWAS. Unfortunately, using the traditional approaches in GWAS, only the strongest associations can be detected; and there are many more SNPs/genes still to be found as associated with disease (Couzin and Kaiser, 2007; Williams, et al., 2007). Lately, several GWAS (Lesnick, et al., 2007; Pattin and Moore, 2008; Torkamani, et al., 2008; Wang, et al., 2007; Wilke, et al., 2008) have proposed the use of prior knowledge in the form of pathway databases, such as the KEGG and Biocarta, or gene ontology databases. On the other hand, (Franke, et al., 2006) suggested the use of protein interaction network information along with pathway-based analysis. For Multiple Sclerosis GWAS data, (Baranzini, et al., 2009) demonstrated the utility of network-based analysis. On top of these pathway and network based analyses of GWAS, here we devised a methodology that also integrates the functional information of a SNP as a third component. As a result of this multidimensional screening approach, our methodology generated a comprehensive list of functionally important KEGG pathways.

5.1 Discussion on rheumatoid arthritis dataset

While most of these associations can be thought as computational predictions, the functional relations of five of these pathways (Jak-STAT signalling, apoptosis, T cell receptor signalling, leukocyte transendothelial migration and cytokine-cytokine receptor interaction) with RA pathogenesis are shown in the reviews by (Plenge, et al., 2007; Raychaudhuri, et al., 2008; Raychaudhuri, et al., 2009).

Additionally, the effect of Toll-like receptor (TLR) and MAPK signaling pathway on RA is known as following: TLRs are membrane-bound receptors which are expressed in innate immune cells, such as macrophages and dendritic cells. TLR signaling plays an important role in the activation and direction of the adaptive immune system by the up-regulation of co-stimulatory molecules of antigen presenting cells. The activation of the TLRs signaling pathway can trigger the activation of the MAPK and NF- κ B pathways. Evidence is emerging that certain TLRs play a role in the pathogenesis of infectious and/or inflammatory diseases. There is considerable evidence from rodent models that activation of the TLRs can induce or exacerbate inflammatory arthritis (Joosten, et al., 2003).

The role of MAPK signaling pathway in the development and progress of RA was shown to be related to cartilage damage, which is a hallmark of RA. Cartilage damage is based on increased proteoglycan loss as well as attachment and invasion of inflammatory tissue into the cartilage, which leads to its structural disintegration. Production of matrix metalloproteinases (MMPs) by synovial tissue appears to be a key prerequisite for synovial tissue to invade and destroy cartilage. MAPK is a crucial signal transduction pathway for inflammation and carries information about inflammatory stimuli to the cell nucleus. Synthesis of MMPs is regulated through multiple MAPK families, suggesting that a blockade of MAPK might have structural benefit in arthritis (Liacini, et al., 2003; Suzuki, et al., 2000). Also, activation of stress kinase pathways ERK, JNK, and p38 MAPK is a typical feature of chronic synovitis during RA, and several proinflammatory mediators use the signaling of these stress kinase pathways (Schett, et al., 2000).

Cytokine-cytokine receptor interaction pathway has been previously identified by two other studies as RA associated and included in the KEGG functional enrichment of known disease genes (Martin, et al., 2010; Zhang, et al., 2010). Even though this term has not been found as significant in our highest scoring sub-network, it has been identified in the functional enrichment of our third highest scoring sub-network. Due to the nature of the search algorithm used by jActive Modules, several of the identified sub-networks overlap extensively in their component genes. Since it is complicated and cumbersome to represent the enrichment analysis of all identified sub-networks, here we have shown only the results from our highest scoring sub-network. In future, we aim to visualize the KEGG enrichment analysis results from all identified 5 top scoring sub-networks in a comprehensive manner.

5.2 Discussion on partial epilepsy dataset

The knowledge of genes, proteins, and pathways changed during the different stages of epilepsy development is crucial to enlighten the pathophysiology of epilepsies and to develop new therapeutic strategies based on drugs with anti-epileptogenic activities. The synchronized neuronal activity and unbalance between inhibitory and excitatory neurotransmission are known as the common features linked to the pathogenesis of epilepsy (Dalby and Mody, 2001). Hence, the mechanisms of action of most clinically used drugs in epilepsies are based on these biological processes. In this regard, voltage-gated ion channels, gabaergic, and glutamatergic systems are the well-known therapeutic targets. In addition to these well studied examples, several groups conduct gene-expression studies, and a few others perform GWAS, CNV studies to uncover the molecular mechanisms involved in epileptogenesis. Here, we would like to discuss our findings in relation to the identified pathways as part of these previous studies.

Aronica *et al.* detected complement and coagulation cascade pathway as a result of gene expression profile analysis of epilepsy-associated gangliogliomas (GG) (Aronica, et al., 2008). As shown in Figure 5.1.a, they report that C1qa, C1qb, C1qc, C1r, C1s, C3, C4a and C7 genes as part of this pathway showed more prominent expression in GG

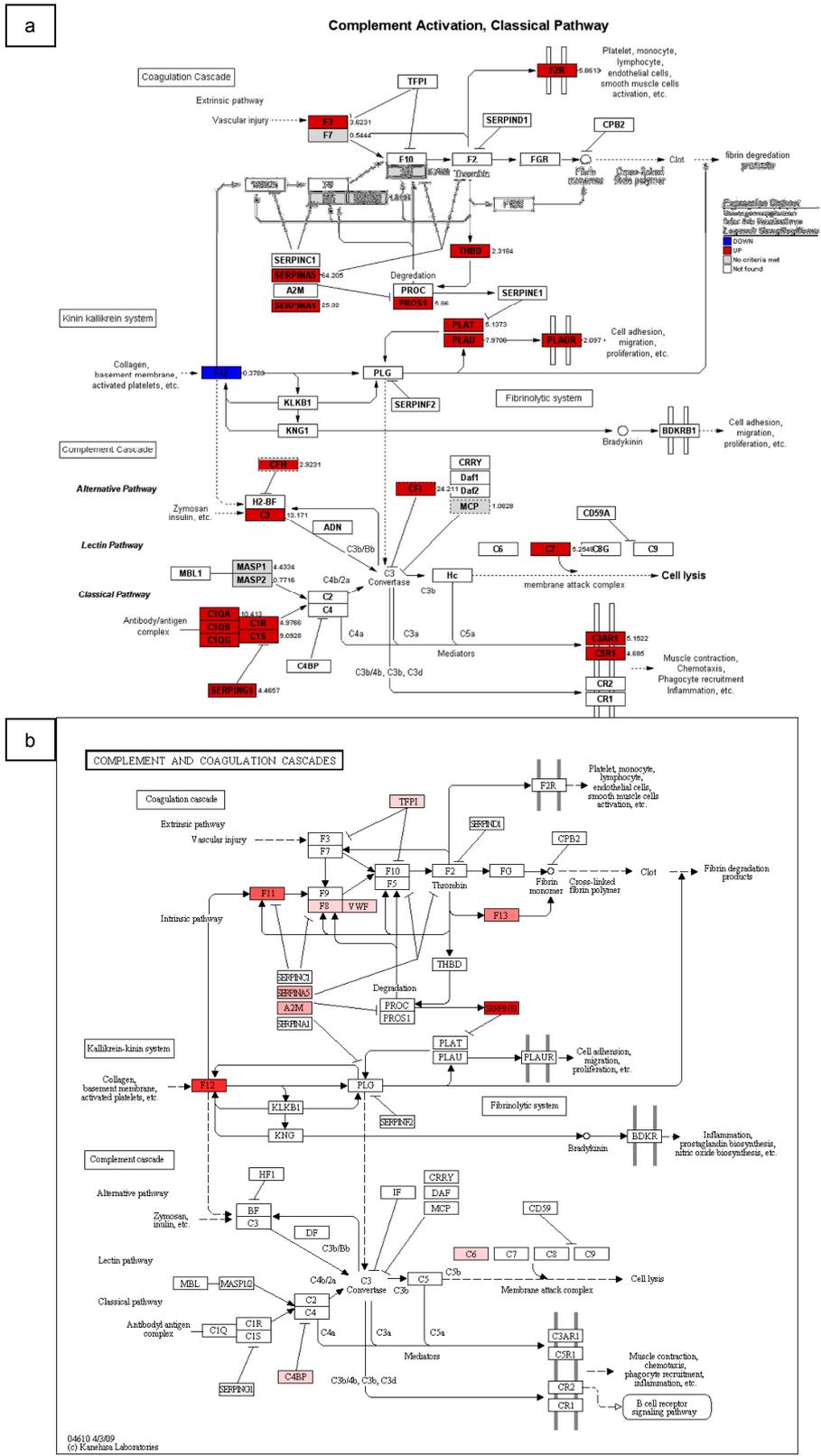


Figure 5.1 The complement and coagulation cascade (a) Up and down-regulated genes are shown in red and in blue, respectively, as a result of microarray analysis for epilepsy-associated gangliogliomas (Aronica, et al., 2008). (b) The shade of red color in genes indicates the number of GWAS targeted SNPs per base pair of the gene. Red

refers to the highest targeted gene, whereas white refers to a gene product, not targeted by the SNPs.

