

Network-Assisted Analyses of Chronic Inflammatory Diseases

Jörn Bethune
Christian-Albrechts-Universität zu Kiel

Dissertation zur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen
Fakultät der Christian-Albrechts-Universität zu Kiel

Selbstständigkeitserklärung

Hiermit erkläre ich, dass diese Abhandlung abgesehen von der Beratung durch meine Betreuer nach Inhalt und Form meine eigene Arbeit ist. Teile der Arbeit wurden bereits als wissenschaftlichen Papers veröffentlicht^{1;2}. Alle Inhalte waren noch nicht Teil eines Prüfungsverfahrens. Diese Arbeit wurde unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft angefertigt.

Kiel, den 30.6.2016, Jörn Bethune Jörn Bethune

Referent/in	<u>Prof. Dr. David Ellinghaus</u>
Koreferent/in	<u>Prof. Dr. Tal Dagan</u>
Tag der mündlichen Prüfung	<u>25.10.2016</u>
Zum Druck genehmigt	<u>Ja</u>
Dekan	<u>Prof. Dr. Oppelt</u>

“Für meine Eltern”

Acknowledgements

I am very grateful to my supervisors David Ellinghaus and Andre Franke for giving me the opportunity to do this PhD thesis and for their continuous support and advice.

Furthermore, I also like to thank Tal Dagan and her group members for keeping me around. This resulted in many pleasant, helpful and interesting exchanges, both on the scientific level and in non-scientific regards. I like to give special thanks to Robin Koch, Fernando Tria, Michael Bartz, Samer Kadib Alban, Carina Kreutzer, Sarah Kovarik, Lisa Stuckenschneider and Hendrik Breitschuh for being absolutely awesome and helpful office mates.

I also like to give special thanks to David Bogumil and Judith Ilhan who gave me many interesting insights and things to laugh about. I also like to thank Ovidiu Popa, Giddy Landan and Tal Dagan for providing me with valuable advice for working with networks. Besides, I also like to thank Tal for letting me teach her students Python programming.

And of course I like to thank the members of my own (official) group. I like to give special thanks to Sören Mucha, Matthias Hübenthal, Zhipei Gracie Du and Priyadarshini Kachroo without whom these three years would have been completely different. Thank you so much! I would also like to thank Elisa Rosati, Mareike Wendorff, Malte Rühlemann, Ingo Thomsen, Silke Szymczak, Lars Kraemer, Teide Boysen, Abdou ElSharawy and Daniela Esser for being really nice people to be around. All the best to you! Many thanks also to Michael Forster for the many wise words that he has shared with me. I like to express further thanks to ITIS featuring Santiago, Markus and Iacopo for really helpful and world-class IT support. And of course I want to thank the administrative staff for keeping everything running. Many thanks to Eike Zell and Christiane Wolf-Schwerin for their continuous help with everything related to paperwork and organisational issues.

We tend to stand on the shoulders of giants. Therefore I like to thank all maintainers and contributors of the Open Source software that was used throughout this thesis.

And finally and most importantly, I like to thank my parents, grandparents, siblings, cousins, aunts and uncles for their continuous support which made all of this possible.

I hope that I have thanked everyone involved. If I should have failed to thank you, please know that you still have my gratitude.

With great appreciation,

Jörn Bethune

1. Abbreviations

Table 1.: List of abbreviations

Abbreviation	Long name
ALL	All five diseases (AS, CD, PS, PSC, UC)
AS	ankylosing spondylitis
BED	Browser Extensible Data
CD	Crohn's disease
CHR	chromosome
CPDB	ConsensusPathDB
DAVID	The Database for Annotation, Visualization and Integrated Discovery (Enrichment tool)
DEPICT	Data-driven Expression Prioritized Integration for Complex Traits (software)
DNA	deoxyribonucleic acid
EGF	epidermal growth factor
EGFR	epidermal growth factor receptor
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait locus
ESR	estrogen receptor
et al.	and others (Latin: et alii)
FDR	false discovery rate
GO	Gene Ontology
GOTERM	Gene Ontology term
GWA	genome-wide association
GWAS	genome-wide association study/studies
HGNC	HUGO Gene Nomenclature Committee
HLA	human leukocyte antigen (synonymous with MHC in humans)

(continued on next page)

(continued from previous page)

HUMAN	In the case of UniProt IDs: A protein being part of the human proteome
IBD	inflammatory bowel disease
ID	identifier
IIBDGC	International IBD genetics consortium
IMSGC	International MS genetics consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCC	largest connected component (in a network)
LD	linkage disequilibrium
MHC	major histocompatibility complex (synonymous with HLA in humans)
MS	multiple sclerosis
NCBI	National Center for Biotechnology Information
OR	odds ratio
PINBPA	Protein interaction network-based pathway analysis (also the name of a Cytoscape plugin)
PIN	protein interaction network
PPI	protein-protein interaction(s)
PSC	primary sclerosing cholangitis
PS	psoriasis
RNA	ribonucleic acid
SBM	Subset-based meta-analysis
SNP	single nucleotide polymorphism
SNV	single nucleotide variation
TF	transcription factor
TFBS	transcription factor binding site

(continued on next page)

(continued from previous page)

UC	ulcerative colitis
UniProt	Universal Protein Resource (uniprot.org)
VEGAS	Versatile Gene-Based Test for Genome-wide Association Studies (software)

Contents

1. Abbreviations	v
1. Introduction	1
1.1. Genome-wide association studies (GWAS)	2
1.2. Inflammation	5
1.3. Complex Diseases	6
1.4. Networks as a Response to Complex Diseases	9
1.5. Introduction to protein-protein interaction networks	12
1.5.1. Characteristics of protein-protein interaction networks	13
1.5.2. Centralities (importance of nodes in networks)	14
1.5.3. Pathways and Modules	15
1.6. Enrichment of terms and tissues based on gene lists	17
1.7. Objectives of this Thesis	17
2. Material & Methods	19
2.1. SNPs	19
2.2. Protein-Protein Interaction Reference Databases	22
2.3. Mapping of SNPs to Genes	23
2.4. ID mapping	27
2.5. Network handling	28
2.5.1. Cytoscape	28
2.5.2. Effect directions of SNPs on the network level	31
2.5.3. Network randomization	31
2.6. Term- and Tissue-Enrichment	33
3. Results	35
3.1. Common Network Modules Across Inflammatory Diseases	36
3.1.1. SNP to Gene Mapping	36
3.1.2. Disease networks	38

3.1.3.	Protein-interaction-network-based pathway analysis (PINBPA)	40
3.1.4.	Gene-wise Overlap Between Diseases in Networks	49
3.2.	Geneset-based networks	55
3.2.1.	Geneset enrichment	56
3.2.2.	Tissue enrichment	62
3.2.3.	Geneset-based networks	68
3.2.4.	Enrichment Analyses	71
3.2.5.	Disease-specific subnetworks	75
3.2.6.	Linker nodes	78
3.3.	Effect directions of Single Nucleotide Polymorphisms	83
3.3.1.	Mapping of SNPs to genes/DNA binding elements	87
3.3.2.	Relationships between transcriptional regulators and SNPs	88
3.3.3.	Purely Protective Binding Factors	89
3.3.4.	Protective genes in detail	94
3.3.5.	Comparison with geneset-based networks	97
3.3.6.	Network analysis	98
4.	Discussion	105
4.1.	Biology of Inflammatory Diseases	105
4.1.1.	Networks of Inflammatory Diseases	107
4.1.2.	Representation of Biology in Networks	109
4.2.	Mapping of SNPs to genes	112
4.3.	Construction of Networks	115
4.3.1.	Finding subnetworks and submodules	116
4.4.	Enrichment Findings	118
4.5.	Methodological Considerations	121
4.5.1.	Automation, Reproducibility and Repeatability	122
4.6.	Future work	123
5.	Conclusions	126
A.	Zusammenfassung	128
B.	Summary	130
C.	Supplementary Chapters	131
C.1.	A few notes on the Microbiome	131
C.2.	Modifications to VEGAS	132

List of figures	134
List of tables	138
Bibliography	140

Chapter 1.

Introduction

“Look at me: still talking when there’s science to do! When I look out there, it makes me glad I’m not you.”

— Mad robot GLaDOS (Lyrics)

Human health is dependent on many factors. On the highest level, these factors can be divided into inborn and environmental factors. Inborn factors include genetics and epigenetics while environmental factors include (but are not limited to) exposure to other organisms as well as to chemicals. These other organisms may be viruses, microorganisms or higher organisms. Chemicals include food and non-food products.

Another environmental factor is radiation. But it is usually not a highly influential factor, even though sunlight is relevant for vitamin D production³ and skin cancer⁴.

There can be an interplay between these factors. Pathogens and symbionts can coevolve with their hosts and take advantage of the genetic architecture that is common in a human population⁵.

Understanding the etiology of a disease can therefore require an understanding of the inborn and environmental factors that drive the disease. However, this thesis only focusses on the genetic aspects of inflammatory diseases. Environmental factors will only be briefly discussed.

1.1. Genome-wide association studies (GWAS)

In order to understand the genetic factors that drive disease, it has to be determined which parts of the genome are responsible for a certain phenotype. The human genome is about 3.2 billion base pairs long⁶ and unfortunately it is far from trivial to understand what these base pairs do.

Almost every human has a unique genomic sequence⁶. But this genomic sequence is still very similar to every other human's genomic sequence. Yet the variations in the human genome can lead to major differences in the development and health of individuals⁷.

The smallest variations in the genome are called *single nucleotide polymorphisms* (SNPs). They describe the variability of a single nucleotide at a specific position in the genome under the condition that each of the observed nucleotides must be present in at least 1% of the population⁸.

In order to assess the relevance of a SNP for a disease, statistical methods can be applied to show that the presence or the absence of a SNP correlates well with the presence or the absence of a disease. A modern statistical method for assessing the relevance of a SNP for several diseases at once is the *Subset-based Meta-analysis* (SBM)⁹.

It should be noted that there is positive and negative association. If the variant that is less frequent in the population is negatively correlated with the presence of a disease, the variant is considered to be protective against that disease. In a similar fashion, if the less frequent variant is positively correlated, it is regarded as risk-increasing for developing the disease (see also Figure 1.1).

When trying to understand which SNPs influence a disease, it usually makes sense to test all known SNPs in the genome for association. Such an endeavor is called a *genome-wide association study* (GWAS). It is commonly said that such a GWAS is *hypothesis-free*¹⁰ because all SNPs in the genome are regarded as potentially equally qualified to contribute to the disease. However, the term *hypothesis-free* is actually not correct in the statistical sense of the word, because hypotheses are statements that are being tested with statistical tests. In a typical genome-wide association study there are usually about one million (or more) statistical hypotheses: For every SNP it is being tested if it is not associated with a disease (Null hypothesis)¹¹.

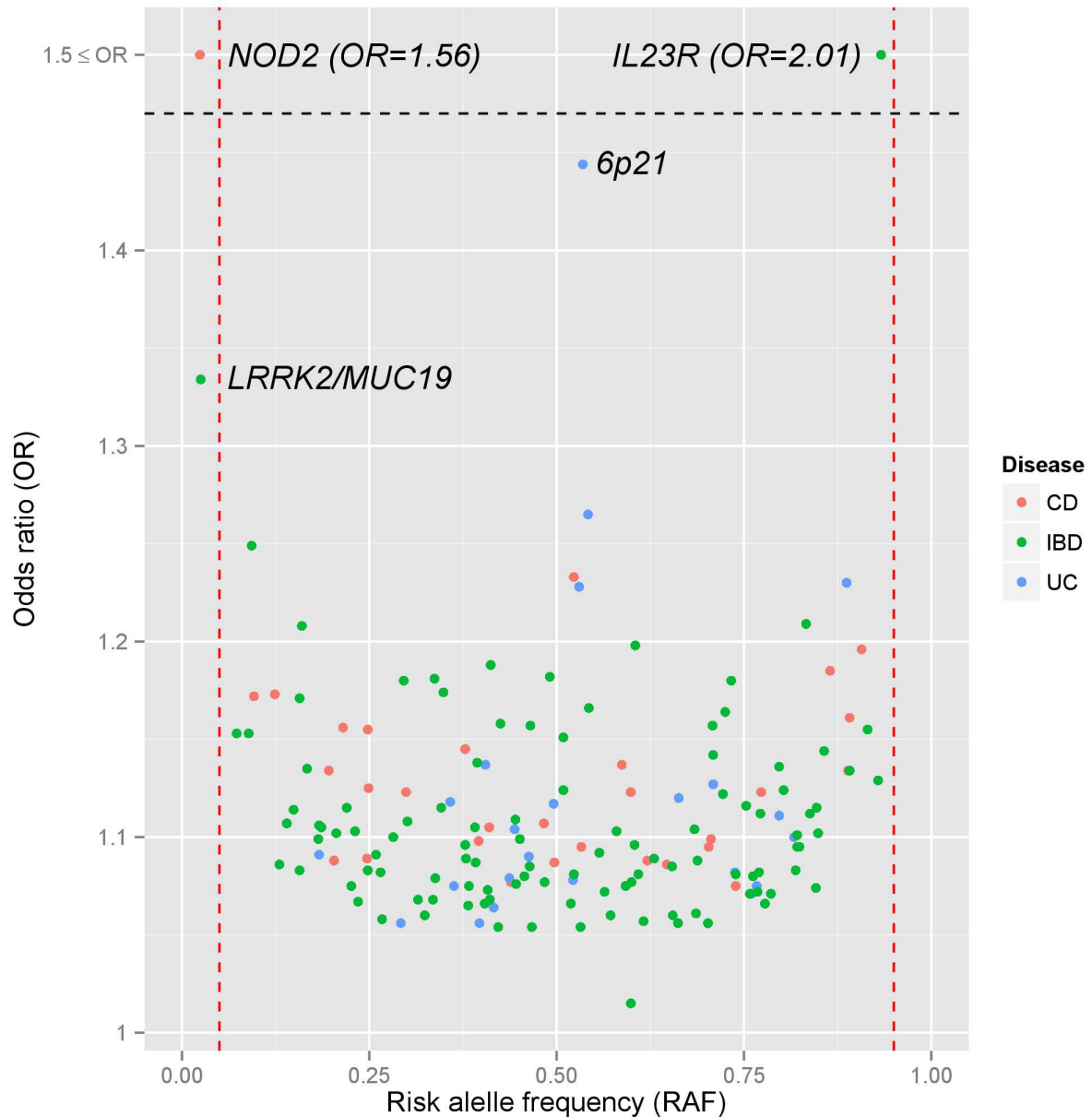


Figure 1.1.: Risk alleles of inflammatory bowel disease (IBD) and its subtypes Crohn's disease (CD) and ulcerative colitis (UC) based on the 163 SNPs from Jostins et al.⁷. The frequency of the risk alleles in the population is shown on the x-axis. The corresponding odds ratio for every risk allele is shown on the y-axis with most alleles being below an OR of 1.3 which reflects on their modest effect sizes. This image was first published by Ellinghaus et al¹ as Figure 1.

When doing statistics, sample size is a major concern. More genotypes of people can lead to stronger statistical evidence of the influence of a SNP on a disease. However, sequencing complete genomes is still relatively expensive and sequencing tens of thousands of people is even more so. To work around this problem, no complete sequencing is done and instead only a subset of SNPs are genotyped. This subset of SNPs consists of so-called marker SNPs. They allow us to infer the remaining SNPs by a process called *imputation*¹².

The idea behind imputation is that SNPs are coupled due to *linkage disequilibrium* (LD). The cause of LD is based on the principle that genetic information in humans is transferred on chromosomes. A chromosome is transferred as one unit of inheritance to the next generation together with all SNPs on it, hence SNPs are coupled¹¹. However, this is not the complete picture. During meiosis the homologous chromosomes pair up and exchange parts of their DNA sequence. This process is called *crossing over*. Through crossing over some SNPs can be transferred from one chromosome to another and the coupling is broken. The greater the distance between two SNPs, the more likely it is that there is a crossing over event between them. It is therefore said that two SNPs, which are close to each other, are in high LD, because it is unlikely that they will be separated by a crossing over event. High LD allows the assumption that if one SNP is being observed, the other will also be present and thus does not need to be genotyped¹¹.

As stated previously, a genome-wide association study involves about a million (or more) statistical tests. This leads to a *multiple-testing problem*: When performing many statistical tests of the same type, some of these tests will incorrectly reject the null hypothesis just by chance. It is therefore important to make this incorrect rejection less likely. A simple approach is to reduce the significance threshold for statistical tests so that the p-value has to be even lower than the typical 5% threshold to reject the null hypothesis. It is common to use the Bonferroni-correction¹² to either reduce the significance threshold or to raise the p-value - which, in the end, leads to the same analytical result. Because a GWAS has about a million statistical tests, the p-value is either multiplied by one million or the significance threshold is divided by the factor one million and thus becomes

$$0.05 \times \frac{1}{1000000} = 5 \times 10^{-8}$$

which is commonly referred to as the *genome-wide significance* threshold^{13;14}.

An important principle in science is "correlation does not imply causation". If a SNP is correlated with a disease, that SNP might still be irrelevant for the disease, because it is in high linkage disequilibrium with the actual causal SNP, i.e. the SNP is commonly inherited together with the causal SNP¹¹. It is therefore necessary to not only consider the SNPs with the lowest disease-association p-value (lead SNPs) but also all SNPs that are in high LD with this SNP (LD SNPs) when investigating the causes of a disease¹.

Another common problem with genome-wide association studies are the estimation of effect sizes. A p-value does provide an answer to the question whether or not a SNP (or any of its LD SNPS) does have an effect on a disease. But the actual impact of that SNP is hard to determine¹⁵. For the diseases that will be presented in section 1.3 many SNPs have very low effect sizes, i.e. their presence does not have a huge impact on the development of a disease. It is usually the combination of several SNPs and environmental factors that causes the disease to actually develop¹.