specimens than in control specimens. Expression of SerpinG1, a C1 inhibitor, and CD59 an inhibitor at the level of the membrane attack complex was also higher, but to a lesser extent compared to C1q genes. They claim that the complement and the IL- β 1 system are indeed activated in different human epilepsy associated lesions (Aronica, et al., 2008; Aronica, et al., 2007; Jamali, et al., 2006) consolidates the preclinical findings. Figure 5.1.a shows that the differentially expressed genes exist in both complement and coagulation cascades, according to the microarray study of epilepsy-associated gangliogliomas. On the other hand, in Figure 5.1.b, we show that mostly the coagulation cascade of this pathway is affected by the genes, which are targeted by the genotyped SNPs (SNPs that are found to be significant for PE in a GWAS (Kasperaviciute, et al., 2010)). As part of the complement cascade, only C6 and C4BP genes (shown in light pink in Figure 5.1.b) are targeted by the genotyped SNPs. This example supports our hypothesis that pathways can be used as markers of diseases. While the transcriptomics data only includes genes with significant changes in expression levels, GWAS data includes genes, affected by several factors. Different factors cripple distinct parts of the pathways, as shown here on complement and coagulation cascades with a comparison of gene expression vs GWAS study.

In parallel with our results, MAPK, Wnt, Notch, TGF-beta, Jak-STAT and Calcium signaling pathways are identified in Okamoto et al's study as regulatory signaling pathways, which include over- and hypo-expressed genes during all experimental times studied after status epilepticus (Okamoto, et al., 2010). Among these pathways, MAPK, Jak-STAT, and TGF-beta were found regulated in pilocarpine-treated animals throughout the epileptogenesis period evaluated. In order to confirm the microarray results, they quantified the differential expression of the selected genes Nestin, CDK1, p18 (INK4c), TGF-b1, IGF-1 and GFAP by real-time PCR and found similar results with their microarray study (Okamoto, et al., 2010). Ye et al analyzed microarray data of temporal epilepsy from Gene Expression Omnibus and reported that the main biological functions shared among the 71 differentially expressed genes included MAPK and Calcium signaling pathways (Zhou, et al., 2011). Jimenez-Mateos et al

investigated hippocampal transcriptome after status epilepticus in mice, and identified MAPK, TGF-beta, Jak-STAT pathways (Jimenez-Mateos, et al., 2008). In their study, the pathway containing the most differentially down-regulated genes in tolerance was calcium signaling, including the genes associated with the plasma membrane (ionotropic glutamate receptor and voltage-dependent calcium channel genes) and genes downstream of endoplasmic reticulum calcium stores. They also validate the expressional changes of a selection of these genes by quantitative PCR (Jimenez-Mateos, et al., 2008). Limviphuvadh et al investigated the genes on the chromosomal region 4p15, which is shown previously as the partial epilepsy with pericentral spike locus (Limviphuvadh, et al., 2010). They detected 14 candidate genes in this region, which are found to be deleted in both of the two independent studies, describing patients that share the epilepsy-like seizures. Among these genes, they report that CCKAR gene functions in the nervous system and they show this gene as part of the calcium signaling pathway (Limviphuvadh, et al., 2010). STAT3 gene, as part of the JAK-STAT pathway was shown to be phosphorylated both at Ser727 by mTORC1 (Yokogami, et al., 2000) and at Tyr705 by Janus kinases (Reich, 2009), leading to the regulation of several genes in varying cellular processes. In an experimental epileptical model based on rats, when rats have been injected with kainates, they have been shown to have STAT1, STAT3 and p42/44 MAPK activated in their hippocampus preceding epileptic seizures (Choi, et al., 2003). On the other hand, CREB is actively expressed in surgically obtained epileptic hippocampus material, leading to proliferating reactive astrocytes specifically localized to the hippocampal sclerosis region (Morimoto, et al., 2004). Our analyses have revealed Wnt and Notch signaling pathways involving CREBBP gene, which can indicate roles for these pathways in seizure formation through the protein-protein interactions of CREB and CREBBP.

In addition to the signaling pathways, cell cycle pathway is identified in 2nd ranking with $P=1.03E-24$, as shown in Table 4.8. Jimenez-Mateos et al also identified this pathway and apoptosis pathway (identified in 8th ranking with $P=3.72E-16$ in our study) (Jimenez-Mateos, et al., 2008). ANAPC4 gene as part of the cell cycle pathway is detected in Limviphuvadh et al's study, as one of the 14 candidate genes (Limviphuvadh, et al., 2010). Among the cell cycle regulators, mutations in Cdk1 inactivates TSC1 leading to the onset of Tuberos Sclerosis Complex and 70% of individuals affected with TSC known to develop epilepsy (Cho, 2011).

5.3 Discussion on intracranial aneurysm dataset

The pathway and network oriented analysis of GWAS data in two different populations together with gene expression data gave us the tools to investigate the pathogenesis of IA. The genes that are found to be targeted by disease predisposing SNPs are shown to be involved in several biological pathways including MAPK signaling, Cell cycle, TGF-beta signaling, Focal adhesion, Adherens junction, Regulation of actin cytoskeleton, and Neurotrophin signaling pathways. Since these pathways are known to have a role in the regulation of cell growth, tissue remodeling, inflammation, and wound healing, they are likely to contribute to the pathophysiology of IA. In addition to these top ten pathways, here, we also would like to discuss in detail the identified signaling pathways from top 20 list, that are functionally relevant to the pathogenesis of IA.

The mitogen-activated protein kinases (MAPKs) are serine-threonine kinases that are involved in intracellular signaling related with several cellular activities such as cell proliferation, differentiation, survival, death and transformation (Kholodenko and Birtwistle, 2009; McCubrey, et al., 2006). Laaksamo et al. studied the expression and phosphorylation of the 3 major MAPKs in unruptured and ruptured human IAs: c-Jun N-terminal kinase (JNK), p38, and extracellular signal-regulated kinase (Laaksamo, et al., 2008). Their study shows that JNK and p38 expression have role in IA growth; and JNK activity and expression have possible role in rupture (Laaksamo, et al., 2008). As shown in Table 4.9, this pathway is identified in 1st and 8th rankings with $P=3.53E-27$, $P=2.70E-18$ in EU and JP populations, respectively. As shown in Figure 5.2 in red, and in Table 4.10, in this pathway, MAP3K7 (TAK1) and NFATC2 genes are identified in our method both by EU and JP GWAS. There are 28 typed SNPs on MAP3K7 gene according to EU GWAS and 2 typed SNPs according to JP GWAS; and among those SNPs, 1 SNP is identified in both studies. As shown in the KEGG pathway map in Figure 5.2, TAK1 gene is shown to have a downstream effect on Wnt signaling and the pathways of proliferation, inflammation, and anti-apoptosis. Additionally, as part of this

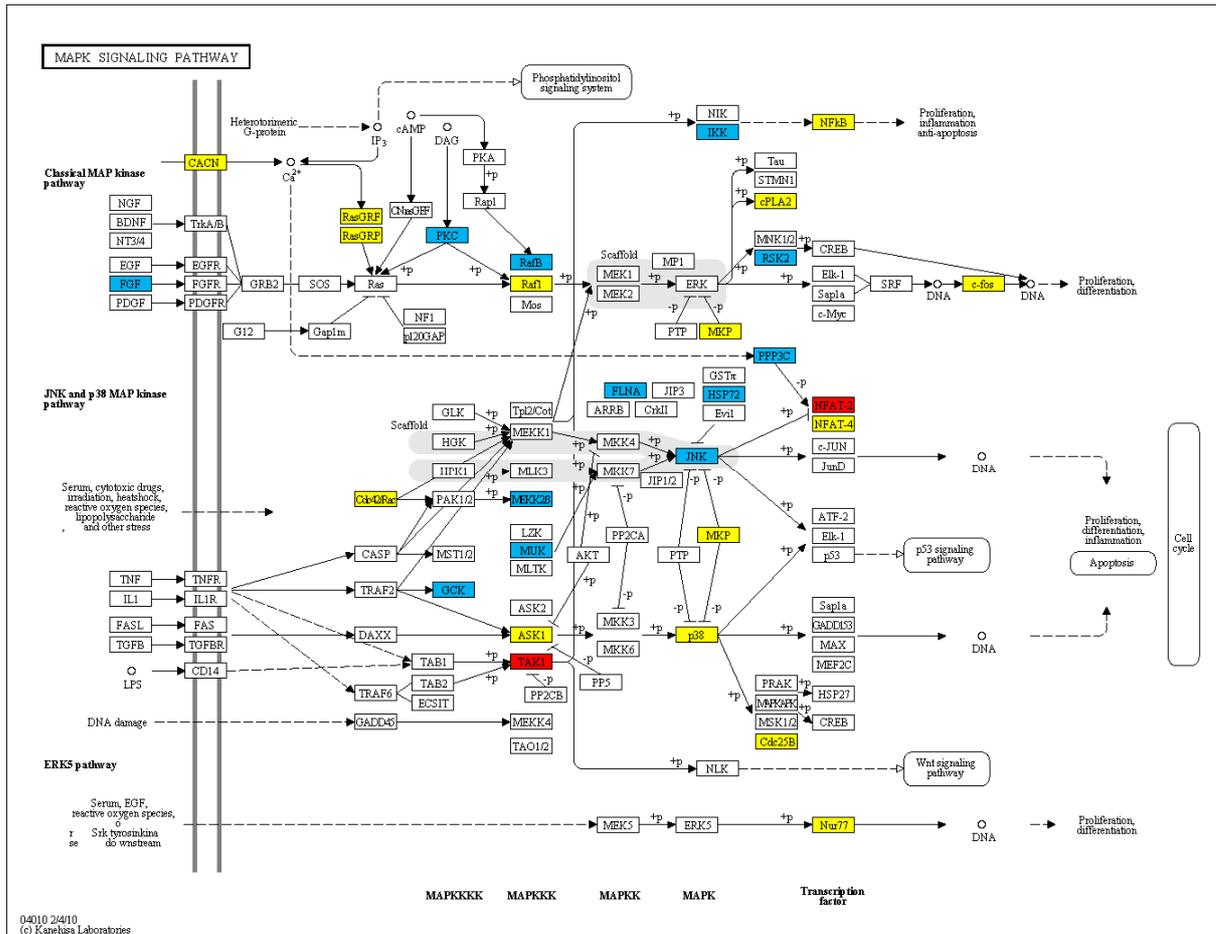


Figure 5.2 KEGG pathway map for MAPK signaling. The set of genes shown in blue includes genes that are found for EU dataset; yellow includes genes that are found for JP dataset; red includes genes that are found both by EU and JP GWAS of IA.

pathway, HSPA1L, PRKCA, BRAF, RPS6KA2, MAP3K2, MAP4K2, PPP3CA, MAPK10, FGF12, FLNB, CHUK, MAP3K12 genes are uniquely found in EU population (shown in blue in Figure 5.2) and DUSP10, RAF1, NR4A1, NFKB1, CACNG2, CDC25B, FOS, PLA2G4A, RPS6KA3, MAP3K5, RASGRP3, RASGRF1, MAPK14, RAC1, NFATC4, CACNA1C genes are uniquely found in JP population (shown in yellow in Figure 5.2).

The transforming growth factor beta (TGF-beta) signaling pathway is known to play a role in aortic aneurysms and also has a possible role in aneurysms in general (Ruigrok, et al., 2008). Additionally, TGF-beta signaling is shown to drive aneurysm progression in multiple disorders, including Marfan syndrome (Holm, et al., 2011). It is reported

that therapies that inhibit this signaling cascade are in clinical trials in mice (Holm, et al., 2011). As shown in Table 4.9, this pathway is identified in 3rd and 9th rankings with $P=6.26E-24$, $P=2.41E-17$ in EU and JP populations, respectively. In our analysis, we detected 15 and 9 SNP targeted genes in EU and JP populations, respectively. As shown in Table 4.10, 5 of these genes (SMAD6, SMAD3, SMAD2, SMURF1, TGFB2) are identified in both populations; and 2 of these 5 genes, SMAD3 and SMAD6, have common typed SNPs. SMAD2 in this pathway harbors 28 typed SNPs in EU population which is not observed in JP population. In Figure 5.3, the KEGG pathway map of TGF-beta signaling shows that SMAD6 gene (shown in red) is targeted by typed SNPs in JP population and it inhibits the formation of SMAD2/3 complex (shown in pink). The colors of the genes in Figure 5.3 indicate the number of targeted SNPs in JP population per base pair of the gene, from red to white. SMURF1 (shown in pink) inhibits TGFBR2 (shown in pink with blue border), that also binds to TGFβ (shown in pink). TGFBR2 gene is found to be differentially expressed. As a downstream effect, SMAD2/3 complex (shown in pink) is affected as well as the transcription factors, co-activators, and co-repressors. As shown in Figure 5.3, this cascade of events leads to angiogenesis, and neogenesis. Our method detected ten additional genes (ACVR2B, SMAD9, SMAD7, GDF5, SMAD4, SMAD1, BMP7, BMPR1B, BMPR1A, BMP6) that are affected in EU population, but not in JP population. These genes are not colored in Figure 5.3.

Several putative risk genes were suspected to play a role in cell-cycle progression, potentially affecting the proliferation and senescence of progenitor-cell populations that are responsible for vascular formation and repair (Yasuno, et al., 2010). As shown in Table 4.9, Cell-cycle pathway is identified in 2nd and 4th rankings with $P=2.35E-25$, $P=2.81E-19$ in EU and JP populations, respectively.

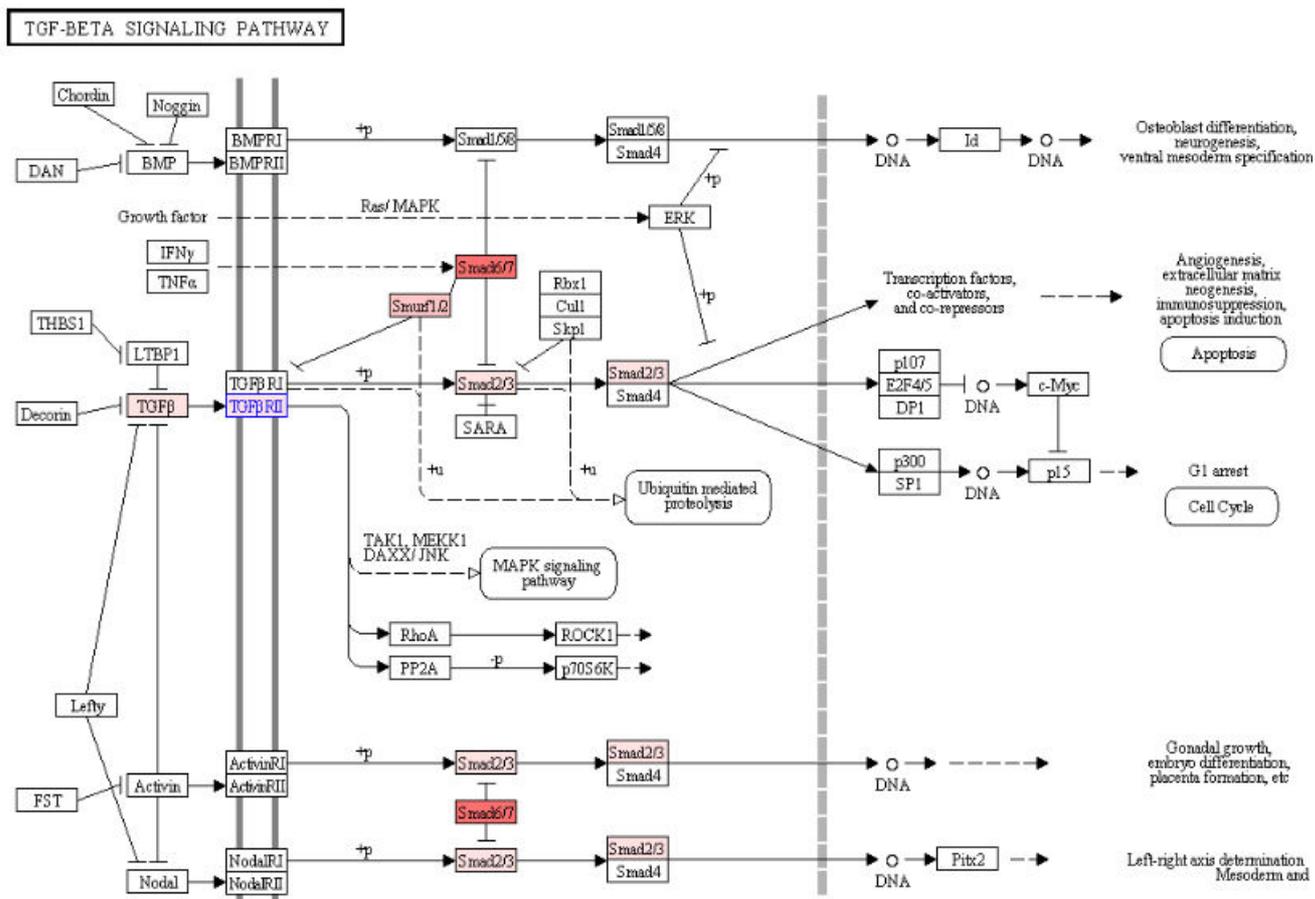


Figure 5.3 KEGG pathway map for TGF-beta signaling pathway. The shade of red color in gene indicates the number of targeted SNPs in JP population per base pair of the gene. Red refers to the highest targeted gene, whereas white refers to a gene product, not targeted by the SNPs. Blue border indicates that the gene is found to be differentially expressed.