Genome-wide association studies are usually huge endeavours which nowadays have complete research consortia behind them like the *International IBD genetics consortium* (IIBDGC) and the *International Multiple Sclerosis Genetics Consortium*¹⁶ (IMSGC). Furthermore, there are archival websites like the GWAS catalogue¹⁷ that provide access to the results of past GWA studies and also make it possible to conduct meta-GWAS studies⁷ with even greater sample sizes than before to increase statistical power.

Finally there is the challenge of interpreting the results of a GWA study. SNPs that lie in coding regions with non-synonymous base-exchanges are comparatively easy to understand. But many times SNPs lie in intronic regions or in intergenic regions. In the latter case it is especially difficult to assess which genes are affected by a SNP.

In any case, genome-wide association studies provide a basis for further analysis to understand phenotypes and especially diseases.

1.2. Inflammation

Humans and other organisms have immune systems to defend themselves against pathogens. The immune system in humans consists of many different cells that all contribute to the defense against invaders¹⁸.

However, the immune system has a difficult task to perform: It should not attack the tissue of the individual that it is supposed to protect¹⁸. Furthermore, humans live in symbiotic relationships with microorganisms¹⁹. The immune system should not attack or react aggressively towards these symbionts which live in specific areas of the human body.

There is a class of diseases that are called chronic inflammatory diseases. In these diseases the immune system overreacts and induces inflammation in disease-specific parts of the body. These diseases are chronic and patients suffer from lifelong impairments.

Inflammation is a process that protects humans from tissue-invading microbes¹⁸. It usually occurs after wounding when microbes enter the wound and immune cells encounter these microbes. Inflammation leads to accumulation of leukocytes, plasma proteins and fluid derived from blood at the site of inflammation¹⁸ to fight the invaders. However, inflammation also comes with side-effects that are problematic for the inflamed tissue: Swelling, pain, redness, heat and loss of function²⁰. Especially when inflammation becomes chronic, tissue damage occurs¹⁸.

Inflammation can be regulated by different different immune cell types. Macrophages can be pro- and anti-inflammatory, depending on the conditions²¹. Monocytes are usually pro-inflammatory²¹. Regulatory T-cells are major regulators that prevent inadequate immune responses²². Regulatory B-cells are a small subpopulation of B-cells that are involved in the downregulation of inflammatory processes²².

Macrophages and Monocytes are part of the innate immune system while the regulatory T- and B-cells are part of the adaptive immune system¹⁸. After resolution of inflammation both types of immune cells aggregate in the tissue and reside there for weeks, so that they can directly attack pathogens and cause inflammation if the same tissue is being invaded again²⁰.

1.3. Complex Diseases

The previous section explained the usefulness but also the harmfulness of inflammation. It is therefore important to understand under which conditions inflammation gets out of control. Based on this understanding it might be possible to develop cures for various diseases.

Understanding inflammatory diseases has turned out to be difficult. The primary reason is that many factors play together and our current understanding of these interplays is very limited on the biological level. In fact, even though we know many genetic factors (SNPs) that correlate well with disease, we do not know what the actual biological consequences of these factors are because most of them lie outside of coding regions. However, many SNPs lie within potential *expression quantitative trait loci* (eQTL)s that may regulate the expression of specific genes via cis- or trans-effects²³.

There are five diseases that are being analysed in this thesis: Crohn's disease (CD), ankylosing spondylitis (AS), psoriasis (PS), primary sclerosing cholangitis(PSC) and ulcerative colitis (UC). Each of these diseases causes chronic inflammation in patients and the etiology of these diseases is still unknown. Table 1.1 shows the affected organs of these diseases. The diseases CD and UC are the most frequent subtypes of inflammatory bowel disease (IBD) which cause inflammation in gastrointestinal organs. Beyond the inflammation itself, the symptoms experienced by the patients include the following:

Ankylosing Spondylitis: Limited motion of the lumbar spine, persistent lower-back pain, limited chest expansion²⁴

Crohn's disease: Can vary by subtype. Possible symptoms include rectal bleeding, diarrhea, abdominal cramping pain (associated with iron deficiency), fatigue, weight loss and fever²⁵

Psoriasis: Psoriasis means itching condition in Greek. The patients experience itchy skin with red scaly plaques¹⁸.

Primary Sclerosing Cholangitis: Symptoms can include fatigue, abdominal pain, jaundice and fever²⁶

Ulcerative colitis: Rectal bleeding, diarrhea, abdominal cramping pain²⁵

It has been shown that these diseases are very similar on the genetic level despite having mostly different affected organs¹ which is the reason why they are being investigated together in this thesis. But it has also been observed that there is a high comorbidity between these diseases²⁶: Patients having one disease have a high chance of developing another within a time frame of five years².

In addition to the known genetic factors there are also the environmental factors which play a role¹. It has been statistically shown that smoking is risk-inducing for Crohn's

Disease	Primary affected (inflamed) organs
ankylosing spondylitis	Axial skeleton (mainly spine)
Crohn's disease	Any part of the gastrointestinal tract from mouth to anus
psoriasis	Skin
primary sclerosing cholangitis	Bile ducts
ulcerative colitis	Colon and rectum

Table 1.1.: Inflamed Organs in different inflammatory diseases

disease while at the same time it has a protective effect against ulcerative colitis²⁷. It is hypothesized that smoking impairs autophagy²⁸ and autophagy appears to be an important process in Crohn's disease²⁹.

There are probably more factors at play here that are related to the lifestyle of advanced "western" human societies. Determining these factors is still a future challenge. In fact, the incidence rate of these diseases has risen in recent decades while at the same time the incidence rates of infectious diseases have fallen due to better hygiene³⁰.

These diseases are called complex diseases because a great number of known genetic factors influence the probability of developing such a disease in addition to environmental factors. Because these diseases have an inflammatory component, they are sometimes categorized as autoimmune diseases³¹. It is not uncommon for autoimmune diseases to be complex polygenetic traits with a major environmental component¹⁸.

There are also less complex forms of some of these diseases. There are monogenic forms of IBD³² where a single gene or in some cases a few genes suffice to cause the disease. These types of non-complex diseases are not investigated in this thesis. Furthermore, these less complex forms are much rarer than their complex counterparts³². Most variants in common complex diseases have only very low effect sizes and these complex diseases usually develop much later in life than their monogenic forms³².

In addition to genetic and environmental factors it has also been shown that epigenetic factors play a role in inflammatory diseases³³ making the diseases even more complex.

Neurological aspects also seem to play a role in the regulation of inflammation. The vagus nerve has been shown to be involved in the regulation of inflammation and might be a promising target for therapy³⁴. However, it is unclear if it is involved in cause of such diseases. But the relevance of the vagus nerve also reflects well on the concept of

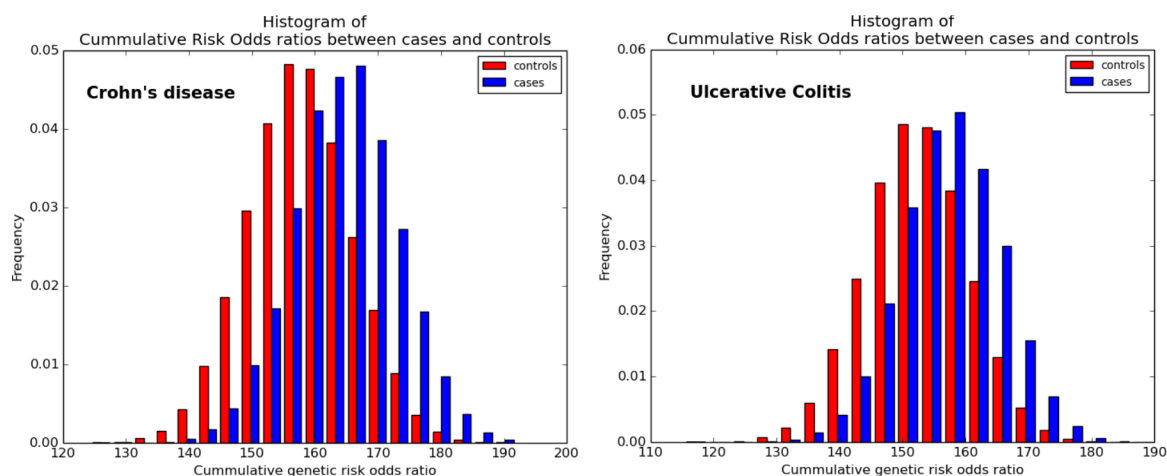


Figure 1.2.: Cumulative genetic risk in IBD patients (blue) and healthy individuals (red) for Crohn's disease (left) and ulcerative colitis (right). A high genetic risk does not always lead to the development of the diseases. And a low genetic risk is no guarantee for health. Further non-genetic factors have to be taken into account to explain the development or lack of development of diseases.

the gut-brain axis which describes the relationship between the great number of neurons in the gastrointestinal system that are connected and interact directly or indirectly with the neurons in the brain³⁵.

With so many factors playing a role in these diseases, genetics alone does not and cannot explain the development of the diseases completely³⁶ (see also Figure 1.2). But genetics might still provide a starting point to address the complexity of these diseases.

It has been observed that many SNPs are relevant to several diseases at once². If an individual has one SNP, the risk for multiple diseases can increase or decrease. Figure 1.3 gives an overview over shared loci of the five diseases studied in this thesis. These five diseases have been selected because of their shared genetics factors and have been investigated in the context of the *cross-disease project*².

1.4. Networks as a Response to Complex Diseases

There are many genetic factors that have been found through GWA studies². These factors are spread all across the genome yet they all contribute to the same phenotypes. With so many factors being relevant, one has to wonder how it all fits together. The explanation to this phenomenon is likely complex because there are many types of interactions that occur in biological systems and that are common in health and disease³⁷.

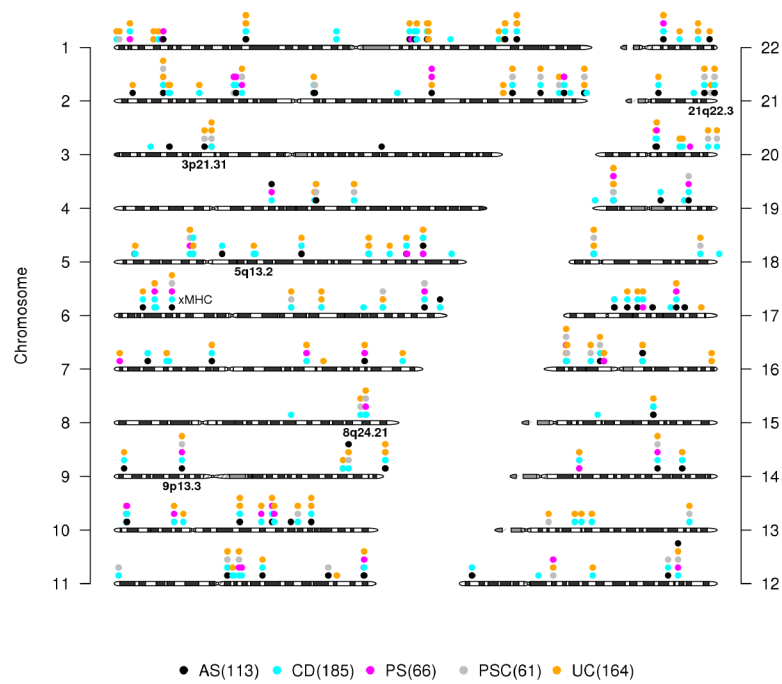


Figure 1.3.: Shared loci among the five diseases studied in this thesis. Image taken from Ellinghaus et al², supplementary figure 2

Human health is maintained through many control mechanisms in the body. For some diseases to develop, several checkpoints have to break down²³ which fits well with the observation that many loci appear to play a role in these diseases. But it is probably a specific combination of variants in the genome that drive the development of diseases¹.

The central actors in the cell are proteins. They physically interact with each other to form functional modules or they chemically modify each other. Proteins are involved in many metabolic, communicative and structural processes⁶. The interactions between proteins are an important aspect of their function. Most proteins do not act in isolation but require other proteins to fulfill their purposes³⁸. This alone accounts for a lot of complexity in biological systems because it is difficult to analyse proteins in isolation.

Proteins do not only interact with each other but also with many other molecules, most notably DNA and histones. They can act as transcription factors for genes and thereby establish a relationship with these genes on a conceptual level⁶. Proteins are also responsible for chemical modifications of DNA and histones. Specifically, these proteins add or remove methyl or acetyl groups from DNA or histones. These modifications are maintained upon DNA duplication and can also be passed on to future generations⁶.

Another important class of molecules in the cell are RNA molecules. These include mRNAs which serve as a template for the translation of proteins but also non-coding RNAs like microRNAs that interfere with mRNAs and thereby prevent translation³⁹. Thus microRNAs can interact with genes or proteins by downregulating their mRNA. There are also *long non-coding RNAs* (lncRNAs) which are defined as RNAs that are at least 200bp long. These perform various functions in the cell⁴⁰, including neutralization of microRNAs⁴⁰.

Finally there is DNA which harbours the template sequences for RNAs (including precursors of mRNA). The previously mentioned genetic variants are found in the DNA. Every variant in the DNA may influence the sequence of a RNA or protein molecule. In addition any variant in the DNA may influence the binding of proteins to the DNA. Common classes of DNA-binding proteins are transcription factors which bind to enhancers, silencer, insulator or promoter regions of a gene⁶. Furthermore, histones also bind to DNA but this binding has to be sequence-agnostic in order to fully condense DNA into chromosomes⁶.

To summarize: Proteins, various RNAs and DNA can interact with each other in different fashions. These interactions are required for normal biological function and they might be disturbed in disease.

There is a principle called *guilt-by-association* which means that if a given protein is known to interact with another protein, then it is likely that both proteins participate in the same or related cellular functions⁴¹. The same principle can also be applied to disease-associated genes/proteins: If two proteins interact and one of them is known to be associated with a disease, then the other protein is also a good candidate for being a disease-associated protein⁴².

To handle this complexity on the scientific level it makes sense to gather all known interactions into a virtual network that represents the overall mechanisms that are relevant to a disease. This is not a new idea and many projects have used networks to represent collections of biological interactions^{43;44;45;46}.

A network bundles all known interactions together and has the potential to reveal how the different genetic factors act together to cause disease. It may be used to explain how interaction chains and interaction subnetworks are disturbed in these complex diseases and they may be used to select therapeutic targets.

1.5. Introduction to protein-protein interaction networks

A network is a very versatile formalism. Networks can represent relationships between entities. The type of entities and the type of relationships can be many different things and networks have been applied to various research areas^{47;46}. But for our purposes we will focus solely on *protein-protein interaction* (PPI) networks. Protein-Protein interaction networks are not the only networks that are relevant in molecular biology but they are the most widely used form of networks in the research area of the inflammatory diseases discussed in this thesis⁴⁸.

A network consists of nodes and edges. A node represents an entity. An edge is a connection between nodes and represents a relationship between entities. In a protein-protein interaction network every node is a protein and an edge between two nodes signifies that there is *evidence* that these two proteins physically interact with each other. This evidence is usually experimental evidence but there are also predictive approaches to derive unknown protein-protein interactions from text mining⁴⁹ or other data sources.

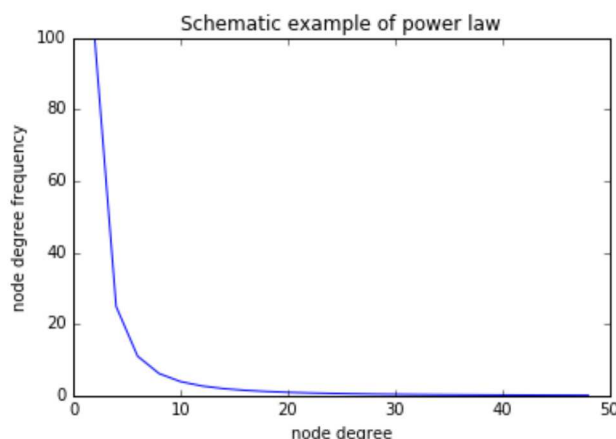


Figure 1.4.: Exemplary power law of node degrees. Many natural networks have many nodes of low degree and a few nodes of really high degree. These networks are called scale-free networks⁴⁶.

In protein-protein interaction networks it might be unknown what a connection actually is for: If someone isolates two proteins that are bound to each other, we only know that these proteins can interact with each other. We might not know what the biological consequences are of this interaction and further research is required to find out.

1.5.1. Characteristics of protein-protein interaction networks

The number of edges attached to a node is called the degree of a node. Nodes that have a high degree are considered to be important nodes because they are directly connected to many other nodes which often implies that a lot of other nodes depend on this single node. Such a node is called a "hub node". It has been observed by Goh et al. that "hub nodes" tend not to be disease-associated nodes⁴⁴ which is probably due to their central role in these networks and a mutation in these genes will likely disturb a great number of pathways and reduce the fitness of the individual to a great extent.

Protein-Protein interaction networks tend to be *scale-free*: The distribution of node degrees follows power laws⁴⁶ which means that the number of nodes of high degree (hubs) is very low compared to the rest of the network. Figure 1.4 illustrates this relationship.

Another common characteristic of protein-protein interaction networks is that they tend to consist of one large connected component and several components of considerably smaller size⁴⁶. Protein-Protein networks are therefore usually well connected and it is possible to reach most other proteins from any protein in the network. Furthermore,

Protein-Protein networks exhibit the *small world phenomenon*: The distance between any two nodes in a connected component is on average 6.8 steps apart⁴⁶.

1.5.2. Centralities (importance of nodes in networks)

A common concept to determine the importance of nodes in the network is to determine their *centrality*⁴⁶. There are different types of centrality. To define these centrality measures more formally, let $G = (V, E)$ be a graph consisting of nodes $v \in V$ and edges $e \in E$ connecting these nodes. All networks presented in this thesis are graphs.

Let A be the adjacency matrix which is defined as

$$A_{ij} = \begin{cases} 1 & v_i \text{ and } v_j \text{ are connected by an edge} \\ 0 & \text{otherwise} \end{cases} \quad (1.1)$$

The simplest centrality measure is degree centrality which corresponds to the degree of a node:

$$\text{degree}(v_i) = \sum_{j=1}^n A_{ij} \quad (1.2)$$

All hub nodes have a high degree centrality.