Calcium is a key signaling ion that controls many different cellular processes, such as gene transcription, synaptic activity, muscle contraction, cell-cell communication, adhesion and cell proliferation (Hofer, 2005; Marambaud, et al., 2009). The calcium signaling pathway has a significant role in regulating a great variety of neuronal processes (Edwards, et al., 2011). As shown in Table 4.9, we identify this pathway in 20th and 12th rankings with $P=3.27E-15$, $P=1.37E-15$ in EU and JP populations, respectively. In this pathway, we suspect a mechanism related to autocoids and GPCRs for IA disease development. As shown in Table 4.10 and in red in Figure 5.4, GPCR, Gq, PLCB1 genes are detected in our methodology both by EU and JP GWAS. These genes are found on our suspected autocoid path in calcium signaling pathway. There are

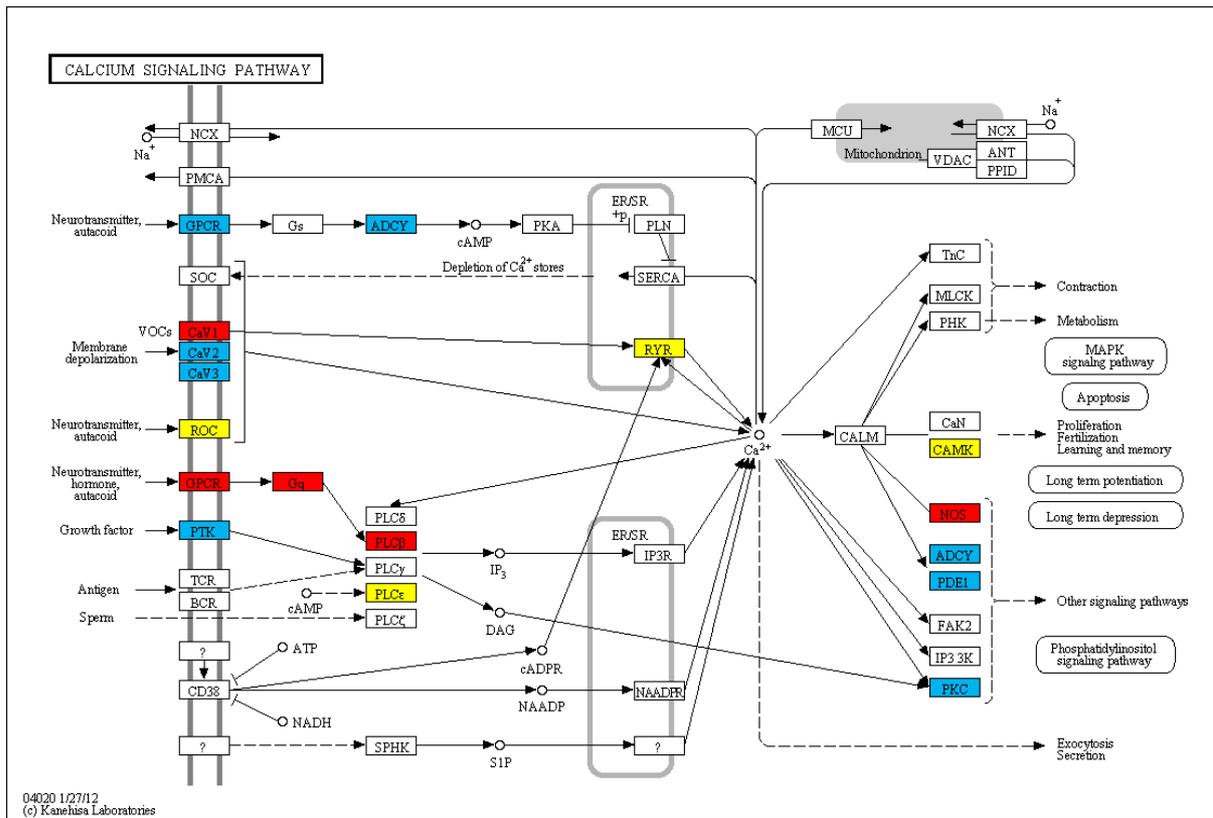


Figure 5.4 KEGG pathway map for calcium signaling pathway. The set of genes shown in blue includes genes that are found for EU dataset; yellow includes genes that are found for JP dataset; red includes genes that are found both by EU and JP GWAS of IA.

44 marker SNPs on *PLCB1* gene according to EU GWAS and 1 marker SNP according to JP GWAS; and none of those SNPs are identified in both studies. As part of our suspected autocoid path, Kuo et al. has shown the association of a polymorphism of *ITPKC* (inositol-trisphosphate 3-kinase C, *IP3-3KC*) with the susceptibility and aneurysm formation in KD patients in a Taiwanese population (Kuo, et al., 2011). *ITPR1* (inositol 1.4.5-trisphosphate receptor, type 1, *IP3R*) is identified in our analysis as part of Calcium signaling pathway and it is also found as differentially expressed between aneurysm patients and controls in JP population. Calcium signaling pathway's high rank in our analysis, and our suspected autocoid path within this pathway also fit well with the recent work which reports that Clazosentan is in phase III trial to reduce vasospasm caused by Endothelin A autocoid (Feigin and Findlay, 2006; Zhou, et al., 2011).

5.4 Discussion on Behçet's disease dataset

Out of the top ten affected pathways, notch signaling, focal adhesion, Jak-STAT signaling, pathways in cancer and long-term potentiation pathways were commonly identified in both Turkish and Japanese populations. Among these pathways, here we would like to emphasize the possible role of Jak-STAT signaling pathway for Behçet's disease development mechanism. Because, it has been reported that different GWAS identified Jak-STAT signaling pathway as highly relevant to human autoimmunity and targeting JAKs is now a reality in immune-mediated disease (O'Shea and Plenge, 2012). This pathway is identified in 4th and 5th rankings with $P=1.28E-20$, $P=2.26E-18$ in TR and JP populations, respectively. As part of this pathway, while 16 genes are identified in TR population and 16 genes are identified in JP population, only three of these genes (IL2RB, OSMR, JAK2) are commonly detected between these two populations. None of these genes are targeted by the same SNP, which is found to be significant in a particular population's GWAS. It is especially interesting for JAK2 gene, which is targeted by 14 genotyped SNPs in Japanese population, and none of these SNPs target the same gene in Turkish population. Hence, if one searches for conserved SNPs between populations, such important clues, illuminating an aspect of disease etiology, might have been missed. Therefore, to understand the underlying mechanism of complex diseases, one should find out affected pathways targeted by several genetic variants. Also, similar to our results on intracranial aneurysm dataset, the identification of the five same pathways out of the top ten pathways in both Turkish and Japanese population showed that our results are independent from disease.

5.5 General Discussion

The identification of significant individual factors causing complex diseases is challenging in genome-wide association studies (GWAS), since each factor would have a modest effect on the disease development mechanism. In this thesis, we hypothesize that the biological pathways that are targeted by these individual factors show higher conservation within and across populations. To test this hypothesis, we searched for the disease related pathways on i) two intracranial aneurysm GWAS in European and Japanese case-control cohorts; ii) two Behçet's disease GWAS in Turkish and Japanese

case–control cohorts. Even though there were a few significantly conserved SNPs within and between populations, seven of the top ten and five of the top ten pathways were significantly identified in both populations for IA and Behçet's disease datasets, respectively. The probability of random occurrence of such an event is $2.44E-36$. Hence, our results indicate that even though each individual has a unique combination of factors involved in disease development mechanism, most of the targeted pathways that need to be altered by these factors are mostly the same.

It is noteworthy to mention that pathway-based analyses, like it is presented here, are limited to our knowledge of cellular processes. The biological functions of most of the genes in the genome are not known. Since network and pathway tools make use of functional information from gene and protein databases, they are biased toward the well-studied genes, interactions, and pathways. Also, variants associated to genes not represented in the protein-protein interaction network were not evaluated in this analysis. Nevertheless, there is scope for the development of related methodologies to increase the power to detect associations in these genes. By combining information from several sources (functional properties of SNPs, genetic association of a SNP with the disease, PPI network), as shown in this paper, such limitations can be overcome. We also would like to point out that our method is not intended to be used for tag SNPs which are associated with a specific phenotype.

CHAPTER 6

6 CONCLUSION

With the fast technological developments and continuous data production in the field of GWAS, more and more datasets are expected to be available in the near future. However, these studies are thought to be undermined in most cases. For GWAS analysis of complex diseases, novel disease-susceptibility genes and mechanisms can only be identified by looking beyond the tip of the iceberg (the most significant SNPs/genes). In this thesis, we described a novel methodology, PANOGA that performs network and pathway-oriented analysis of GWAS datasets via incorporating the functional information of the genotyped SNP. We tested PANOGA on rheumatoid arthritis, partial epilepsy, intracranial aneurysm and Behçet's disease datasets. In order to determine the biological significance of our results, we compared our findings with known disease related pathways in literature and with disease associated gene list obtained from OMIM, retrieved from literature using the NCBI PubMed module, or downloaded from Pharmaccogenomics Knowledge Base website. Our results show that incorporating SNP functional properties, protein-protein interaction networks, and pathway classification tools into GWAS can dissect leading molecular pathways, which cannot be picked up using traditional analyses. We hope that such developments of pathway and network-based approaches that also integrate prior biological knowledge for mining the associations of a group of SNPs, will take us one step closer to unravel the complex genetic structure of common diseases.

Using intracranial aneurysm datasets, we have described the advantages of a network and pathway-oriented analysis of GWAS data on different populations. Starting with two independent GWAS, which are conducted on two different populations, we have

shown that most of the affected pathways are shared between populations. But, in different populations, different SNP targeted genes are found to be affected in these commonly found pathways. In other words, same pathways can be targeted via independent genes in different populations. Even though there are not so many common disease predisposing SNPs and commonly targeted genes between two populations, the identification of 7 common pathways in the top 10 pathways showed the relevance of our pathway oriented approach. We have shown that while the shared pathways between the EU and the JP populations explain the general mechanisms of IA disease development; the pathways that are identified by population specific GWAS also need to be examined to gain a more comprehensive understanding of IA pathogenesis.

As a future work, we plan to fully automate our protocol and convert to a webserver such that it takes GWAS data as an input and generates disease specific pathway terms. Since the PANOGA protocol, which is developed throughout this thesis is quite modular, each main step of this method can be further improved. For example, currently available human PPI networks are far from being comprehensive. As the coverage and accuracy improves for these networks, the usage of such high quality PPI networks could dissect molecular pathways, which are not associated with the disease before. Another example might be the adaptation of pathway topology based approaches to the pathway identification step of PANOGA. Differently from traditional over-representation analysis approaches, these approaches incorporate the topological measures of the pathways while assigning a set of genes into pathways. Also, the incorporation of a functional enrichment approach that considers the shared genes between pathways (dependence between pathways), significance values of the genes, might further improve our methodology. In addition to the GWAS datasets, the tremendous boost in the “omics” technologies such as transcriptomics, proteomics and metabolomics also makes it possible to generate a global picture of system characteristics. Recently, miRNA expression datasets also became popular to understand the effect of the targeted modulation of gene regulation on complex disease mechanisms. Due to the modularity of PANOGA protocol, such different types of datasets can be easily incorporated to our system. Hence, the pathway level integration of all these different types of information might be a valuable approach to illuminate disease development mechanisms in a more comprehensive manner.

To conclude, as exemplified with GWAS datasets throughout this thesis, the affected pathways can be used as marker pathways for diseases to explain universal disease development mechanisms. Each population may search for disease causing factors targeting the genes within these affected pathways. Rather than the population, the same method can be extended to individuals to identify modifications occurring on the genes within these pathways. Hence, we can determine individual reasons for disease development which can be exploited for drug development and personalized therapeutical applications. To understand individual disease development mechanisms, these marker pathways can be scanned for an individual for alterations in the functions of the genes contained within. Thus, determining the disease-causing factors will provide a valuable insight for individualized therapy targets that would rectify the impact of these function altering factors.

REFERENCES

- Adeyemo, A. and Rotimi, C. (2010) Genetic Variants Associated with Complex Human Diseases Show Wide Variation across Multiple Populations, *Public Health Genomi*, **13**, 72-79.
- Adzhubei, I.A., *et al.* (2010) A method and server for predicting damaging missense mutations, *Nat Methods*, **7**, 248-249.
- Akiyama, K., *et al.* (2010) Genome-wide association study to identify genetic variants present in Japanese patients harboring intracranial aneurysms, *J Hum Genet*, **55**, 656-661.
- Albert, R. (2005) Scale-free networks in cell biology, *J Cell Sci*, **118**, 4947-4957.
- Alexa, A., Rahnenfuhrer, J. and Lengauer, T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure, *Bioinformatics*, **22**, 1600-1607.
- Altshuler, D., *et al.* (2000) The common PPAR gamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes, *Nat Genet*, **26**, 76-80.
- Amberger, J., Bocchini, C. and Hamosh, A. (2011) A New Face and New Challenges for Online Mendelian Inheritance in Man (OMIM (R)), *Hum Mutat*, **32**, 564-567.
- Aronica, E., *et al.* (2008) Gene expression profile analysis of epilepsy-associated gangliogliomas, *Neuroscience*, **151**, 272-292.
- Aronica, E., *et al.* (2008) Gene expression profile analysis of epilepsy-associated gangliogliomas, *Neuroscience*, **151**, 272-292.
- Aronica, E., *et al.* (2007) Complement activation in experimental and human temporal lobe epilepsy, *Neurobiol Dis*, **26**, 497-511.
- Askland, K., Read, C. and Moore, J. (2009) Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission, *Hum Genet*, **125**, 63-79.
- Bakir-Gungor, B. and Sezerman, O.U. (2011) A New Methodology to Associate SNPs with Human Diseases According to Their Pathway Related Context, *PLoS One*, **6**.
- Bakir-Gungor, B. and Sezerman, O.U. (2012) Identification of SNP Targeted Pathways From Genome-wide Association Study (GWAS) Data, Nature Protocol Exchange. DOI:10.1038/protex.2012.019.
- Bali, D., *et al.* (1999) Genetic analysis of multiplex rheumatoid arthritis families, *Genes Immun*, **1**, 28-36.

- Barabasi, A.L. (2009) Scale-Free Networks: A Decade and Beyond, *Science*, **325**, 412-413.
- Barabasi, A.L., Gulbahce, N. and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease, *Nature reviews. Genetics*, **12**, 56-68.
- Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization, *Nature reviews. Genetics*, **5**, 101-113.
- Baranzini, S.E., *et al.* (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis, *Hum Mol Genet*, **18**, 2078-2090.
- Barrenas, F., *et al.* (2009) Network properties of complex human disease genes identified through genome-wide association studies, *PLoS One*, **4**, e8090.
- Barton, A., *et al.* (2009) Identification of AF4/FMR2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further pan-autoimmune susceptibility genes, *Hum Mol Genet*, **18**, 2518-2522.
- Bauer, S., *et al.* (2008) Ontologizer 2.0 - a multifunctional tool for GO term enrichment analysis and data exploration, *Bioinformatics*, **24**, 1650-1651.
- Bebek, G., *et al.* (2012) Network biology methods integrating biological data for translational science, *Brief Bioinform.*
- Begovich, A.B., *et al.* (2004) A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis, *Am J Hum Genet*, **75**, 330-337.
- Behçet, H. (1937) Über rezidivierende aphthose durch ein virus verursachte Geschwure amMund, am Auge und an den Genitalien, *Dermatologische Wochenschrift*, **105**, 1152–1157.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing, *J Roy Stat Soc B Met*, **57**, 289-300.
- Bilguvar, K., *et al.* (2008) Susceptibility loci for intracranial aneurysm in European and Japanese populations, *Nat Genet*, **40**, 1472-1477.
- Bindea, G., *et al.* (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, *Bioinformatics*, **25**, 1091-1093.
- Bonetta, L. (2010) Protein-protein interactions: Interactome under construction, *Nature*, **468**, 851-854.
- Box, N.F., *et al.* (2001) Melanocortin-1 receptor genotype is a risk factor for basal and squamous cell carcinoma, *J Invest Dermatol*, **116**, 224-229.

- Boyle, E.I., *et al.* (2004) GO::TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics*, **20**, 3710-3715.
- Calabrese, R., *et al.* (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins, *Hum Mutat*, **30**, 1237-1244.
- Calzone, L., *et al.* (2008) A comprehensive modular map of molecular interactions in RB/E2F pathway, *Mol Syst Biol*, **4**, 173.
- Cantor, R.M., Lange, K. and Sinsheimer, J.S. (2010) Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application, *Am J Hum Genet*, **86**, 6-22.
- Chial, H. (2008) Rare genetic disorders: Learning about genetic disease through gene mapping, SNPs, and microarray data, *Nature Education* **1**(1).
- Chelala, C., Khan, A. and Lemoine, N.R. (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms, *Bioinformatics*, **25**, 655-661.
- Chen, L.S., *et al.* (2010) Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data, *Am J Hum Genet*, **86**, 860-871.
- Cho, C.H. (2011) Frontier of Epilepsy Research - mTOR signaling pathway, *Exp Mol Med*, **43**, 231-274.
- Choi, J.S., *et al.* (2003) Upregulation of gp130 and differential activation of STAT and p42/44 MAPK in the rat hippocampus following kainic acid-induced seizures, *Mol Brain Res*, **119**, 10-18.
- Cline, M.S., *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape, *Nat Protoc*, **2**, 2366-2382.
- Couzin, J. and Kaiser, J. (2007) Closing the net on common disease genes (vol 316, pg 820, 2007), *Science*, **317**, 320-320.
- Couzin, J. and Kaiser, J. (2007) Genome-wide association. Closing the net on common disease genes, *Science*, **316**, 820-822.
- Cowan, L.D. (2002) The epidemiology of the epilepsies in children, *Ment Retard Dev Disabil Res Rev*, **8**, 171-181.
- Dalby, N.O. and Mody, I. (2001) The process of epileptogenesis: a pathophysiological approach, *Curr Opin Neurol*, **14**, 187-192.
- de Kovel, C.G.F., *et al.* (2010) Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies, *Brain*, **133**, 23-32.