Other centrality measures are betweenness centrality, closeness centrality and eigenvector centrality. A node has a high betweenness centrality when many shortest paths between any pairs of nodes in the network goes through the node itself:

$$\text{betweenness}(v_i) = \sum_{s,t \in V} v_i(s, t) \quad (1.3)$$

Where

$$v_i(s, t) = \left\{ \begin{array}{ll} 1 & \text{if shortest path from } v_s \text{ to } t \text{ goes through } v_i \\ 0 & \text{otherwise} \end{array} \right\}$$

A node has a high closeness centrality if it is possible to reach most other nodes in the network with only few steps:

$$\text{closeness}(v_i) = n \cdot \left(\sum_{j \in V} \gamma_{ij} \right)^{-1} \quad (1.4)$$

Where γ_{ij} is the length of the shortest path between node v_i and node v_j

The Eigenvector centrality defines the importance of a node based on the importance of the nodes connected to that node:

$$\text{eigen}(v_i) = \frac{1}{\rho} \sum_{j \in \Gamma(i)} \text{eigen}(v_j) = \frac{1}{\rho} \sum_{j=1}^n A_{ij} \cdot \text{eigen}(v_j) \quad (1.5)$$

Where ρ is the largest eigenvalue of A and $\Gamma(i)$ is the set of neighbours of node v_i .

There are also other centrality measures^{50,46}. But they have not been used in this thesis.

1.5.3. Pathways and Modules

It should be noted that currently we do not have complete knowledge of all interactions in cellular systems. Therefore any networks that we construct will likely be incomplete⁵¹. In addition to this incompleteness there are many false-positive interactions in the databases⁴⁸ which increase the complexity of the *in-silico* networks. Another problem is the lack of temporal and spatial resolution of these interactions so that it is often not known under which conditions an interaction actually takes place⁴⁸.

When we construct networks the objective is rarely to get a complete picture of a biological system, but rather an approximation that is good enough to get insight into key aspects of biology.

But this incompleteness does not mean that there are only few interactions. When constructing networks in an automated fashion the resulting networks tend to be very big and complex. While some of the complexity does indeed reflect biological reality, it is advisable to reduce the networks so that they ideally focus on a single functional concept that can be analysed.

One such functional concept is a *pathway*. A set of different pathway analysis methods exist. They take different views on what constitutes a pathway³⁸ but in the end they either consider a pathway to be a specific set of nodes and/or a specific set of interactions that may have a known functional description. It has been reported that the results obtained from pathway-based analyses differ greatly between methods^{52;53}.

Another similar concept is that of a submodule. Modules in networks are subnetworks that consist of nodes that have been selected as a group by an algorithm⁴⁸. One example of an algorithm is implemented by the jActiveModules Cytoscape plugin which tries to find groups of connected nodes that have highly significant gene-wise association p-values⁵⁴.

Jia et al. describe five different dimensions of finding modules (subnetworks) in a network⁴⁸:

- Binary categories (disease-associated or not disease-associated) *versus* quantitative (disease association score)
- Network topology oriented *versus* node weight oriented ("genetically oriented"), *or* a combination of both
- Global *versus* local search in the network space (limiting the maximum number of steps in a network algorithm)
- Prioritization of single genes *versus* finding a combination of several genes for further analysis
- Direct *versus* indirect interactions. Indirect interactions take into account the number of paths between two non-connected nodes. More paths between two nodes might indicate a strong relationship.

Depending on the available data some options may not be available but overall there are many possible choices.

1.6. Enrichment of terms and tissues based on gene lists

A common result of bioinformatics analyses are lists of genes. These genes might share common characteristics like being associated with the same disease, being inside a network module or being differentially expressed.

Given that there are about 20.000 genes in the human genome⁵⁵, the genes on the list may be unknown to the researcher. Enrichment tools provide an automated approach to give an overview about what the genes on a list have in common and what kind of biological phenomena they are involved in.

Enrichment tools make use of existing gene annotations like the Gene Ontology^{56;57} and the KEGG pathway database^{58;59}. But they can make also use of gene expression data to perform tissue enrichment⁵¹. These tools determine a background distribution for every annotation to assess how likely it is for a random gene to have a specific annotation⁶⁰. If significantly more genes from the input gene list are annotated with a specific term than expected by chance, it is said that the gene list is enriched for that term. The statistics behind these methods usually use a χ^2 or hypergeometric test³⁸. Tools like DAVID use a slightly modified version of the Fisher's exact test⁶¹.

Depending on the experimental setup, the gene list might be subject to biases, like certain genes having a greater chance to end up on the list because the genotyping chip has a special focus on certain genes. In such a case a specific background has to be provided to the enrichment tool to take into account these biases.

1.7. Objectives of this Thesis

A great number of factors contribute to the etiology of the five inflammatory diseases presented previously. This thesis investigates how the known genetic factors could potentially work together to cause disease. For this, the genetic factors will be mapped to genes that are potentially affected by these SNPs. The genes and their products

(proteins) will be mapped into reference networks to get a better understanding on what is common and what is different between these diseases and to potentially gain further insights into mechanisms of these diseases.

While the idea to project GWAS results into networks is far from new⁴⁸, the combination of these specific five diseases is a novelty. A variety of approaches will be tried to create networks that aim to represent and investigate the biology behind these diseases.

Chapter 2.

Material & Methods

“But there’s no sense crying over every mistake. You just keep on trying ‘til you run out of cake. And the science gets done. And you make a neat gun for the people who are still alive.”

— Mad robot GLaDOS (Lyrics)

This thesis features the combination of many approaches and tools to go from GWAS data to network submodules. Figure 2.1 provides a generalized overview which kind of steps have been taken in this thesis. It does not provide details about the specific tools because these differ from section to section and many more combinations of tools within this workflow are conceivable.

2.1. SNPs

The SNPs of five different diseases have been investigated in this thesis: Crohn’s disease, ankylosing spondylitis, psoriasis, primary sclerosing cholangitis and ulcerative colitis. Different parts of this thesis used different approaches. The first approach used VEGAS⁶² to map all known SNPs to genes. These SNPs included also SNPs with a high (non-significant) p-value, because VEGAS can combine non-significant SNP p-values to significant gene-wise p-values. Therefore 130,215 SNPs (without MHC region: 124,489 SNPs) have been used together with their association p-values for each disease. The relevant results can be found in section 3.1.

Later analyses only worked with the lead SNPs instead of all SNPs. These lead SNPs were taken from the paper by Ellinghaus et al.² (Figure 2.2L). In total 210 of the 244 association signals had their lead variant within the 10kb boundary surrounding a gene². The list of SNPs contains only 16 coding variants: 14 missense , 1 frameshift and 1 splice donor².

It was previously observed that many genetic factors are shared among diseases^{7;10}. To get a clearer picture on the common disease association of each SNP, the subset-based meta-analysis method⁹ has been used by Ellinghaus et al.². The subset-based meta-analysis works with a list of SNPs and association p-values from different diseases for each SNP. The method tries all combinations of diseases and calculates a p-value for every combination for a specific SNP being associated with exactly this combination of diseases. In the end the lowest p-value of all combinations is taken for further analysis together with the set of diseases that formed the most significant combination. Figure 2.2R shows to which extent the SNPs are shared among the diseases based on the result of the subset-based meta-analysis (ignoring effect directions).

In addition to the SNP p-values from the subset-based meta-analysis, the SNP p-values from the five individual diseases were also taken into account.

Most SNPs used in this study were detected with the "ImmunoChip" which is a genotyping chip that focusses on immune system-related SNPs¹⁰ and allows cost-effective genotyping of many individuals. However, this chip also introduces a bias because it was designed to detect immune system-related genetic signals and it might fail to capture important signals from other parts of the genome. In total the ImmunoChip allows genotyping of 37,377 LD-independent markers² which represent 195,806 SNPs and 718 small insertion-deletions¹⁰.

Every SNP from Ellinghaus et al.² has an odds ratio for every disease. Odds ratios are calculated as follows⁶³:

$$\frac{D^+}{D^-} : \frac{H^+}{H^-} \quad (2.1)$$

Where D denotes the number of people in the study population who have the disease while H denotes the number of people in the study population that are healthy. The

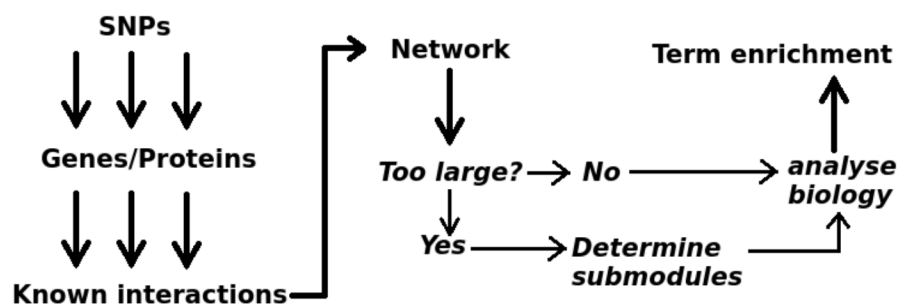


Figure 2.1.: Generalized workflow to work with GWAS data to construct networks. The starting point are SNPs. Depending on the employed method, these SNPs might have to be genome-wide significant. Next it is attempted to determine which genes are affected by these SNPs. These genes are then treated as proteins. They are mapped into reference protein-protein interaction networks which have been previously filtered to only contain edges of a certain minimum confidence value. Then a disease-specific subnetwork is created that may contain nodes that are not associated with a disease but which are directly connected to disease-associated genes. If the resulting network is too large, it has to be split into modules for closer analysis by using an appropriate method. Once a network of manageable size is obtained, it can be analyzed on the biological level by visual inspection and/or by using enrichment analyses.

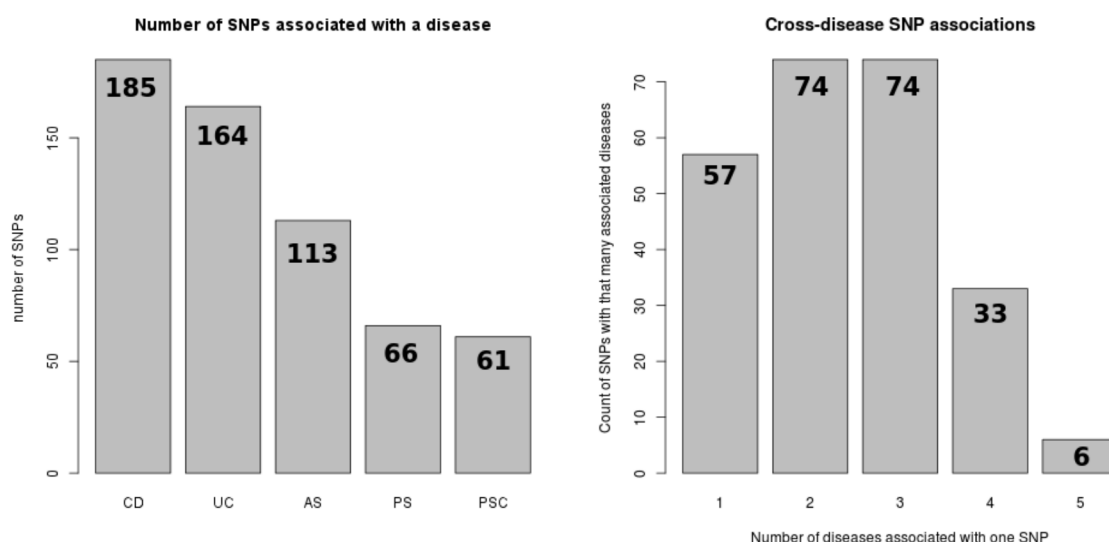


Figure 2.2.: SNP disease-association statistics based on SNPs taken from Ellinghaus et al.². These SNPs were used for the analyses starting at section 3.2.

Left: Number of known SNPs associated with specific diseases. A SNP can be associated with more than one disease. In total 244 lead SNPs from the five diseases were used.

Right: Histogram of the number of diseases that SNPs are associated with. Six SNPs are associated with all five diseases while 57 SNPs are not shared with any of the other four diseases.

CPDB version	number of nodes	number of edges
29 (90 % confidence)	16,620 (9,533)	444,311 (80,422)
30	17,105 (6,815)	471,097 (52,647)
31	17,460 (9,732)	551,362 (81,736)

Table 2.1.: Sizes of interactomes in the protein-protein interaction network database ConsensusPathDB. The numbers in parenthesis are the number of nodes and edges that are left when all edges are removed that have a confidence value less than 95 percent. In the case of version 29 the first row gives the number of nodes and edges for a minimum confidence of 90 percent.

+ modifier selects the number of people that carry a specific allele and the – modifier selects the number of people who do not carry the allele.

An odds ratio of greater 1 indicates risk while an odds ratio smaller than 1 indicates protection. The odds ratios are calculated with the frequency of the minor allele in the population.

2.2. Protein-Protein Interaction Reference Databases

There are different protein-protein interaction databases that can be used for obtaining interaction data to construct networks. There are also metadatabases that consolidate these individual databases into a big dataset.

ConsensusPathDB

The *ConsensusPathDB* (CPDB) is a metadatabase of protein-protein interactions. It consolidates 18 protein-protein interaction databases^{64;65}.

To account for false positive interactions, the CPDB uses the IntScore algorithm⁶⁶ which assesses the plausibility of an interaction based on three topology criteria and three annotation criteria. Every interaction in the CPDB is annotated with a confidence score between zero and one. The results presented in section 3.1 rely on interactions with a minimum confidence of 90 percent while later sections work with a minimum confidence of 95 percent. Given that this work spanned several years, different releases of the CPDB were downloaded and used to construct networks. The corresponding sections in the

results part mention the exact version that was used for a particular analysis. Table 2.1 lists the sizes of the different reference networks.

iRefIndex

The iRefIndex database⁶⁷ is another popular protein-protein interaction metadatabase that consolidates the information from several other databases. One central aim of the iRefIndex metadatabase is to establish unified identifiers for all interactions taken from ten major PPI databases.

The iRefIndex database was used by the International Multiple Sclerosis Genetics Consortium for their analysis procedure¹⁶. For this reason the iRefIndex database (version 13.0) has also been used to test a reimplementations of that procedure.

2.3. Mapping of SNPs to Genes

Various approaches have been used to determine the genes/proteins that are affected by disease-associated SNPs. Simple approaches like taking the closest gene or taking the closest genes within 0.1cM distance have been abandoned in favor of more sophisticated approaches like VEGAS and DEPICT.

VEGAS

VEGAS is a popular^{68;69;16} tool for mapping SNP disease association p-values to gene disease association p-values. VEGAS can be used as a web service but it can also be installed locally. For this thesis VEGAS version 0.8.27 has been downloaded and used.

VEGAS takes a list of SNPs and a genomic reference (by default hg18) as input and produces a table of genes with p-values as output. VEGAS combines SNPs that can be assigned to the same gene and calculates a single p-value. This approach has the advantage that when a SNP has several non-genome-wide significantly associated SNPs, the combination of these SNPs might be sufficient to calculate a significant gene p-value.

The gold standard for assigning several SNPs to a gene is already implemented in the PLINK software package⁷⁰ based on permutations⁶². The disadvantage of this approach

is the long computation time and VEGAS uses Monte Carlo simulations to get a heuristic result instead. It uses a vector-based χ^2 test with one degree of freedom where every component of that vector represents one SNP and the test assesses the significance of a gene.

Additional changes had to be made for the VEGAS software to run with our data. The details are explained in section C.2.

VEGAS has been used for this work (section 3.1) until DEPICT emerged as an alternative.

DEPICT

Another approach to get from SNPs to genes is implemented by the DEPICT software by Pers et al.⁵¹. DEPICT maps SNPs to genes by prioritising genes from every locus based on how similar these genes are to other genes from other loci. The similarity of these genes is based on the number of shared annotations and the similarity of expression profiles. This method is based on the principle that truly associated genes should share functional annotations^{71;45}. Or in other words: The genes in a group of disease-related genes are probably annotated with similar terms.

DEPICT combines geneset enrichment with mapping of SNPs to genes. To account for the incompleteness of gene annotations, DEPICT extends existing genesets (sets of genes with a common/shared annotation) by finding genes that have similar gene expression patterns. With this method 14,461 reconstituted genesets were constructed that cover the complete genome. These reconstituted genesets are precomputed. Furthermore, DEPICT makes use of 37,427 microarrays to identify tissue/cell types to identify highly expressed genes for specific tissues.



Direct collaboration with Tune Pers was undertaken to let DEPICT run with our data using an Immunochip-specific background for the enrichment to avoid biases. Table 2.2 lists the notable options used in the jobfiles when running DEPICT.

DEPICT uses the hg19 genome reference and accounts for multiple testing by calculating the false discovery rate for enrichment results. It was used with the 244 SNPs from the cross-disease study. Several runs were made for disease-specific p-values and Subset-based meta-analysis p-values. DEPICT uses plink2 internally to identify genome-wide significant loci.

GWAS FILE SETTINGS

association_pvalue_cutoff 1e-5

PLINK SETTINGS

genotype_data_plink_prefix [...] / ALL.chr_merged.phase3_shapeit2 
 _mvncall_integrated 
 _v5.20130502.genotypes

DEPICT SETTINGS

step_write_plink_input yes
 step_write_plink_output yes
 step_run_plink yes
 step_construct_depict_loci yes
 step_depict_geneprio yes
 step_depict_gsea yes
 step_depict_tissueenrichment yes

MISC SETTINGS



collection_file ld0.5_collection_depict_150315.txt.gz
 reconstituted_genesets_file GPL570-GPL96-GPL1261 
 -GPL1355TermGeneZScores 
 -MGI_MF_CC_RT_IW_BP_KEGG_z_z.binary
 max_top_genes_for_gene_set 10
 nr_repititions 20
 nr_permutations 500
 hla_start_bp 25000000
 hla_end_bp 35000000

Table 2.2.: Important options that were used when running DEPICT. The list of SNPs differed between each run.