- Dennis, G., *et al.* (2003) DAVID: Database for annotation, visualization, and integrated discovery, *Genome Biol*, **4**.
- Dermitzakis, E.T. and Clark, A.G. (2009) Life After GWA Studies, *Science*, **326**, 239-240.
- Dhandapany, P.S., *et al.* (2009) A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia, *Nat Genet*, **41**, 187-191.
- Duan, D.M., *et al.* (2011) Label-free high-throughput microRNA expression profiling from total RNA, *Nucleic Acids Res*, **39**.
- Edwards, Y.J.K., *et al.* (2011) Identifying Consensus Disease Pathways in Parkinson's Disease Using an Integrative Systems Biology Approach, *PLoS One*, **6**.
- Elbers, C.C., *et al.* (2009) Using Genome-Wide Pathway Analysis to Unravel the Etiology of Complex Diseases, *Genet Epidemiol*, **33**, 419-431.
- Fei, Y.P., *et al.* (2009) Identification of novel genetic susceptibility loci for Behcet's disease using a genome-wide association study, *Arthritis Res Ther*, **11**.
- Feigin, V.L. and Findlay, M. (2006) Advances in subarachnoid hemorrhage, *Stroke*, **37**, 305-308.
- Feigin, V.L., *et al.* (2005) Risk factors for subarachnoid hemorrhage - An updated systematic review of epidemiological studies, *Stroke*, **36**, 2773-2780.
- Feldman, I., Rzhetsky, A. and Vitkup, D. (2008) Network properties of genes harboring inherited disease mutations, *Proc Natl Acad Sci U S A*, **105**, 4323-4328.
- Flicek, P., *et al.* (2010) Ensembl's 10th year, *Nucleic Acids Res*, **38**, D557-562.
- Franke, L., *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes, *Am J Hum Genet*, **78**, 1011-1025.
- Frayling, T.M. (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology, *Nat Rev Genet*, **8**, 657-662.
- Frazer, K.A., *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs, *Nature*, **449**, 851-861.
- Gaal, E.I., *et al.* (2012) Intracranial aneurysm risk locus 5q23.2 is associated with elevated systolic blood pressure, *Plos Genet*, **8**, e1002563.
- Gehlenborg, N., *et al.* (2010) Visualization of omics data for systems biology, *Nat Methods*, **7**, S56-S68.
- Gibson, G. (2010) Hints of hidden heritability in GWAS, *Nat Genet*, **42**, 558-560.

- Gieteling, E.W. and Rinkel, G.J. (2003) Characteristics of intracranial aneurysms and subarachnoid haemorrhage in patients with polycystic kidney disease, *J Neurol*, **250**, 418-423.
- Goldstein, D.B. (2007) Replicating genome-wide association studies - Response, *Science*, **318**, 391-391.
- Goldstein, D.B. and Hirschhorn, J.N. (2004) In genetic control of disease, does 'race' matter?, *Nat Genet*, **36**, 1243-1244.
- Gourfinkel-An, I., *et al.* (2001) Genetics of inherited human epilepsies, *Dialogues Clin Neurosci*, **3**, 47-57.
- Gregersen, P.K., *et al.* (2009) REL, encoding a member of the NF-kappa B family of transcription factors, is a newly defined risk locus for rheumatoid arthritis, *Nat Genet*, **41**, 820-U877.
- Guo, Y., *et al.* (2012) Two-stage genome-wide association study identifies variants in CAMSAP1L1 as susceptibility loci for epilepsy in Chinese, *Hum Mol Genet*, **21**, 1184-1189.
- Hardy, J. and Singleton, A. (2009) CURRENT CONCEPTS Genomewide Association Studies and Human Disease, *New Engl J Med*, **360**, 1759-1768.
- Hatemi, G. and Yazici, H. (2011) Behcet's syndrome and micro-organisms, *Best Pract Res Cl Rh*, **25**, 389-406.
- Hauser, W.A. (1994) The prevalence and incidence of convulsive disorders in children, *Epilepsia*, **35 Suppl 2**, S1-6.
- Hofer, A.M. (2005) Another dimension to calcium signaling: a look at extracellular calcium, *J Cell Sci*, **118**, 855-862.
- Holm, T.M., *et al.* (2011) Noncanonical TGF beta Signaling Contributes to Aortic Aneurysm Progression in Marfan Syndrome Mice, *Science*, **332**, 358-361.
- Holmans, P., *et al.* (2009) Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder, *Am J Hum Genet*, **85**, 13-24.
- Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res*, **37**, 1-13.
- Huang, D.W., *et al.* (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists, *Genome Biol*, **8**, -.
- Hugot, J.P., *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease, *Nature*, **411**, 599-603.

- Hunter, D.J. (2005) Gene-environment interactions in human diseases, *Nat Rev Genet*, **6**, 287-298.
- Ideker, T., *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics*, **18 Suppl 1**, S233-240.
- Jamali, S., *et al.* (2006) Large-scale expression study of human mesial temporal lobe epilepsy: evidence for dysregulation of the neurotransmission and complement systems in the entorhinal cortex, *Brain*, **129**, 625-641.
- Jeong, H., *et al.* (2000) The large-scale organization of metabolic networks, *Nature*, **407**, 651-654.
- Jimenez-Mateos, E.M., *et al.* (2008) Hippocampal transcriptome after status epilepticus in mice rendered seizure damage-tolerant by epileptic preconditioning features suppressed calcium and neuronal excitability pathways, *Neurobiol Dis*, **32**, 442-453.
- Joosten, L.A., *et al.* (2003) Toll-like receptor 2 pathway drives streptococcal cell wall-induced joint inflammation: critical role of myeloid differentiation factor 88, *J Immunol*, **171**, 6145-6153.
- Juvela, S. (2000) Risk factors for multiple intracranial aneurysms, *Stroke; a journal of cerebral circulation*, **31**, 392-397.
- Juvela, S., Poussa, K. and Porras, M. (2001) Factors affecting formation and growth of intracranial aneurysms: a long-term follow-up study, *Stroke; a journal of cerebral circulation*, **32**, 485-491.
- Kanehisa, M., *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets, *Nucleic Acids Res*, **40**, D109-D114.
- Karchin, R. (2009) Next generation tools for the annotation of human SNPs, *Brief Bioinform*, **10**, 35-52.
- Kasperaviciute, D., *et al.* (2010) Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study, *Brain*, **133**, 2136-2147.
- Kelder, T., *et al.* (2010) Finding the Right Questions: Exploratory Pathway Analysis to Enhance Biological Discovery in Large Datasets, *Plos Biol*, **8**.
- Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics*, **21**, 3587-3595.
- Khatri, P., Sirota, M. and Butte, A.J. (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges, *Plos Comput Biol*, **8**.
- Kholodenko, B.N. and Birtwistle, M.R. (2009) Four-dimensional dynamics of MAPK information-processing systems, *Wires Syst Biol Med*, **1**, 28-44.
- Kiberstis, P. and Roberts, L. (2002) It's not just the genes, *Science*, **296**, 685-685.