When running, DEPICT makes use of a precomputed gene-to-geneset matrix which contains a probability for every gene to be part of a specific reconstituted geneset. This matrix was created based on gene expression and existing gene annotations. To construct this matrix, every existing geneset was taken and a common expression profile of all genes in that geneset were calculated. Then the expression profile of every gene outside of this geneset was correlated with this common expression profile and the correlation was used to give the gene a probability to be part of that geneset.

The gene prioritisation (SNP mapping) works as follows: DEPICT tries to find the most likely combination of genes across all GWAS loci that are most similar to each other. This similarity is defined through the reconstituted genesets and the tissues where genes share high expression.

GoShifter

A rather unconventional approach to map SNPs to genes was performed by using the tool GoShifter⁷² with ENCODE *transcription factor binding sites* annotations (TFBS). The ENCODE transcription factor binding sites were downloaded on the 8th of September 2015. The two annotation files used for the mapping were ENCFF029ZUJ.bigBed (proximal TFBS) and ENCFF787QYS.bigBed (distal TFBS). The bigBed files were converted to BED files by using the tool bigBedToBed⁷³. All genomic coordinates referred to genome build hg19.

The annotations for the transcription factor binding sites contain the transcription factors that bind to a genomic region and which might exhibit a different binding affinity if a SNP lies within the binding region. It would have been helpful if the annotations also included the genes that are regulated by the binding regions but this information was not available in this context. However, in hindsight it is probably possible to at least map proximal TFBS regions to the regulated genes. Thus this mapping is indirect: A mapping to the potential regulator is performed but the regulated gene itself is not known.

Transcription factor binding sites are especially interesting because they might act as *expression quantitative trait loci* (eQTLs). A different binding affinity of a transcription factor might directly affect the transcription and therefore the expression levels of a gene⁷⁴.

GoShifter works by taking lead SNPs and determines the region in LD around each lead SNP. All annotations in that region are overlaid with all LD SNPs and the number of overlaps is determined as a base value. Then the coordinates of all annotations are shifted by a random amount with wrapping around at the boundaries of the linkage region to preserve local genomic characteristics⁷². After every shift the number of overlaps is recorded. At the end a distribution of overlap counts is compared to the base value to assess the probability of the real LD SNPs overlapping these annotations by chance.

GoShifter also provides a stratified test to test dependence of different annotations against each other. This test works by moving all annotations that should be tested directly adjacent to each other in a linkage block and then the shifting of the other annotation type is performed as normal. Next a background distribution is determined for the number of times a SNP overlaps both annotations versus a SNP overlapping only the of the primary (shifting) annotation.

Because GoShifter does not report the original overlapping annotations when assessing the significance of an overlap, further backmapping had to be done by using the SNP coordinates from the 1000 genomes project and the locations of the ENCODE TFBSs. For every detected overlap the transcription factors were determined and used in later analyses.

The risk status of every SNP was taken from supplementary table 3 of the paper by Ellinghaus et al.². Depending on the SNPs that are linked to a single gene, that gene was assigned one of the following risk statuses:

Protective: Gene is only linked to protective variants

Risk: Gene is only linked to risk variants

Both: Gene is linked to risk and protective variants

None: Gene has no SNPs associated with the currently considered disease

Each of these classes was assigned for each of the five diseases.

2.4. ID mapping

Because different analysis steps/tools use different types of IDs, a mapping between different ID types was required for several procedures. The most frequent conversions were needed from HGNC gene names to UniProt IDs. HGNC Gene names are used by VEGAS and DEPICT and UniProt IDs are used by the ConsensusPathDB. The DAVID webtool accepts a variety of ID types but for this thesis UniProt IDs were supplied.

ID mapping data was downloaded from genenames.org (using BioMart) and UniProt.org (using the FTP server) and custom scripts were written in Python to map IDs from one type to another. Sometimes ID mapping was also performed within Cytoscape by importing mapping tables and matching with existing node IDs.

2.5. Network handling

2.5.1. Cytoscape

Cytoscape is a tool to work with digital networks with a special focus on networks in molecular biology⁷⁵. Cytoscape is a graphical application and supports various visualization, exploration, analysis and processing methods for networks. Throughout this work Cytoscape versions 3.0 to 3.3 have been used.

Cytoscape also can make use of many plugins provided by the network research community. One plugin used in this thesis is jActiveModules⁵⁴. This plugin was originally developed to find submodules in gene expression networks. To determine these submodules, the user has to provide gene-wise p-values and jActiveModules transforms these p-values into z-scores (z_i) for every gene-node i using the inverse cumulative normal distribution. To calculate a score Z_A for a subnetwork with a given set of k nodes, the following formula is used:

$$Z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \quad (2.2)$$

And to calibrate this score against the background distribution the following formula is used:

$$S_A = \frac{Z_A - \mu_k}{\sigma_k} \quad (2.3)$$

Where μ_k and σ_k are the mean and standard deviation of randomly sampled networks respectively.

Because finding the best submodule is NP-hard, jActiveModules uses a simulated annealing heuristic to find good submodules. During the search every node is either marked as active or inactive. Initially, every node has a chance of 50% to be part of the active set. The search is performed by going through N iterations (for a given N). At each iteration one node is picked randomly and its state is toggled (active or inactive). Then the scores of every connected component consisting only of active nodes

is calculated (with equation 2.3). The maximum score is recorded and compared with the maximum score from the previous iteration. If the new score is higher, the node stays toggled. Otherwise is stays toggled with the following probability:

$$\exp(s_i - s_{i-1}/T_i) \quad (2.4)$$

Where s_i, s_{i-1} are the highest module scores in iterations $i, i - 1$ respectively and T_i is the temperature value from the simulated annealing at iteration i .

After the N iterations the highest score and the corresponding connected component is returned. This process can be repeated to find several submodules.

Gene-wise overlap between diseases

To determine subnetworks that are specific to single diseases or specific subsets of diseases, every node in the CPDB reference network (version 29) was annotated with VEGAS gene-wise p-values. A Cytoscape plugin was written to annotate every node with the number of significant disease associations and to annotate each node with the diseases themselves. Then the network was simplified by removing edges that had nodes for which the diseases of these nodes contradicted each other.

Formally "contradiction" can be expressed as follows: Let e be an edge connecting two nodes n_1, n_2 . Let $d(n_1), d(n_2)$ be the sets of diseases associated with these nodes respectively, then edge e is kept in the network if the following relationship is true:

$$d(n_1) \subseteq d(n_2) \vee d(n_1) \supseteq d(n_2) \quad (2.5)$$

That is, the diseases of one node have to be a superset of the diseases of another node to ensure that only nodes are connected that could form a disease-specific subnetwork.

The nodes were then filtered into different categories by the number of diseases that are associated with every node to form disease-subset-specific subnetworks.

Geneset-based networks

The tool DEPICT was used to determine the most consistent sets of genes based on the list of 244 cross-disease SNPs (See section 2.3 for details). DEPICT was run with each disease-specific list of SNPs together with their p-values to determine the disease-specific p-values. In addition, DEPICT was run with the list of SNPs with the SBM p-values.

The protein-protein interactions from the ConsensusPathDB version 30 were filtered to have a minimum confidence value of 95 % and the genes from DEPICT were mapped into the network using Cytoscape. In addition, the genes from the paper "Genetic insights into common pathways and complex relationships among immune-mediated diseases"¹⁰ were mapped into the network to see how well DEPICT was able to determine already known disease genes.

For every node the associated diseases were determined and each node was annotated with a stripe chart using the Enhanced Graphics Cytoscape plugin⁷⁶. The resulting network was laid out manually for better visual inspection.

REST-based scripting

As an easier alternative to writing plugins, Cytoscape 3 originally had JavaScript-based scripting support which was also used to automate some parts of this work. This scripting interface was deprecated in later versions of Cytoscape 3 because of incompatibilities with the OSGi classloader which is a fundamental component of Cytoscape 3.

Keino et al. created the CyREST plugin for Cytoscape which exposes a REST interface to the most common basic actions that would normally be performed with a script⁷⁷. With the help of this plugin a wrapper of the REST interface was written in Clojure⁷⁸. This made it possible to remote-control many parts of Cytoscape and also to automate many tasks used to create the results in section 3.2 by writing regular Clojure code instead of using raw REST calls.

The code is located in the supplementary folder `cyREST-clojure_scripts`.

2.5.2. Effect directions of SNPs on the network level

While Cytoscape provided many convenient facilities to work with networks, it had one fundamental flaw: As a graphical application it is not possible to create a reliable data processing pipeline from start to finish. Furthermore, Cytoscape provides no good mechanisms to document the individual transformation and processing steps so that it is easier later on to understand what was done and to automatically redo these steps.

The networkx python library (version 1.11) provides many facilities to work with networks and is fully scriptable⁷⁹. The Jupyter notebook (version 4.1.0) provides an infrastructure to write code in many different programming languages while at the same time providing an easy way to write the rationale for every piece of code directly next to it⁸⁰. The Jupyter notebook improves on methods like knitR and Sweave by not requiring a recompilation of the whole document when a change is made or when a new piece of code or text is added.

The Jupyter notebook was used together with networkx to investigate how the effect directions of SNPs might play out on the network level (see section 3.3 for results).

The networks were created by using the ConsensusPathDB version 31 with interactions that had at least a confidence value of 95 % which resulted in a global network of 9,732 nodes and 81,736 edges.

The disease genes and their risk status were taken from the procedure described in section 2.3. They were mapped into the network and different centrality measures and assortativity values were calculated using the built-in functionality of networkx.

2.5.3. Network randomization

In order to assess how similar the networks are to random networks, random networks have been generated with different approaches. It is possible to generate random networks based on picking random nodes or picking random edges while maintaining the overall number of nodes and/or edges to keep the networks comparable. After generating the random networks, a distribution over the metrics of the random networks is generated and it is investigated in which quantile the metrics of the real network(s) lie. If they are less than the 2.5 % quantile or more than 97.5 % quantile, they are significantly different from random networks.

To follow the example of the IMSGC, random networks were generated by randomly sampling nodes and comparing the size of the largest connected component (nodes and edges) of the real disease networks with the random networks. This randomization and test procedure was written in D⁸¹ and 10,000 random networks were generated for each of the five diseases. The randomization tests were done based on both, the ConsensusPathDB (version 29) and the iRefIndex database version 13.

Edge randomisation is more meaningful than node randomization, because biological networks tend to follow a power law which means that they contain only a few nodes of very high degree. These nodes have a great influence on the connectivity of networks. Randomly picking nodes will likely lead to an underrepresentation of these high-degree nodes and it is therefore preferable to use edge randomisation which ensures that every node maintains its degree, but the edges themselves point to other nodes.

Formally this procedure can be described as follows:

Let $G = (V, E)$ be a graph where V is the list of nodes and E is the list of edges. Each edge $e_i = (s_i, t_i)$ has a source s_i and a target node t_i . To randomise the edges, two lists are generated:

$$S = (s_i \mid i \in 1, \dots, |E|) \quad (2.6)$$

$$T = (t_i \mid i \in 1, \dots, |E|) \quad (2.7)$$

Where S is the list of source nodes and T is the list of target nodes. Based on T a new list T' is generated by shuffling the entries in T and then a new list of edges is generated:

$$E' = ((S_i, T'_i) \mid i \in 1, \dots, |E|, s_i \neq t_i, \nexists i, j : E'_i = E'_j \wedge i \neq j) \quad (2.8)$$

The edge randomisation was implemented in Python. If an edge was randomly generated that did not fulfill the conditions in equation 2.8, it was attempted to find a replacement edge that fulfills the conditions. Most of the time this was possible but in general $|E| \approx |E'|$ is only an approximation (observed worst case: Out of 1546 edges, 9 edges had to be dropped).

Ten thousand randomizations were performed for the disease networks to assess their dissimilarity to random networks based on the ConsensusPathDB reference network. The sizes of the greatest connected components of each random and non-random network were determined and the position of every real network within the distribution of random networks was checked for significance.

2.6. Term- and Tissue-Enrichment

DAVID

DAVID is a web-based enrichment tool. It is possible to upload lists of genes and test these lists for enriched annotations. DAVID provides annotations from nine different backend databases⁶¹.

Throughout this work DAVID was used with the default annotations which also included Gene Ontology terms and KEGG pathways. To perform the enrichment analyses, the tool "functional annotation clustering" was used to obtain clusters of enriched terms.

DAVID uses a slightly modified Fisher's exact test (called EASE score). In this test the number of genes from the input list that are within a geneset (or pathway) is reduced by one before performing the normal Fisher's exact test calculations.

DAVID calculates a p-value and a Benjamini-Hochberg-corrected p-value for every enriched term. It also clusters enriched terms together when there is a large number of genes that share several terms. These clusters often have a common theme (like immune-system processes). Each cluster gets a score and enrichment results are sorted by this score.

DEPICT

DEPICT combines geneset enrichment with mapping of SNPs to genes. Please see section 2.3 for the details on how DEPICT performs SNP mapping.

DEPICT determines enriched tissues and enriched genesets for a given list of SNPs. The enrichment results for tissues and for genesets were further categorized by how many diseases share a tissue or geneset enrichment respectively. These enriched terms and

tissues were sorted by their score and heatmaps were created with R using `heatmap.2` from the `gplots` package.

GoShifter

GoShifter⁷² (Genomic Annotation Shifter) was primarily developed by Trynka et al. to test whether a set of SNP overlaps with a set of genomic annotations and whether this overlap is not by chance. To preserve local genomic structures, GoShifter randomly shifts around annotated regions within linkage boundaries ($r^2 > 0.8$) and determines all LD SNPs for a given lead SNP within that region. The shifting is wrapped around at the boundaries of the region to ensure that no annotations are lost.

GoShifter determines a null distribution to calculate the expected number of overlaps under random conditions. The real number of overlaps is later compared to the distribution of random overlaps and a score is calculated that indicates how likely it is that the real overlap is by chance. However, no clear cutoff is given. In addition to the individual scores for every lead SNP, GoShifter also calculates a global p-value for the combination of the list of lead SNPs and the set of annotations to indicate if the annotations are likely to be relevant for the list of SNPs.

GoShifter uses the hg19 genomic reference and also the SNP information from the 1000 genomes project under standard conditions. For this work the GoShifter development version 0.2 was used (personal communication with Gosia Trynka).

KEGG

KEGG is the Kyoto Encyclopedia of Genes and Genomes^{58 59}. It contains many curated pathways but also offers pathway enrichment and pathway search functionality on the website. KEGG has been used to identify pathways when manually looking up the functionality of various genes.

Chapter 3.

Results

“Now, these points of data make a beautiful line. And we’re out of beta. We’re releasing on time! So I’m GLaD I got burned! Think of all the things we learned! For the people who are still alive.”

— Mad robot GLaDOS (Lyrics)

There are many different approaches to work with GWAS data to create networks⁴⁸. This is also shown in the upcoming sections where different methods have been applied to investigate inflammatory diseases.

The first approach is based on an analysis workflow presented in the paper *“Network-Based Multiple Sclerosis Pathway Analysis with GWAS Data from 15,000 Cases and 30,000 controls”*¹⁶. This workflow was reimplemented and adapted for the cross-disease project.

The next network approach was based on the DEPICT software by Tune Pers et al.⁵¹. DEPICT tries to find plausible combinations of genes from GWAS data based on preexisting gene annotation. The genes were then mapped into reference networks.

Lastly the direction of effects of SNPs on the network level have been investigated to understand better how protective SNPs may influence the susceptibility for disease.

3.1. Common Network Modules Across Inflammatory Diseases

Many SNPs used in this thesis are related to several diseases at once (for example rs13407913², rs2812378²). Network-based approaches can be used to investigate how these multi-disease relations manifest themselves on the network level i.e. which modules within a network are specific to single diseases and which modules are shared. This in turn might provide better insights into the common mechanisms behind these inflammatory diseases but also into the mechanisms that are specific to single diseases.

The International Multiple Sclerosis Genetics Consortium developed a workflow to derive network submodules from GWAS data in multiple steps¹⁶. This workflow has been adapted and reimplemented for this thesis, to work with our multi-disease SNP sets from the five diseases.

3.1.1. SNP to Gene Mapping

VEGAS is a tool to determine gene-level p-values from SNP-level p-values. VEGAS was run with 130,215 SNPs with disease-specific p-values for each disease. A total of 5,726 SNP are located in the MHC region and VEGAS was used with SNP lists containing the MHC region as well as SNP lists without them. It should be noted that these lists contain SNPs which are not all genome-wide significantly associated. But VEGAS can combine several non-significant SNP p-values into a significant p-value for a gene (see methods).

Table 3.1 displays the number of genes VEGAS listed based on the p-values of the SNPs of each individual disease. There are major differences between the number of genes determined when MHC SNPs were used and when they were excluded (Table 3.2). Some of the gene-wise p-values that VEGAS produced are equal to zero (Table 3.3) which is a phenomenon that the authors of VEGAS describe to be due to "computational reasons".

A comparison with the 1,443 IBD genes from Jostins et al.⁷ showed that 906 of the 5,827 genes detected by VEGAS are common between the two studies. When the MHC region is left out, the overlap with the genes by Jostins et al. is 877 genes.

Disease	No MHC		With MHC	
	significant genes	total genes	significant genes	total genes
All	3,612	12,428	3,888	12,705
AS	2,201	12,420	2,465	12,686
CD	2,832	12,407	3,115	12,693
PS	1,488	12,410	1,706	12,675
PSC	1,527	12,408	1,789	12,674
UC	2,472	12,420	2,712	12,685

Table 3.1.: Number of genes reported by VEGAS with and without MHC region. The significance threshold is five percent. The corresponding genes were then used for further analysis. "All" denotes the combination of SNPs from all five diseases.

Disease	Common genes	Additional genes without MHC	Additional genes with MHC
All	2,841	771	1,047
AS	1,164	1,037	1,301
CD	2,218	614	897
PS	563	925	1,143
PSC	1,035	492	754
UC	2,343	129	369

Table 3.2.: Numbers of significant genes detected by VEGAS (significance threshold of 5%) for different diseases. VEGAS was used with SNP lists lacking and containing MHC region SNPs in different runs. The "Common genes" are genes detected by VEGAS with both lists. As indicated above the results from the non-MHC runs differ from the with-MHC runs to a notable extent.

Disease	Common genes	Additional non-MHC genes	Additional MHC genes
All	174	420	698
AS	21	48	295
CD	315	112	219
PS	30	42	189
PSC	14	48	309
UC	116	100	217

Table 3.3.: Counts of VEGAS gene-wise p-values that are exactly zero. VEGAS was run with SNP lists containing and lacking the MHC region. The "Common genes" are genes detected by VEGAS with both lists. According to the paper by Liu et al.⁶², VEGAS produces these p-values because of "computational reasons".

Disease	Number of nodes	Number of edges
AS	1122	1279
with MHC	1234	1546
CD	1504	2463
with MHC	1616	2840
PS	753	751
with MHC	852	964
PSC	808	923
with MHC	923	912
UC	1295	2214
with MHC	1403	2564

Table 3.4.: Sizes of disease-specific subnetworks containing only nodes that have been identified by VEGAS as significantly associated (significance threshold 5%). The protein-protein interactions themselves have been taken from the ConsensusPathDB (version 29) and were filtered to have at least a confidence value of 90%. Each of these subnetworks had one large connected component, some very small independent subnetworks and many unconnected nodes (not shown).

The significantly associated genes from all runs were then used for further analysis.

3.1.2. Disease networks

In order to see how the genes that were detected by VEGAS interact with each other, the protein-protein interaction reference ConsensusPathDB⁶⁴⁶⁵ version 29 has been downloaded and filtered down to only contain those interactions that have a confidence value of at least 90 percent. The remaining interactions have been combined into a network of 9,533 nodes and 80,422 edges inside Cytoscape.

The genes from VEGAS were mapped into the network. For every disease and every gene and each of the two MHC configurations it was determined if the VEGAS p-value was below 5% and if that was the case, the gene was included in the disease-specific subnetwork. Table 3.4 lists the sizes of the subnetworks.

Further analyses also took into account so-called *linker genes* which are genes/proteins that have a known interaction with a disease gene but are themselves not known to be associated with the disease. When including linker genes the networks become a

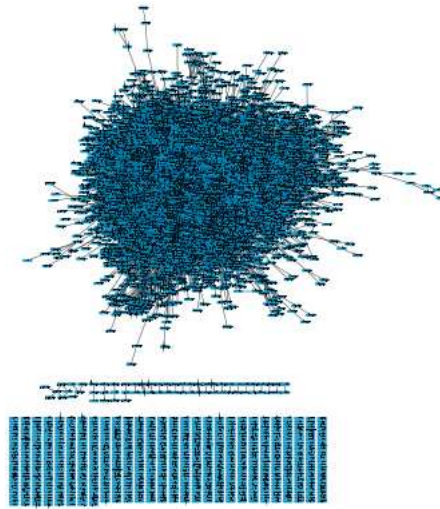


Figure 3.1.: Network of all disease genes with their direct neighbours which themselves might not be known to be associated with any of the five diseases. The network consists of 8,105 nodes and 52,523 edges. The network itself is very large and hard to investigate visually. In network science such networks are called "hairballs" because of their visual complexity⁸².

considerably larger (see Figure 3.1) and an analysis becomes more complex. In principle linker genes may actually turn out to be disease genes^{83;84}. They will be taken into account in section 3.2.6 but for this part they will not be shown because the data is too noisy.

Permutation tests

In accordance to the procedure from the paper by the Multiple Sclerosis Genetics Consortium, distributions of random networks were generated by randomly sampling $n_{d,M}$ proteins from the ConsensusPathDB where $n_{d,M}$ is equal to the number of significant genes in disease d with MHC-configuration M (including MHC genes or leaving them out). For each of these randomly sampled proteins a subnetwork was created and the number of edges and the number of nodes in the largest connected component was determined for comparison with the real subnetworks. This was done to see if the real networks are different from randomly generated networks. For every distribution 10,000 random samples were chosen.

Figures 3.2 and 3.3 depict the number of edges and the number of nodes in the largest connected components in different quantiles compared to the real number of edges and nodes. It can be observed that the subnetworks of the real networks were significantly

different from all random subnetworks (greater number of nodes and edges than all random subnetworks).

The IMSGC did not use the ConsensusPathDB for their workflow. They used the iRefIndex⁶⁷ protein-protein meta-database. For comparison the same permutation tests were done with the iRefIndex meta-database and the real networks were significantly different from the random networks.

Node randomization has the fundamental problem that it does not reflect the distribution of node degrees well. Biological networks are often scale-free and therefore the degrees are far from evenly distributed (see section 1.5.1). When randomizing networks it is much better to randomize the edges and maintain the degree of every node.

An additional edge randomization (10,000 permutations) for every disease subnetwork with each MHC configuration was done. With one exception, the largest connected component of the random networks was always larger than the largest connected component of the corresponding real network. The only exception was the PS network. But this network still had a smaller largest connected component than 99% of all random PS-networks and thus the real networks are all significantly different from the random networks. Table 3.5 contains the percentiles of the random networks.

3.1.3. Protein-interaction-network–based pathway analysis (PINBPA)

After the confirmation of the non-randomness of the disease subnetworks, each of these subnetworks had to be investigated more closely. Given the size of the subnetworks (Table 3.4), it is hardly feasible to understand them by visual inspection. Thus it makes sense to find submodules of smaller size that can be characterized more closely.

The Cytoscape plugin jActiveModules⁵⁴ was used with every disease subnetwork and up to 20 submodules were generated. Table 3.6 lists the scores of the best modules and figures 3.4 to 3.14 show the submodules themselves.

To get an idea about what each module could represent on the biological level, every module was tested for term enrichment with the DAVID webtool⁶¹. The figures 3.4 to 3.14 summarize the enrichment results in their caption texts.

Disease	MHC	Real network	0 %	25 %	50 %	75 %	100 %
AS	no	575	590	614	618	623	641
AS	yes	671	685	712	717	721	739
CD	no	913	925	944	949	952	966
CD	yes	1009	1028	1050	1055	1059	1073
PS	no	346	341	362	365	369	379
PS	yes	420	424	449	453	456	470
PSC	no	352	366	391	397	402	426
PSC	yes	466	472	502	507	512	530
UC	no	785	795	815	819	823	836
UC	yes	868	883	907	911	915	927

Table 3.5.: Percentiles of the random networks obtained from edge endpoint randomisation. All numbers correspond to the number of nodes in the *largest connected component* (LCC) in each network. With the exception of PS (without MHC) all non-random networks have a smaller LCC than all random networks. But more than 99 percent of the PS-derived random networks have a LCC greater than the non-random PS network and therefore all non-random networks are indeed non-random.

Module	jActiveModules score	Figure
AS module 1	7.16	3.4
CD module 1	7.89	3.5
PS module 1	5.48	3.6
PS module 2	5.36	3.7
PSC module 1	5.52	3.8
UC module 1	6.63	3.9
All five diseases module 1	10.44	none (too big)
AS module 1 with MHC	8.67	3.10
CD module 1 with MHC	8.80	3.11
PS module 1 with MHC	7.56	3.12
PSC module 1 with MHC	7.71	3.13
UC module 1 with MHC	6.52	3.14

Table 3.6.: Scores of the best modules produced by the jActiveModules Cytoscape plugin. According to personal communications with the IMSGC, a module score of at least 3.0 was considered to be the threshold for a sound module.

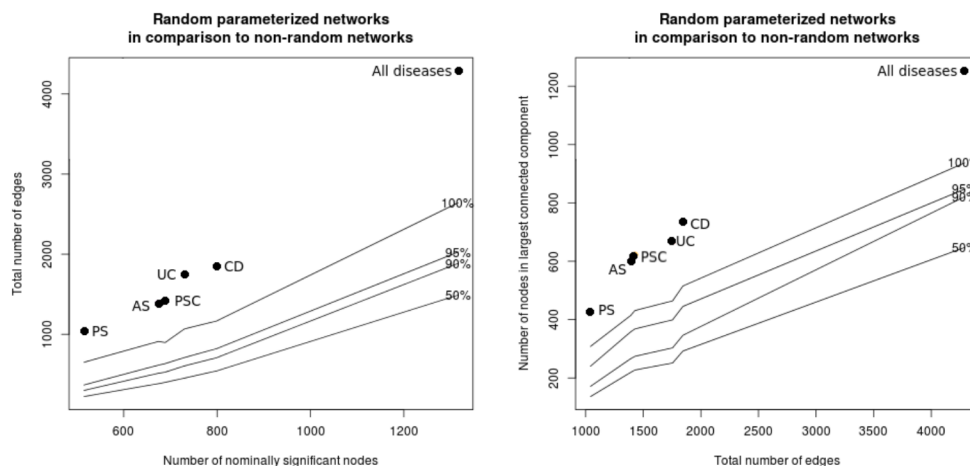


Figure 3.2.: Permutation tests of disease-specific networks without MHC region. For every disease with n disease-specific significant genes (VEGAS p-value < 0.05), n random proteins were sampled from the ConsensusPathDB (10,000 times). These proteins were used to construct subnetworks. The sizes of the subnetworks and the sizes of the largest connected components were determined to create random distributions for comparison with the real disease subnetworks. In any case the real subnetworks (dots) were significantly different from the random networks. The lines indicate the percentiles of the random networks.

Left plot: Number of nodes and the corresponding number of interactions in random networks and in the real networks.

Right plot: Number of edges and the corresponding number of nodes in the largest connected component in random networks and in the real networks (the same networks as in the left plot).

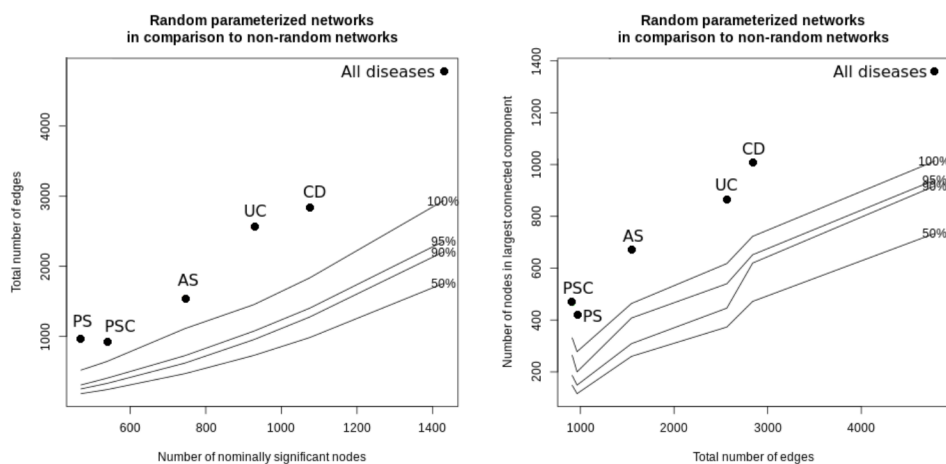


Figure 3.3.: Permutation tests analogous to figure 3.2 but with subnetworks that include genes from the MHC region. The real subnetworks are significantly different from all random networks.

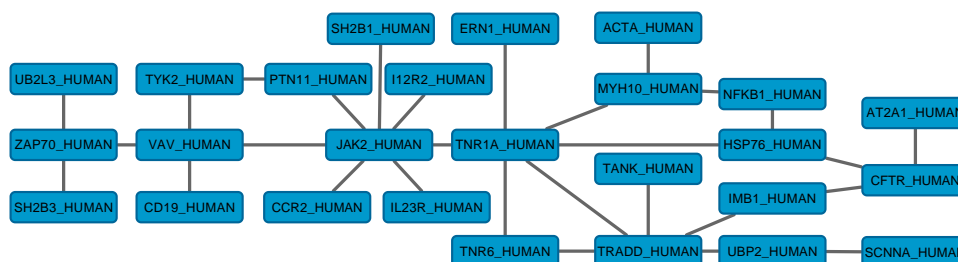


Figure 3.4.: Best jActiveModule for ankylosing spondylitis with a score of 7.16 – The module consists of 26 nodes and 29 edges. Major enrichment terms are SHC2 domain and intracellular signaling cascade. SH2 is known to be a common element of intracellular signaling cascades and the human genes ZAP70, TKY2, PTN11, SH2B1, JAK2, SH2B3 and VAV are annotated with this term. Further terms include ATP binding (10 of 26 nodes) and various binding with other biochemical molecules. Cell death and apoptosis is another prominent theme of this module.

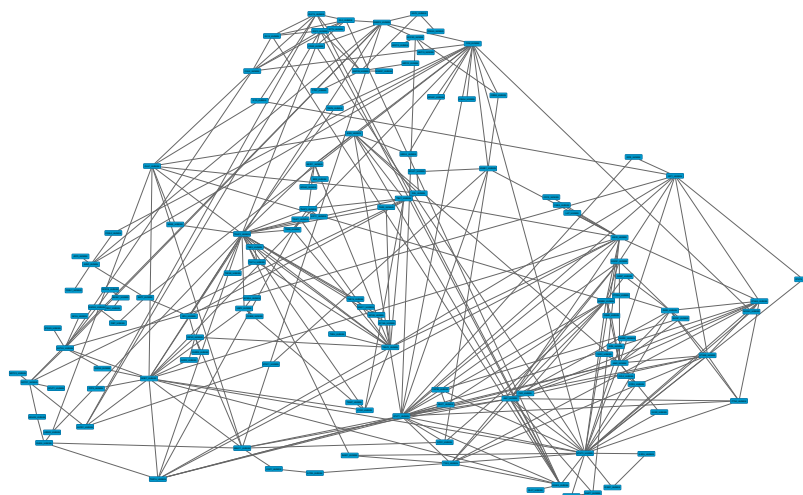


Figure 3.5.: Best jActiveModule for Crohn's disease with a score of 7.89 – The module consists of 126 nodes and 293 edges. The network is too large to show it in full detail. The network is enriched for Jak-Stat signaling and SH2 domains. It appears to play a role in the regulation of the I- κ B kinase cascade. 33 out of the 126 nodes are annotated with ATP-binding and regulation of apoptosis/cell death. Regulation of cytokine production is related to 18 genes. T-cell activation is relevant for twelve genes. More terms include response to wounding, inflammation, response to bacterium (NOD2 and twelve others).

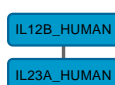


Figure 3.6.: Best jActiveModule for psoriasis with a score of 5.48. No enriched terms were found which is probably due to the small size of this module. However, it is known that these two molecules interact with each other to form the IL23 interleukin⁸⁵.

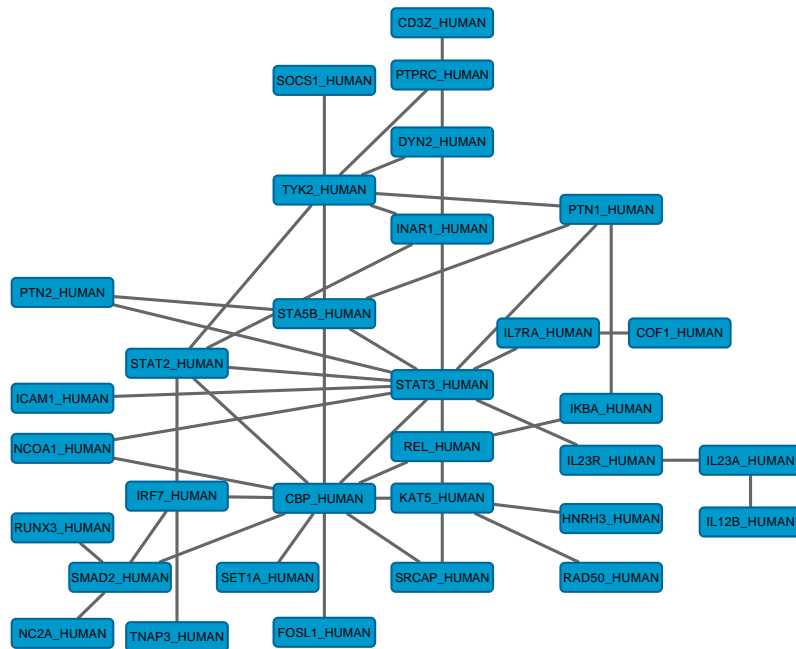


Figure 3.7.: Second-best jActiveModule for psoriasis with a score of 5.36 – This module consists of 32 nodes and 46 edges. This module is enriched for transcriptional regulation and DNA binding. The Jak-Stat pathway and the SH2 domain are a common annotation term. 6 genes are annotated with T-cell activation: STA5B, IL7RA, ICAM1, IL23A, INAR1, IL12B.

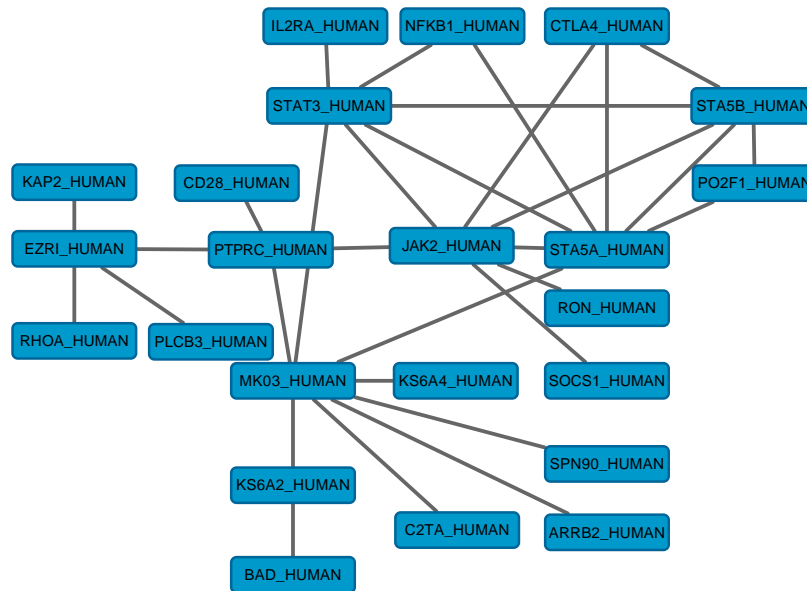


Figure 3.8.: Best jActiveModule for primary sclerosing cholangitis with a score of 5.52 – This module consists of 23 nodes and 31 edges. The genes in this module are annotated with Jak-Stat signaling, SH2 domain, regulation of growth, regulation of inflammatory response, positive regulation of the differentiation of various immune cells.