- Kjeldsen, M.J., *et al.* (2001) Genetic and environmental factors in epilepsy: a population-based study of 11900 Danish twin pairs, *Epilepsy Res*, **44**, 167-178.
- Knight, J.C. (2010) Understanding human genetic variation in the era of high-throughput sequencing, *Embo Rep*, **11**, 650-652.
- Krischek, B. and Inoue, I. (2006) The genetics of intracranial aneurysms, *J Hum Genet*, **51**, 587-594.
- Krischek, B., *et al.* (2008) Network-based gene expression analysis of intracranial aneurysm tissue reveals role of antigen presenting cells, *Neuroscience*, **154**, 1398-1407.
- Krischek, B. and Noue, I. (2006) The genetics of intracranial aneurysms, *J Hum Genet*, **51**, 587-594.
- Ku, C.S., *et al.* (2010) The discovery of human genetic variations and their use as disease markers: past, present and future, *J Hum Genet*, **55**, 403-415.
- Kuo, H.C., *et al.* (2011) ITPKC Single Nucleotide Polymorphism Associated with the Kawasaki Disease in a Taiwanese Population, *PLoS One*, **6**.
- Kurreeman, F.A.S., *et al.* (2007) A candidate gene approach identifies the TRAF1/C5 region as a risk factor for rheumatoid arthritis (vol 4, artn no e278, 2007), *Plos Med*, **4**, 2013-2013.
- Laaksamo, E., *et al.* (2008) Involvement of mitogen-activated protein kinase signaling in growth and rupture of human intracranial aneurysms, *Stroke*, **39**, 886-892.
- Lage, K., *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders, *Nat Biotechnol*, **25**, 309-316.
- Lam, H.Y.K., *et al.* (2012) Detecting and annotating genetic variations using the HugerSeq pipeline, *Nat Biotechnol*, **30**, 226-229.
- Lamond, A.I. (2002) Molecular biology of the cell, 4th edition, *Nature*, **417**, 383-383.
- Lander, E.S. and Schork, N.J. (1994) Genetic Dissection of Complex Traits, *Science*, **265**, 2037-2048.
- Lauren, H.B., *et al.* (2010) Transcriptome Analysis of the Hippocampal CA1 Pyramidal Cell Region after Kainic Acid-Induced Status Epilepticus in Juvenile Rats, *PLoS One*, **5**.
- Lee, P.H. and Shatkay, H. (2008) F-SNP: computationally predicted functional SNPs for disease association studies, *Nucleic Acids Res*, **36**, D820-D824.
- Lee, P.H. and Shatkay, H. (2009) An integrative scoring system for ranking SNPs by their potential deleterious effects, *Bioinformatics*, **25**, 1048-1055.

- Lesnick, T.G., *et al.* (2007) A genomic pathway approach to a complex disease: axon guidance and Parkinson disease, *PLoS Genet*, **3**, e98.
- Liacini, A., *et al.* (2003) Induction of matrix metalloproteinase-13 gene expression by TNF-alpha is mediated by MAP kinases, AP-1, and NF-kappaB transcription factors in articular chondrocytes, *Exp Cell Res*, **288**, 208-217.
- Limviphuvadh, V., *et al.* (2010) Is LGI2 the candidate gene for partial epilepsy with pericentral spikes?, *J Bioinform Comput Biol*, **8**, 117-127.
- Loots, G. and Ovcharenko, I. (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes, *Bioinformatics*, **23**, 122-124.
- Low, S.K., *et al.* (2012) Genome-wide association study for intracranial aneurysm in the Japanese population identifies three candidate susceptible loci and a functional genetic variant at EDNRA, *Hum Mol Genet*, **21**, 2102-2110.
- Lukasiuk, K., *et al.* (2006) Epileptogenesis-related genes revisited, *Prog Brain Res*, **158**, 223-241.
- Lukasiuk, K. and Pitkanen, A. (2004) Large-scale analysis of gene expression in epilepsy research: Is synthesis already possible?, *Neurochem Res*, **29**, 1169-1178.
- Luo, W., *et al.* (2009) GAGE: generally applicable gene set enrichment for pathway analysis, *BMC Bioinformatics*, **10**, 161.
- MacGregor, A.J., *et al.* (2000) Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins, *Arthritis Rheum*, **43**, 30-37.
- Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks, *Bioinformatics*, **21**, 3448-3449.
- Manolio, T.A. (2010) Genomewide Association Studies and Assessment of the Risk of Disease, *New Engl J Med*, **363**, 166-176.
- Marambaud, P., Dreses-Werringloer, U. and Vingtdeux, V. (2009) Calcium signaling in neurodegeneration, *Mol Neurodegener*, **4**.
- Martin, J.E., *et al.* (2010) Identification of the Oxidative Stress-Related Gene MSRA as a Rheumatoid Arthritis Susceptibility Locus by Genome-Wide Pathway Analysis, *Arthritis Rheum*, **62**, 3183-3190.
- McCarthy, M.I., *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges, *Nat Rev Genet*, **9**, 356-369.
- McCubrey, J.A., LaHair, M.M. and Franklin, R.A. (2006) Reactive oxygen species-induced activation of the MAP kinase signaling pathways, *Antioxid Redox Sign*, **8**, 1775-1789.

- Mefford, H.C., *et al.* (2010) Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies, *Plos Genet*, **6**.
- Meguro, A., *et al.* (2010) Genetics of Behcet disease inside and outside the MHC, *Ann Rheum Dis*, **69**, 747-754.
- Menon, R. and Farina, C. (2011) Shared molecular and functional frameworks among five complex human disorders: a comparative study on interactomes linked to susceptibility genes, *PLoS One*, **6**, e18660.
- Merikangas, K.R. and Risch, N. (2003) Genomic priorities and public health, *Science*, **302**, 599-601.
- Mi, H.Y., *et al.* (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium, *Nucleic Acids Res*, **38**, D204-D210.
- Migita, K., *et al.* (2011) CP690,550 inhibits oncostatin M-induced JAK/STAT signaling pathway in rheumatoid synoviocytes, *Arthritis Res Ther*, **13**.
- Mizuki, N., *et al.* (2010) Genome-wide association studies identify IL23R-IL12RB2 and IL10 as Behcet's disease susceptibility loci, *Nat Genet*, **42**, 703-U783.
- Morimoto, K., Fahnestock, M. and Racine, R.J. (2004) Kindling and status epilepticus models of epilepsy: rewiring the brain, *Prog Neurobiol*, **73**, 1-60.
- Myles, S., *et al.* (2008) Worldwide population differentiation at disease-associated SNPs, *Bmc Med Genomics*, **1**.
- Nahed, B.V., *et al.* (2007) Genetics of intracranial aneurysms, *Neurosurgery*, **60**, 213-225.
- Nahed, B.V., *et al.* (2007) Genetics of intracranial aneurysms, *Neurosurgery*, **60**, 213-225; discussion 225-216.
- Nam, D., *et al.* (2006) ADGO: analysis of differentially expressed gene sets using composite GO annotation, *Bioinformatics*, **22**, 2249-2253.
- Neiberger, H.L., *et al.* (2010) GSEA-SNP identifies genes associated with Johne's disease in cattle, *Mamm Genome*, **21**, 419-425.
- O'Shea, J.J. and Plenge, R. (2012) JAK and STAT Signaling Molecules in Immunoregulation and Immune-Mediated Disease, *Immunity*, **36**, 542-550.
- Okamoto, O.K., *et al.* (2010) Whole transcriptome analysis of the hippocampus: toward a molecular portrait of epileptogenesis, *Bmc Genomics*, **11**, 230.
- Ouattara, D.A., *et al.* (2012) Metabolomics-on-a-chip and metabolic flux analysis for label-free modeling of the internal metabolism of HepG2/C3A cells, *Mol Biosyst*.

- Palmer, C.N., *et al.* (2006) Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis, *Nat Genet*, **38**, 441-446.
- Pandolfo, M. (2011) Genetics of Epilepsy, *Semin Neurol*, **31**, 506-518.
- Patterson, S.D. and Aebersold, R.H. (2003) Proteomics: the first decade and beyond, *Nat Genet*, **33**, 311-323.
- Pattin, K.A. and Moore, J.H. (2008) Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases, *Hum Genet*, **124**, 19-29.
- Peng, G., *et al.* (2010) Gene and pathway-based second-wave analysis of genome-wide association studies, *Eur J Hum Genet*, **18**, 111-117.
- Pepin, M., *et al.* (2000) Clinical and genetic features of Ehlers-Danlos syndrome type IV, the vascular type, *N Engl J Med*, **342**, 673-680.
- Pera, J., *et al.* (2010) Gene Expression Profiles in Human Ruptured and Unruptured Intracranial Aneurysms What Is the Role of Inflammation?, *Stroke*, **41**, 224-231.
- Pitkanen, A. and Sutula, T.P. (2002) Is epilepsy a progressive disorder? Prospects for new therapeutic approaches in temporal-lobe epilepsy, *Lancet Neurol*, **1**, 173-181.
- Plenge, R.M., *et al.* (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis, *Nat Genet*, **39**, 1477-1482.
- Poduri, A. and Lowenstein, D. (2011) Epilepsy genetics - past, present, and future, *Curr Opin Genet Dev*, **21**, 325-332.
- Prasad, A.N., Prasad, C. and Stafstrom, C.E. (1999) Recent advances in the genetics of epilepsy: Insights from human and animal studies, *Epilepsia*, **40**, 1329-1352.
- Purcell, S., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am J Hum Genet*, **81**, 559-575.
- Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey, *Nucleic Acids Res*, **30**, 3894-3900.
- Ramos, H., Shannon, P. and Aebersold, R. (2008) The protein information and property explorer: an easy-to-use, rich-client web application for the management and functional analysis of proteomic data, *Bioinformatics*, **24**, 2110-2111.
- Raychaudhuri, S., *et al.* (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis, *Nat Genet*, **40**, 1216-1223.
- Raychaudhuri, S., *et al.* (2009) Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk, *Nat Genet*, **41**, 1313-U1376.