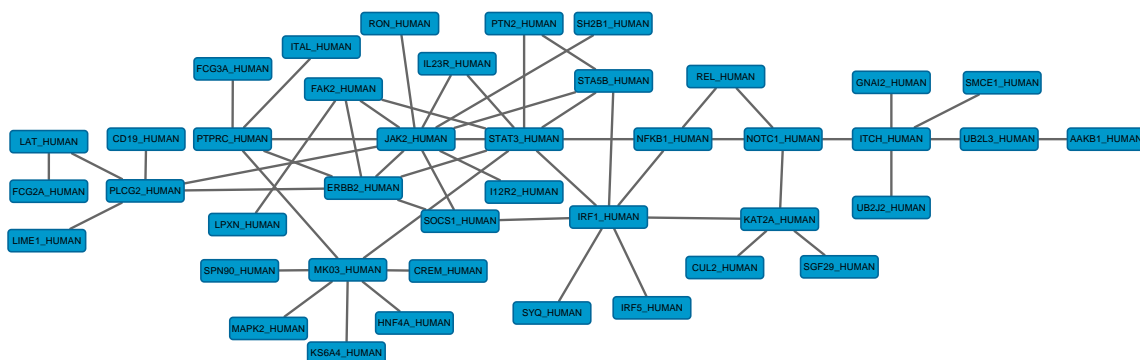


Figure 3.9.: Best jActiveModule for ulcerative colitis with a score of 6.63 – This module consists of 41 nodes and 55 edges. Annotations for this module include enzyme/protein-kinase binding, Jak-Stat signaling pathway, response to wounding, defense response, inflammatory response, SH2 domain, response to organic substance (ten nodes).

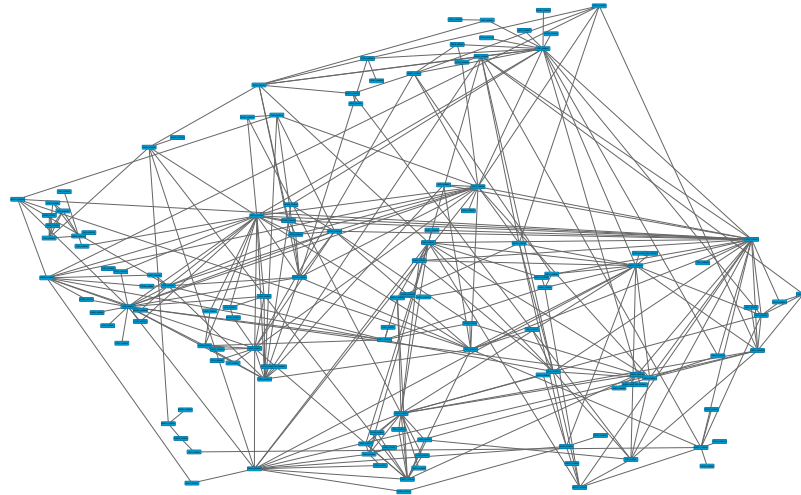


Figure 3.11.: Best jActiveModule for Crohn's disease with MHC with a score of 8.08 – This module consists of 125 nodes and 317 edges and is too large to show it in full detail. Enrichment terms include response to molecule of bacterial origin and response to lipopolysaccharide. The regulation of protein kinase cascades and cell communication/signal transduction are further terms as well as apoptosis and cell death. 35 of the genes are annotated with regulation of cell proliferation and 34 are annotated with regulation of cell death (overlap: 22 genes). Another theme are transcriptional regulation, Jak-Stat signaling, the SH2 domain and immune cell regulation.

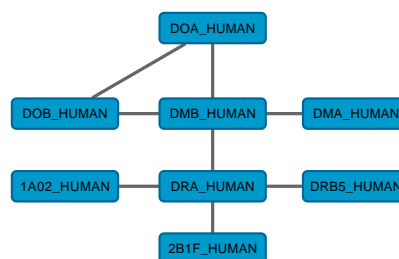


Figure 3.12.: Best jActiveModule for psoriasis with MHC with a score of 7.56 – This module consists of eight nodes and eight edges. The DAVID webtool had difficulties mapping five of the eight IDs. These five IDs belong to HLA-related (MHC) genes. The remaining three genes were annotated with terms like MHC protein complex, allograft rejection, Type 1 diabetes mellitus, immune response, glycoprotein, transmembrane.

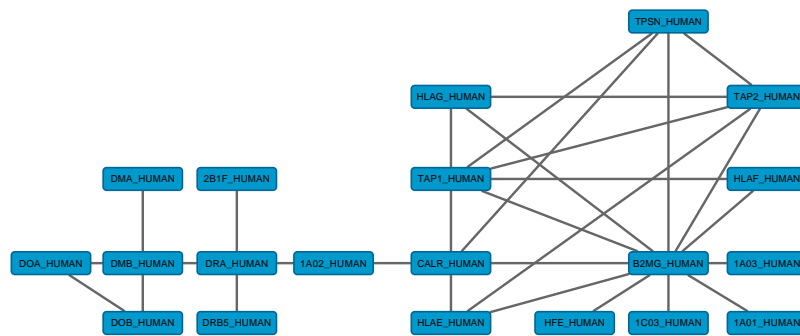


Figure 3.13.: Best jActiveModule for primary sclerosing cholangitis with MHC with a score of 7.71 – This module consists of 20 nodes and 30 edges. Notable enriched terms include antigen processing and presentation, various MHC terms, immune response, host-virus interaction, membrane and transport, TAP/1/2 binding.

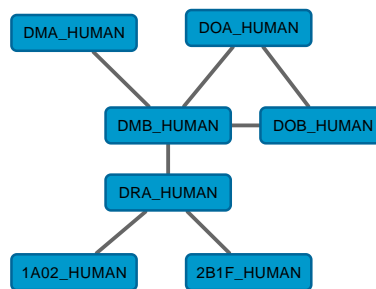


Figure 3.14.: Best jActiveModule for ulcerative colitis with MHC with a score of 6.52 – This module consists of 7 nodes and 7 edges. The second-best module (not shown) consists of 101 nodes and 182 edges. The DAVID enrichment results included the Jak-Stat pathway, SH2 domain, response to organic substance, response to ethanol (STA5B, FAK2, STAT3, corrected p-value 20 %).



Figure 3.15.: Color legend for nodes that are associated with different numbers of diseases.

PINBPA

The workflow in the paper "Network-Based Multiple Sclerosis Pathway Analysis with GWAS Data from 15,000 Cases and 30,000 Controls" requires many individual steps that need to be combined together. Wang et al. wrote the PINBPA plugin for Cytoscape to encapsulate most of this workflow as a Cytoscape plugin⁸⁶. Repeating these steps with the help of this plugin yielded subnetworks that were considerably larger than the networks obtained with jActiveModules.

A central problem with the PINBPA plugin is that it is designed to follow the procedures in the IMSGC paper¹⁶ very closely and does not offer adaptation to different workflows, like analyzing five diseases at once. Because the source code is not available (even not on request), it was not possible to adapt the plugin to my own needs and therefore the analysis with PINBPA has been discontinued.

3.1.4. Gene-wise Overlap Between Diseases in Networks

In order to better understand which parts of a network are relevant for a subset of the diseases and which parts are shared by all or most of the diseases, edges were pruned so that only those nodes were connected that had a consistent set of diseases associated with them (see section 2.5.1 for details).

After pruning the network (by removing 162 edges and 1135 singleton nodes), 351 nodes and 523 edges remain. Figure 3.16 shows that about one quarter of the genes are associated with all five diseases while another quarter is associated with exactly one disease. About one sixth of the nodes are associated with either exactly four, three or two diseases.

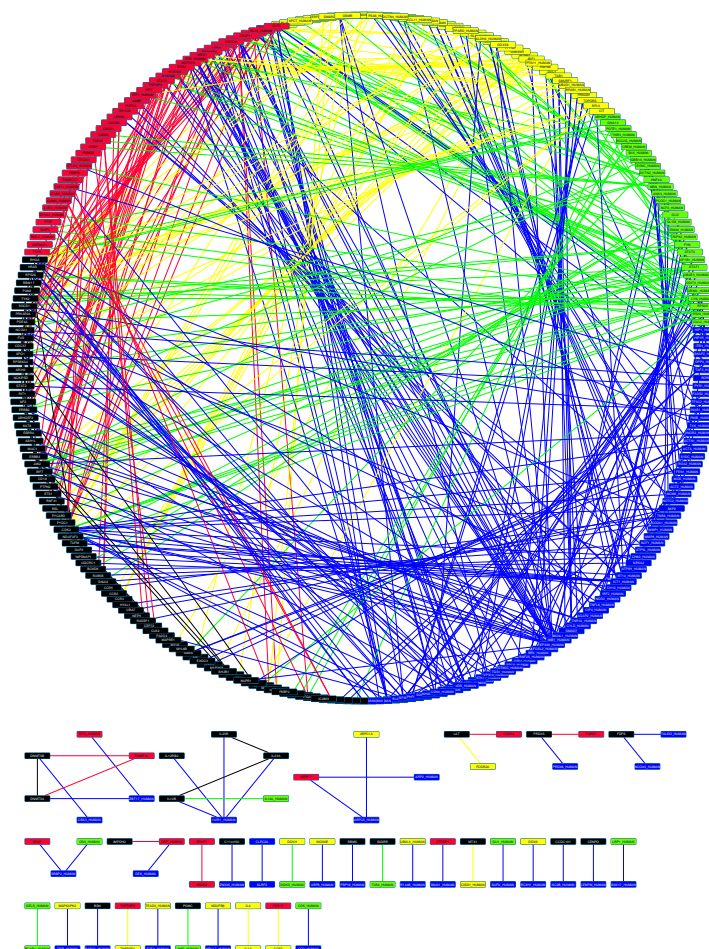


Figure 3.16.: Interactions between proteins that are associated with different numbers of diseases. In total the network consists of 351 nodes and 524 edges. Colors correspond to the legend in figure 3.15. The edges have the color of the connecting node that has the least diseases associated with it.

Number of diseases	5	4	3	2	1	Number of nodes
5	40	116	95	85	145	98
4	116	10	37	40	86	56
3	95	37	3	19	67	58
2	85	40	19	1	53	40
1	145	86	67	53	19	99

Table 3.7.: Symmetric matrix of the number of edges between nodes that have different numbers of diseases associated with them. For every edge the relationship in equation (2.5) holds true. In addition the number of genes that are associated with a specific number of diseases is shown on the right. These numbers are based on the network in figure 3.17.

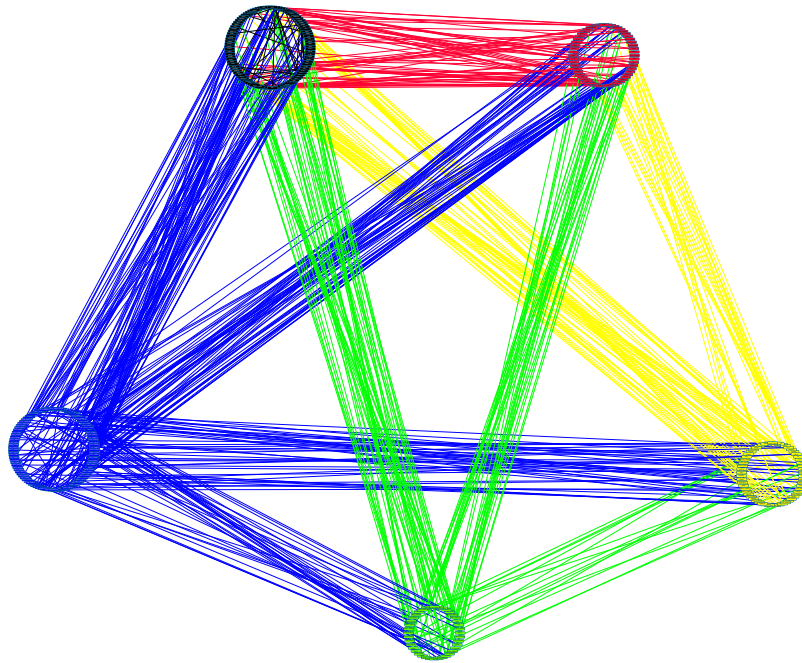


Figure 3.17.: Interactions between proteins that are associated with different numbers of diseases. In total the network consists of 351 nodes and 524 edges. See table 3.7 for details on the number of edges between each cluster. Colors correspond to the legend in figure 3.15. The edges have the color of the connecting node that has the least diseases associated with it.

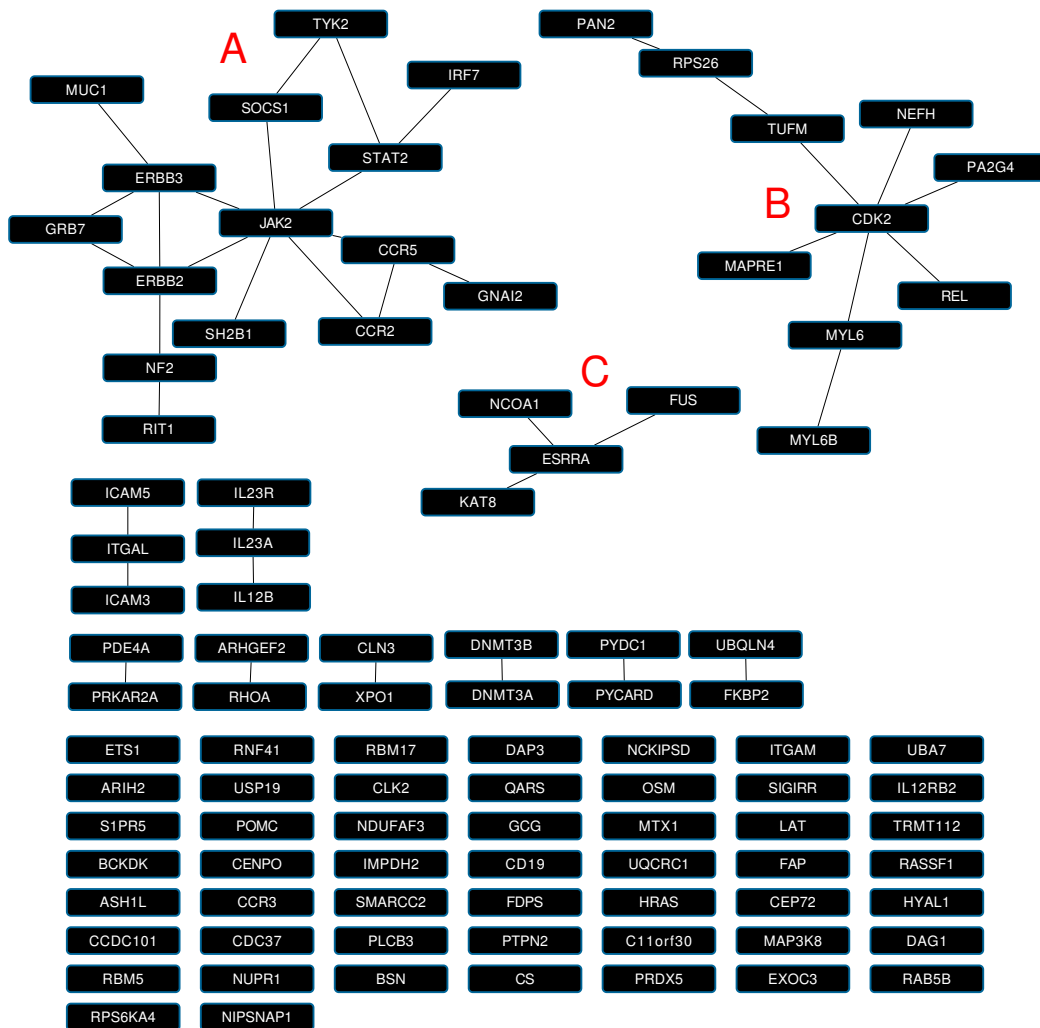


Figure 3.18.: Nodes associated with all five diseases. The network consists of 98 nodes and 40 edges. Subnetwork A is enriched for Jak-Stat signaling. Subnetwork B is enriched for cell cycle regulation. Subnetwork C is enriched for DNA binding and histone modifications.

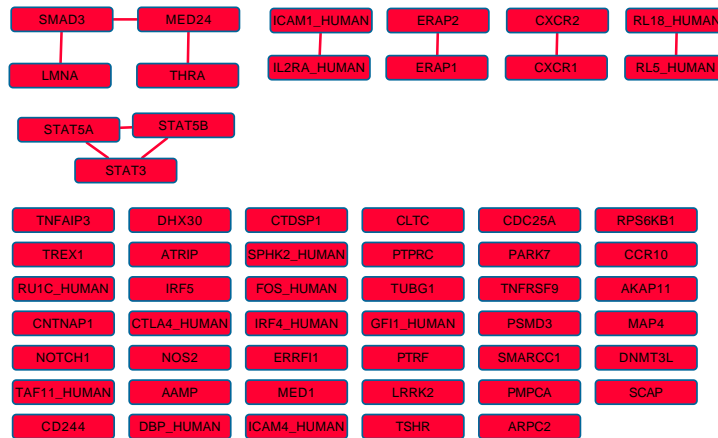


Figure 3.19.: Nodes associated with exactly four diseases. The network consists of 55 nodes and ten edges. The connected component consisting of LMNA, SMAD3, MED24 and THRA are associated with AS, CD, PS, and UC. ICAM1 and IL2RA are not associated with UC. ERAP1 and ERAP2 are not associated with CD. CXCR2 and CXCR1 are not associated with PS. RL18 and RL5 are not associated with UC. The STATs are not associated with AS.

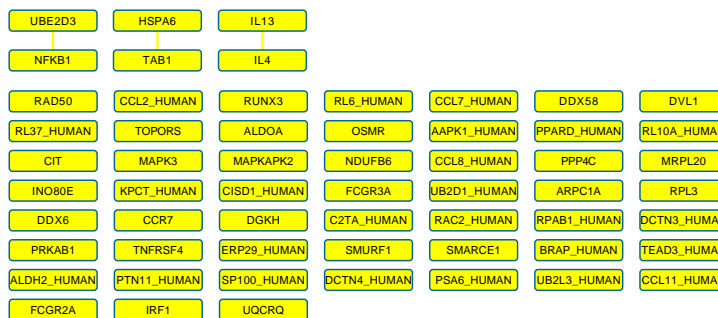


Figure 3.20.: Nodes associated with exactly three diseases. The network consists of 58 nodes and three edges. UBE2D3 and NFKB1 are associated with AS, PSC and UC. HSPA6 and TAB1 are associated with AS, CD and UC. IL4 and IL13 are associated with CD, PS and UC.

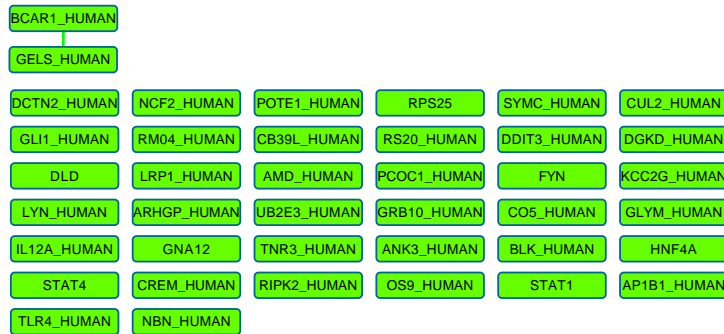


Figure 3.21.: Nodes associated with exactly two diseases. The network consists of 30 nodes and one edge. BCAR1 and GELS are associated with CD and PSC.

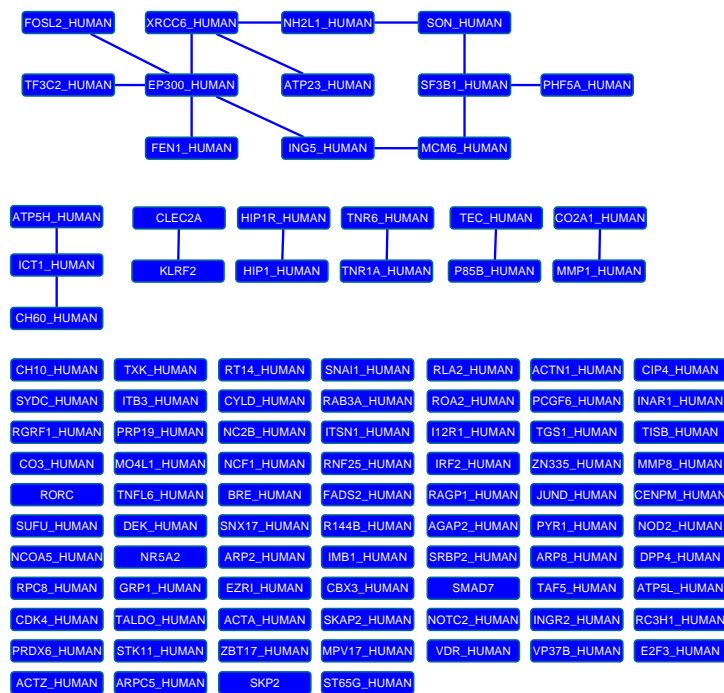


Figure 3.22.: Nodes associated with exactly one disease. The network consists of 99 nodes and 19 edges. All nodes in the largest complex are associated with Crohn's disease. An enrichment analysis and a manual investigation of this complex did not detect a common theme of these connected genes. The complex consisting of ATP5H, ICT1 and CH60 is also associated with CD. CLEC2A and KLRF2 are associated with PSC. The following pairs of genes are all associated with AS: HIP1R and HIP1, TNFR6 and TNFR1A, TEC and P85B, CO2A1 and MMP1.

Disease	Number of SNPs	Number of FDR-significant reconstituted geneset terms	p-values
All 5	244	1981	$\leq 1.12^{-3}$
CD	185	1730	$\leq 1.08^{-3}$
UC	164	1447	$\leq 1.53^{-3}$
AS	113	815	$\leq 8.50^{-4}$
PS	66	1006	$\leq 7.35^{-4}$
PSC	61	508	$\leq 2.88^{-4}$

Table 3.8.: Numbers of enriched genesets determined with DEPICT given different sets of SNPs. In the case of "All 5", a combined SBM-based SNP p-value was used. A geneset is a predefined set consisting of genes which are themselves similar or related in some regard with each other. DEPICT uses its own reconstituted genesets which have been created by extending existing genesets with further predicted members. The FDR threshold is < 0.01 . The p-value limits listed indicate the highest p-values under this FDR threshold.

A closer investigation of the distribution of the edges (see Figure 3.17 and table 3.7) showed that interactions between genes of equal diseases status are much rarer than interactions between genes with different diseases status. Nevertheless, the subnetworks consisting of nodes with equal disease counts have been created (Figures 3.22 to 3.18) and even though they are mostly unconnected, there are some interesting connected components in them. For instance, figure 3.18A shows a component that has Jak-Stat signaling-related genes as well as other signaling-related genes.

3.2. Geneset-based networks

A completely different approach to obtain networks from GWAS data is based on the tool DEPICT⁵¹. DEPICT tries to link lead SNPs to genes with the help of predefined genesets. See the section 2.3 for details on the method and the used parameters.

DEPICT was run with all 244 SNPs from the cross-disease dataset (SBM p-value) but it was also run with all subsets of SNPs that are specific for each single disease.

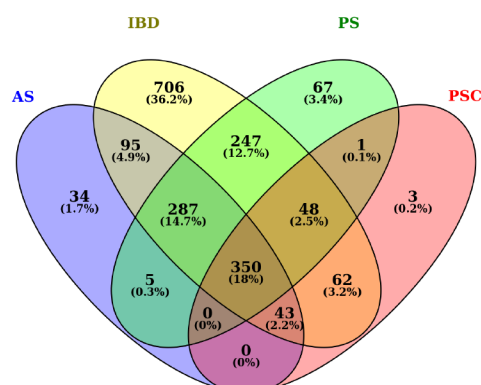


Figure 3.23.: Overlap of significant geneset terms between different diseases according to DEPICT.

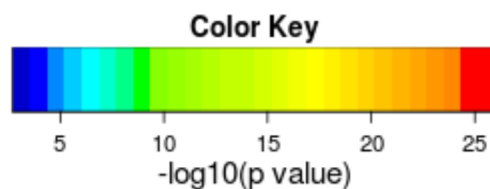


Figure 3.24.: Heatmap color key for the figures 3.25 to 3.29 and the figures 3.31 to 3.35. The colors correspond to the DEPICT scores of individual genes leading to enrichment in the heatmap cells.

3.2.1. Geneset enrichment

DEPICT reported more than 500 FDR-significant genesets for every single disease and for the combination of all five diseases together (table 3.8). Some of these enriched genesets are shared between the different diseases. Figure 3.23 gives an overview on the number of shared genesets.

In order to better understand what is specific and what is shared between diseases, the enriched genesets for every disease have been sorted by p-value and have been separated into one of the following categories: *specific for one disease*, *specific for 2 diseases*, ..., *specific for 5 diseases*. The ten most significant reconstituted genesets for every disease and for every category have been gathered and have been visualized in heatmaps (Figure 3.25 to Figure 3.29).

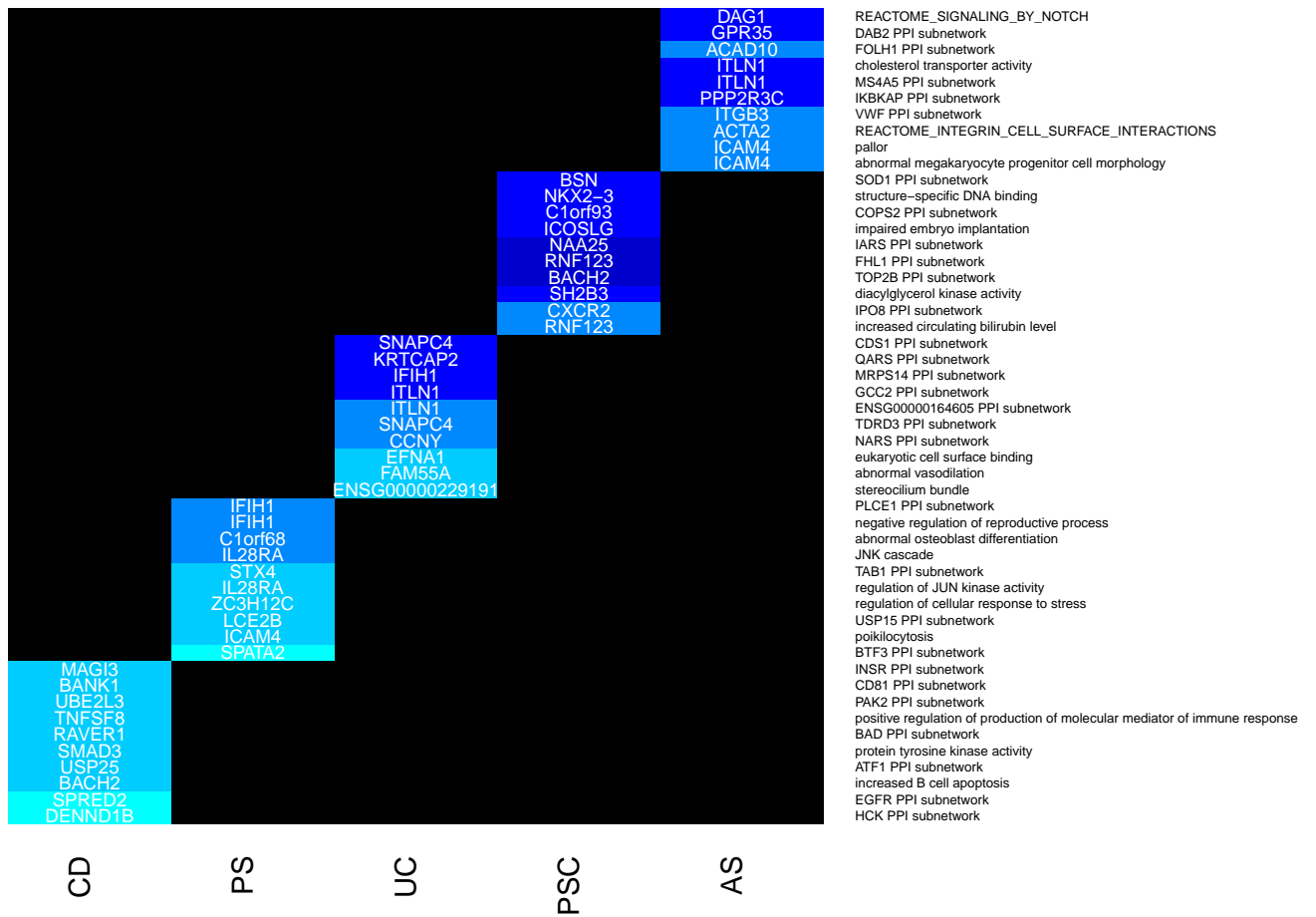


Figure 3.25.: Enriched genesets for every disease that are specific for a single disease (according to DEPICT). The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the most to the enrichment of the term in a disease.

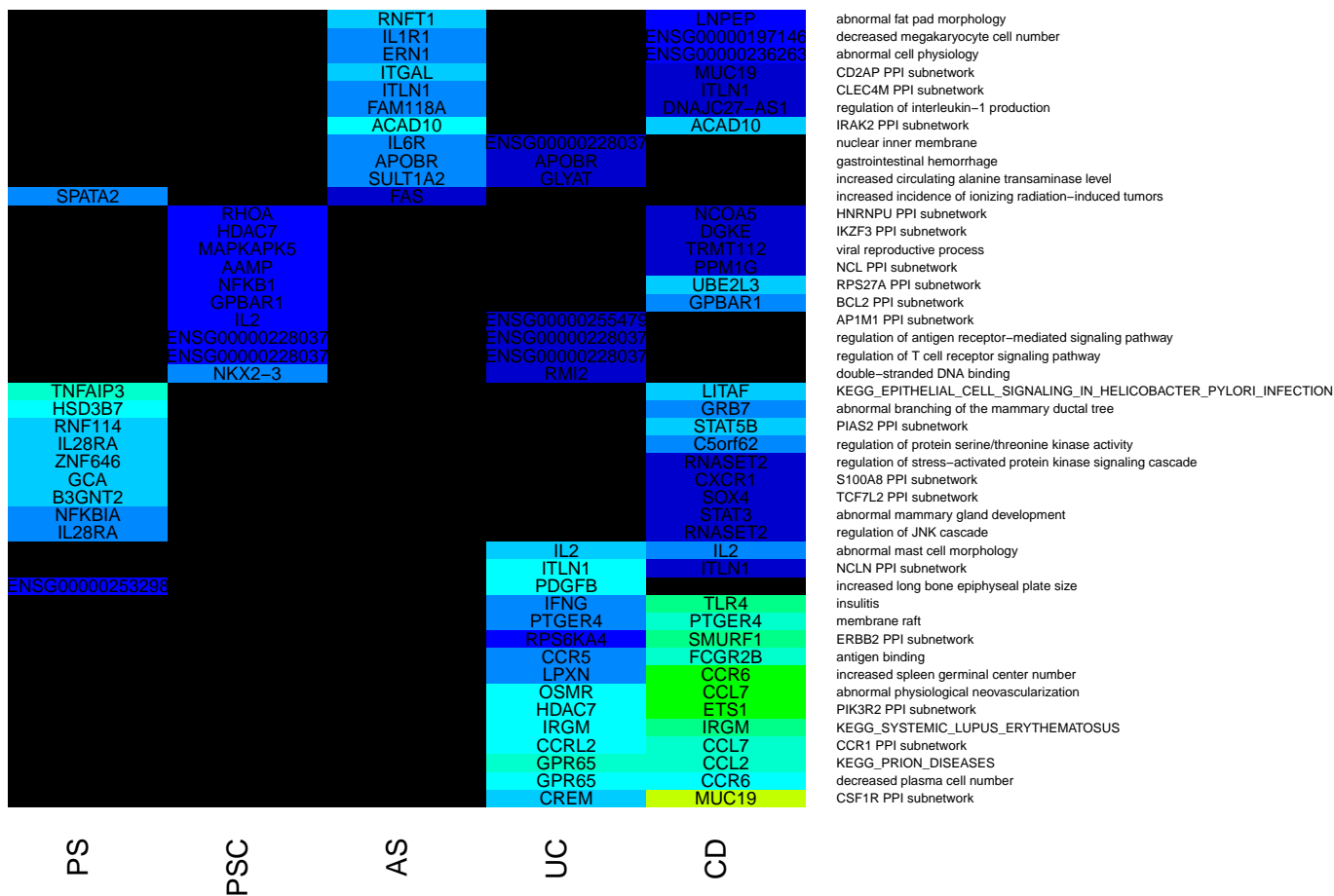


Figure 3.26.: Enriched genesets for every disease that are specific for exactly 2 diseases (according to DEPICT). The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the most to the enrichment of the term in a disease.

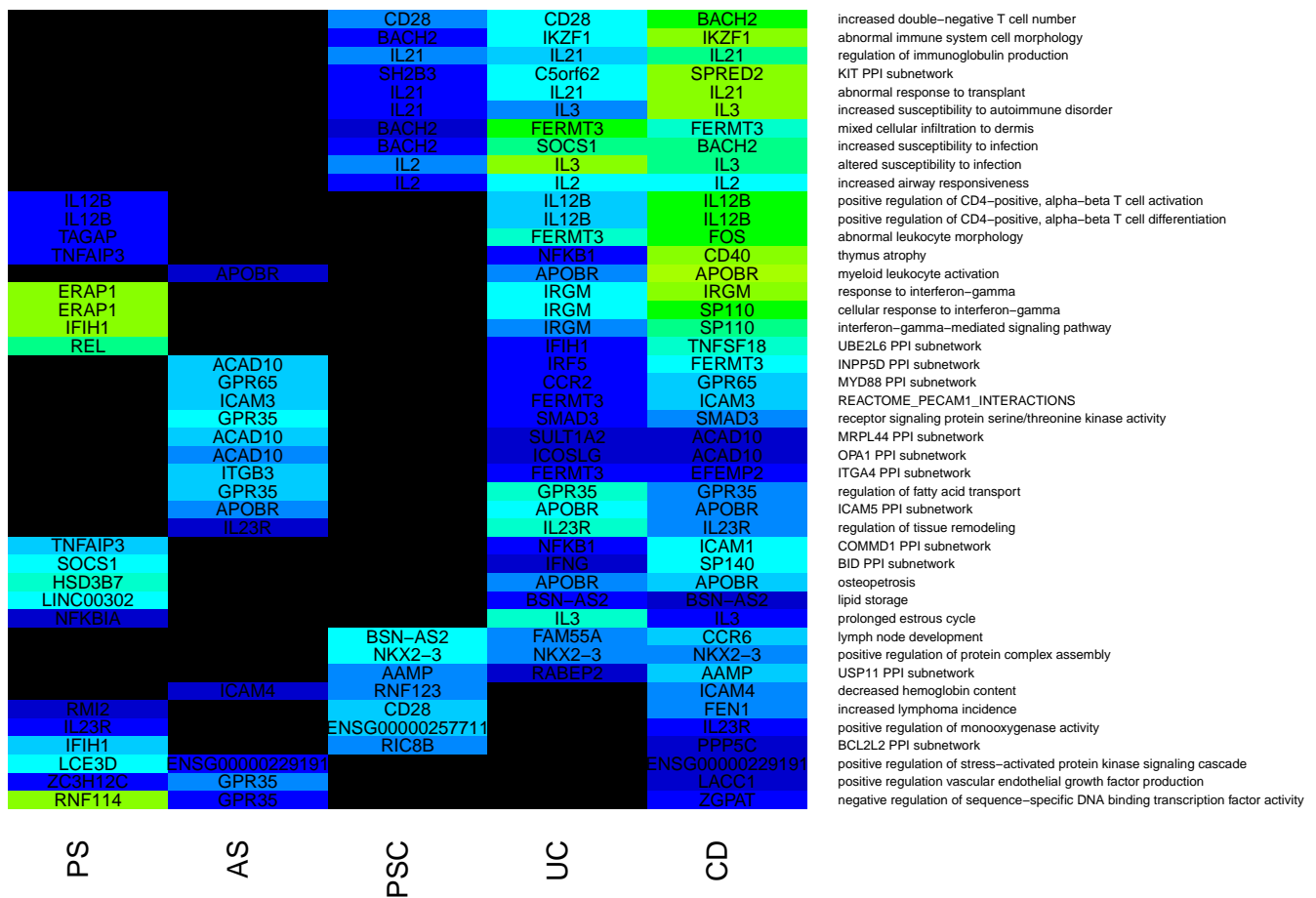


Figure 3.27.: Enriched genesets for every disease that are specific for exactly 3 diseases (according to DEPICT). The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the most to the enrichment of the term in a disease.