- Reich, N.C. (2009) STAT3 Revs Up the Powerhouse, *Sci Signal*, **2**.
- Reid, C.A., Berkovic, S.F. and Petrou, S. (2009) Mechanisms of human inherited epilepsies, *Prog Neurobiol*, **87**, 41-57.
- Remmers, E.F., *et al.* (2010) Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behcet's disease, *Nat Genet*, **42**, 698-U678.
- Remmers, E.F., *et al.* (2007) STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus, *New Engl J Med*, **357**, 977-986.
- Rinkel, G.J.E., *et al.* (1998) Prevalence and risk of rupture of intracranial aneurysms - A systematic review, *Stroke*, **29**, 251-256.
- Roeder, K., *et al.* (2006) Using linkage genome scans to improve power of association in genome scans, *Am J Hum Genet*, **78**, 243-252.
- Rogic, S. and Pavlidis, P. (2009) Meta-analysis of kindling-induced gene expression changes in the rat hippocampus, *Front Neurosci*, **3**, 53.
- Rosenberg, N.A., *et al.* (2010) Genome-wide association studies in diverse populations, *Nat Rev Genet*, **11**, 356-366.
- Rual, J.F., *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, **437**, 1173-1178.
- Ruigrok, Y.M. and Rinkel, G.J. (2010) From GWAS to the clinic: risk factors for intracranial aneurysms, *Genome Med*, **2**, 61.
- Ruigrok, Y.M. and Rinkel, G.J.E. (2008) Genetics of intracranial aneurysms, *Stroke*, **39**, 1049-1055.
- Ruigrok, Y.M., *et al.* (2008) Genes involved in the transforming growth factor beta signalling pathway and the risk of intracranial aneurysms, *J Neurol Neurosurg Ps*, **79**, 722-724.
- Saccone, S.F., *et al.* (2010) SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study, *Nucleic Acids Res*, **38**, W201-W209.
- Saccone, S.F., *et al.* (2008) Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence, *Bioinformatics*, **24**, 1805-1811.
- Sander, J.W. (2003) The epidemiology of epilepsy revisited, *Curr Opin Neurol*, **16**, 165-170.
- Schett, G., *et al.* (2000) Activation, differential localization, and regulation of the stress-activated protein kinases, extracellular signal-regulated kinase, c-JUN N-terminal

kinase, and p38 mitogen-activated protein kinase, in synovial tissue and cells in rheumatoid arthritis, *Arthritis Rheum*, **43**, 2501-2512.

Seal, R.L., *et al.* (2011) genenames.org: the HGNC resources in 2011, *Nucleic Acids Res*, **39**, D514-D519.

Shahrara, S., *et al.* (2007) Differential expression of the FAK family kinases in rheumatoid arthritis and osteoarthritis synovial tissues, *Arthritis Res Ther*, **9**.

Shannon, P., *et al.* (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome Res*, **13**, 2498-2504.

Sharma, A. (2012) Genome-Wide Expression Analysis in Epilepsy: A Synthetic Review, *Curr Top Med Chem*, **12**, 1008-1032.

Shriner, D., *et al.* (2007) Problems with genome-wide association studies, *Science*, **316**, 1840-1841.

Smid, M. and Dorssers, L.C.J. (2004) GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms, *Bioinformatics*, **20**, 2618-2625.

Soh, D., *et al.* (2010) Consistency, comprehensiveness, and compatibility of pathway databases, *BMC Bioinformatics*, **11**.

Steinlein, O.K. (2004) Genes and mutations in human idiopathic epilepsy, *Brain Dev*, **26**, 213-218.

Stelzl, U., *et al.* (2005) A human protein-protein interaction network: A resource for annotating the proteome, *Cell*, **122**, 957-968.

Stranger, B.E., Stahl, E.A. and Raj, T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics, *Genetics*, **187**, 367-383.

Suzuki, M., *et al.* (2000) The role of p38 mitogen-activated protein kinase in IL-6 and IL-8 production from the TNF-alpha- or IL-1 beta-stimulated rheumatoid synovial fibroblasts, *Febs Lett*, **465**, 23-27.

Tarca, A.L., *et al.* (2009) A novel signaling pathway impact analysis, *Bioinformatics*, **25**, 75-82.

Taylor, C.L., *et al.* (1995) Cerebral arterial aneurysm formation and rupture in 20,767 elderly patients: hypertension and other risk factors, *J Neurosurg*, **83**, 812-819.

Thomson, W., *et al.* (2007) Rheumatoid arthritis association at 6q23, *Nat Genet*, **39**, 1431-1433.

Torkamani, A., Topol, E.J. and Schork, N.J. (2008) Pathway analysis of seven common diseases assessed by genome-wide association, *Genomics*, **92**, 265-272.

- Tu, Z., *et al.* (2006) An integrative approach for causal gene identification and gene regulatory pathway inference, *Bioinformatics*, **22**, e489-496.
- Vallabhajosyula, R.R., *et al.* (2009) Identifying Hubs in Protein Interaction Networks, *Plos One*, **4**, -.
- Walsh, L.E. and McCandless, D. (2001) Inherited epilepsies, *Semin Pediatr Neurol*, **8**, 165-176.
- Wang, K., Li, M. and Bucan, M. (2007) Pathway-Based Approaches for Analysis of Genomewide Association Studies, *Am J Hum Genet*, **81**.
- Wang, K., Li, M.Y. and Hakonarson, H. (2010) Analysing biological pathways in genome-wide association studies, *Nat Rev Genet*, **11**, 843-854.
- Wang, Y.Y., *et al.* (2010) Global Expression Profiling in Epileptogenesis: Does It Add to the Confusion?, *Brain Pathol*, **20**, 1-16.
- Weeks, D.E. and Lathrop, G.M. (1995) Polygenic Disease - Methods for Mapping Complex Disease Traits, *Trends Genet*, **11**, 513-519.
- Weng, L.J., *et al.* (2011) SNP-based pathway enrichment analysis for genome-wide association studies, *BMC Bioinformatics*, **12**.
- Wilke, R.A., Mareedu, R.K. and Moore, J.H. (2008) The Pathway Less Traveled: Moving from Candidate Genes to Candidate Pathways in the Analysis of Genome-Wide Data from Large Scale Pharmacogenetic Association Studies, *Curr Pharmacogenomics Person Med*, **6**, 150-159.
- Williams, S.M., *et al.* (2007) Problems with genome-wide association studies, *Science*, **316**, 1840-1842.
- Wingender, E., *et al.* (2000) TRANSFAC: an integrated system for gene expression regulation, *Nucleic Acids Res*, **28**, 316-319.
- Wu, G., *et al.* (2010) A comprehensive molecular interaction map for rheumatoid arthritis, *Plos One*, **5**, e10137.
- Wu, J.M., *et al.* (2006) KOBAS server: a web-based platform for automated annotation and pathway identification, *Nucleic Acids Res*, **34**, W720-W724.
- Xu, Z.L. and Taylor, J.A. (2009) SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies, *Nucleic Acids Res*, **37**, W600-W605.
- Yaspan, B.L. and Veatch, O.J. (2011) Strategies for pathway analysis from GWAS data, *Curr Protoc Hum Genet*, **Chapter 1**, Unit1 20.
- Yasuno, K., *et al.* (2010) Genome-wide association study of intracranial aneurysm identifies three new risk loci, *Nat Genet*, **42**, 420-U469.

- Yokogami, K., *et al.* (2000) Serine phosphorylation and maximal activation of STAT3 during CNTF signaling is mediated by the rapamycin target mTOR, *Curr Biol*, **10**, 47-50.
- Zagulska-Szymczak, S., Filipkowski, R.K. and Kaczmarek, L. (2001) Kainate-induced genes in the hippocampus: lessons from expression patterns, *Neurochem Int*, **38**, 485-501.
- Zeeberg, B.R., *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data, *Genome Biol*, **4**.
- Zhang, B., Kirov, S. and Snoddy, J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts, *Nucleic Acids Res*, **33**, W741-W748.
- Zhang, K., *et al.* (2011) ICSNPathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework, *Nucleic Acids Res*, **39**, W437-443.
- Zhang, K.L., *et al.* (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 Res*, **38**, W90-W95.
- Zhang, L.C., *et al.* (2010) A towards-multidimensional screening approach to predict candidate genes of rheumatoid arthritis based on SNP, structural and functional annotations, *Bmc Med Genomics*, **3**, -.
- Zhernakova, A., *et al.* (2007) Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases, *Am J Hum Genet*, **81**, 1284-1288.
- Zhou, Y., *et al.* (2011) [Bioinformatic analysis of genes related to temporal epilepsy], *Nan Fang Yi Ke Da Xue Xue Bao*, **31**, 180-183.
- Zhou, Y.L., Martin, R.D. and Zhang, J.H. (2011) Advances in Experimental Subarachnoid Hemorrhage, *Acta Neurochir Suppl*, **110**, 15-21.
- Zinovyev, A., *et al.* (2008) BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks, *Bioinformatics*, **24**, 876-877.