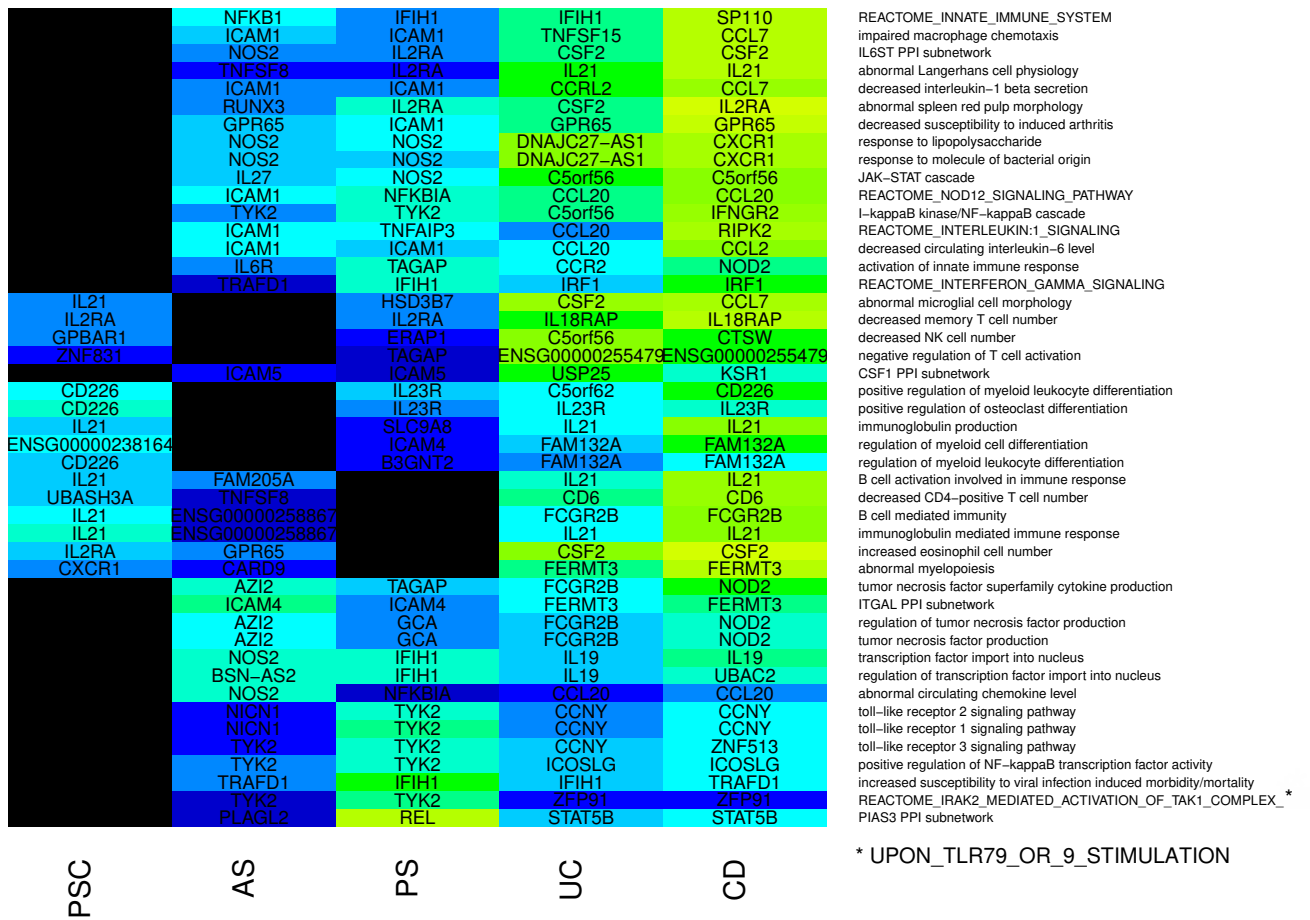


Figure 3.28.: Enriched genesets for every disease that are specific for exactly 4 diseases (according to DEPICT). The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the most to the enrichment of the term in a disease.

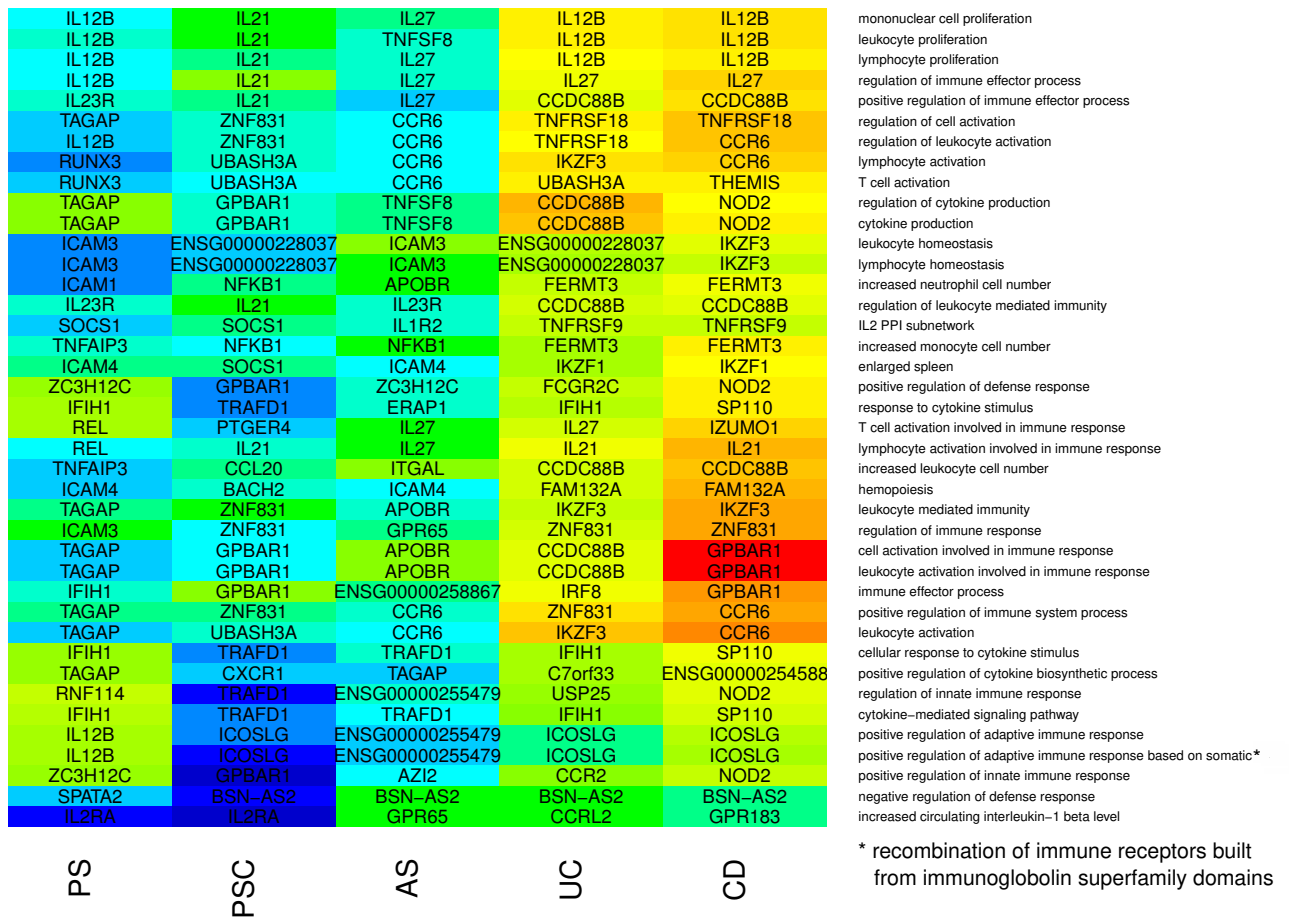


Figure 3.29.: Enriched genesets for every disease that are specific for exactly 5 diseases (according to DEPICT). The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the most to the enrichment of the term in a disease.

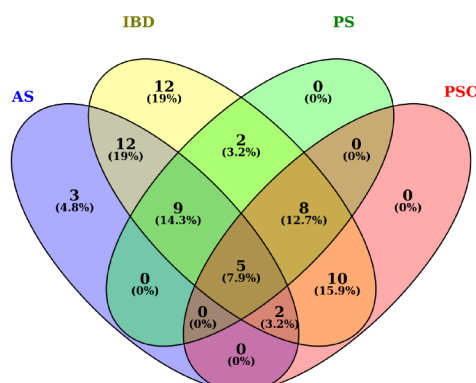


Figure 3.30.: Overlap of significant tissues between different diseases according to DEPICT.

Disease	Number of SNPs	Number of FDR-significant tissues	p-values
All 5	244	65	≤ 0.01
CD	185	57	$\leq 4.02^{-3}$
UC	164	58	$\leq 5.68^{-3}$
AS	113	32	$\leq 1.06^{-3}$
PS	66	25	$\leq 5.77^{-4}$
PSC	61	26	$\leq 2.06^{-3}$

Table 3.9.: Numbers of enriched tissues determined with DEPICT given different sets of SNPs. In the case of "All 5", a combined SBM-based SNP p-value was used. The FDR threshold is < 0.01 .

3.2.2. Tissue enrichment

In addition to geneset enrichment, DEPICT performs tissue enrichment to determine which tissues are most relevant for the set of genes determined by DEPICT. For every disease more than 25 tissues were identified as FDR-significantly enriched (table 3.9). The overlap of tissues among different diseases is shown in figure 3.30. Similar to the enriched genesets described previously, it was determined which tissues are relevant in exactly 1, 2, 3, 4 or 5 diseases. The most significant of these tissues have been plotted as heatmaps in figure 3.31 to figure 3.35.

Because some of the tissues are shared between diseases and because there is a limited number of FDR-significant tissues for every disease, not every heatmap shows $5 \times 10 = 50$ tissues.

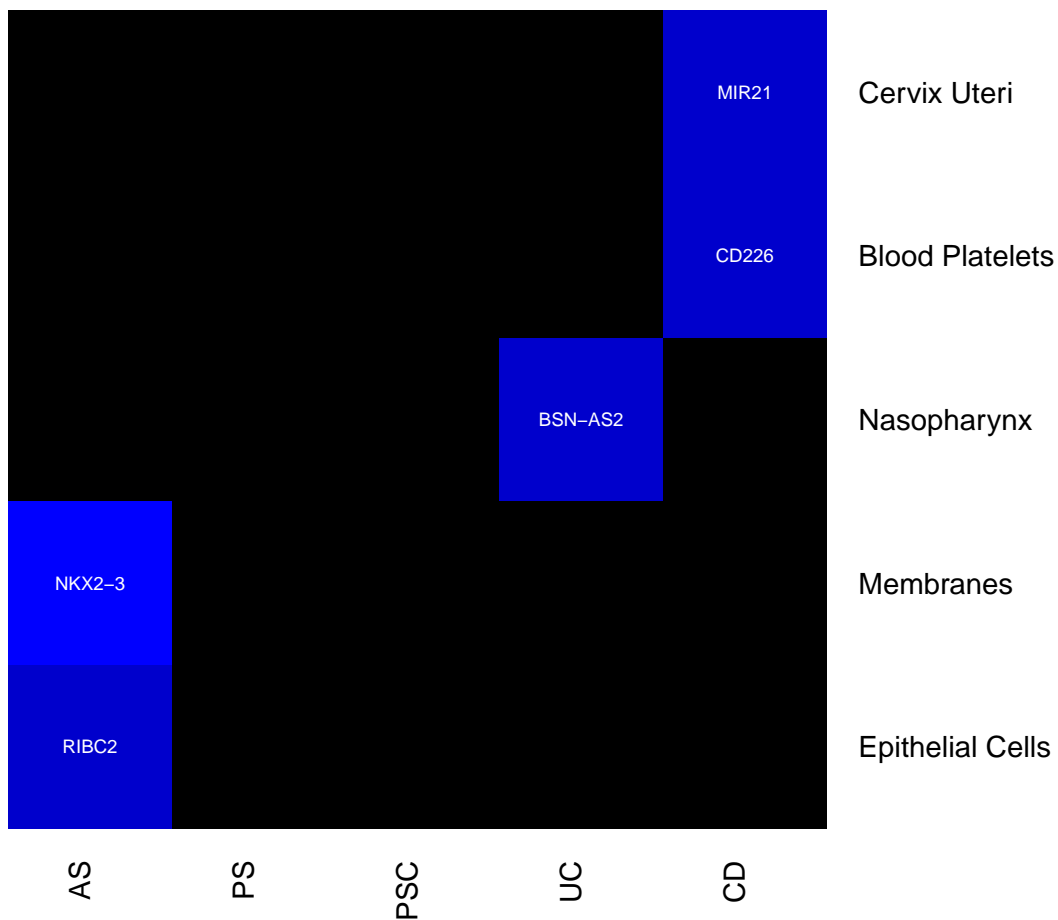


Figure 3.31.: Enriched tissues for every disease that are specific for a single disease. No significant tissues were detected that are exclusively specific for psoriasis or PSC. The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the strongest signal to the enrichment of the tissue in a disease.

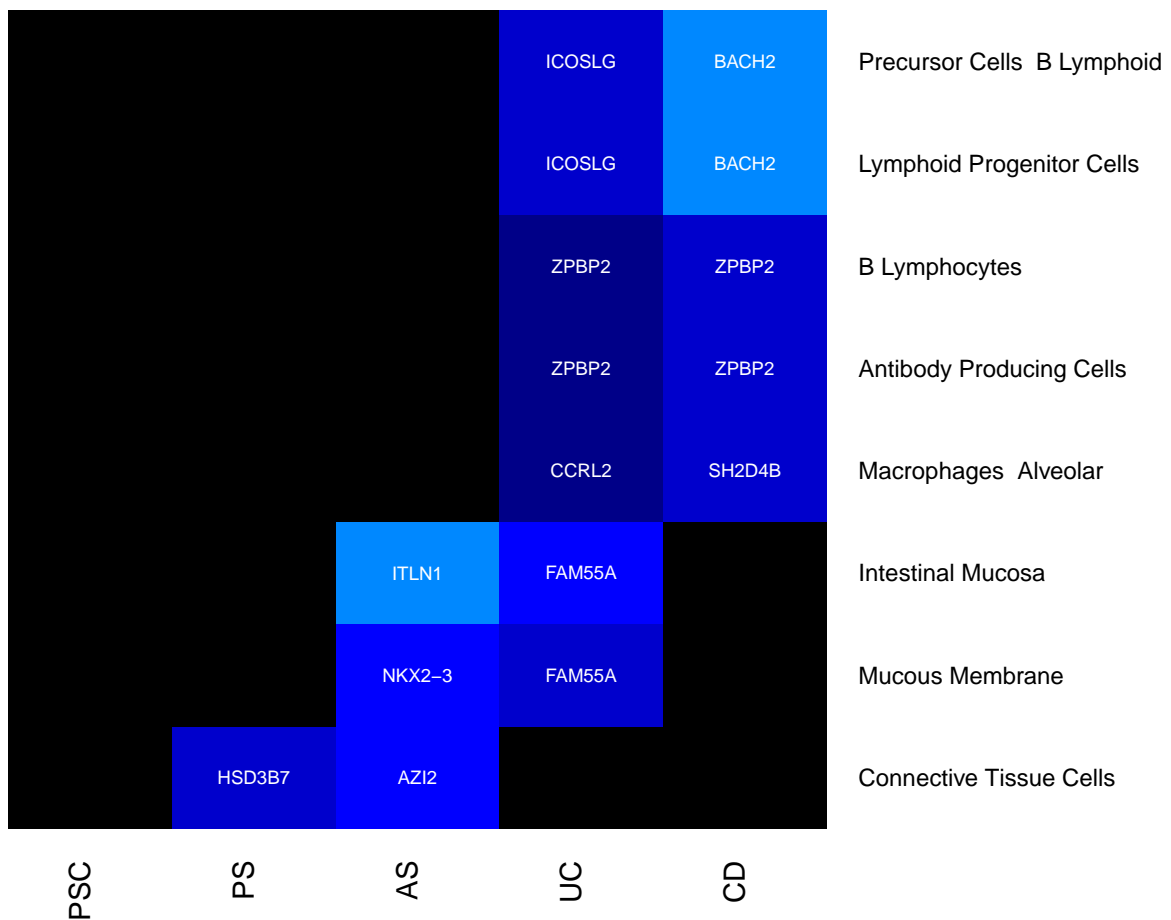


Figure 3.32.: Enriched tissues for every disease that are specific for a exactly 2 diseases. The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the strongest signal to the enrichment of the tissue in a disease.

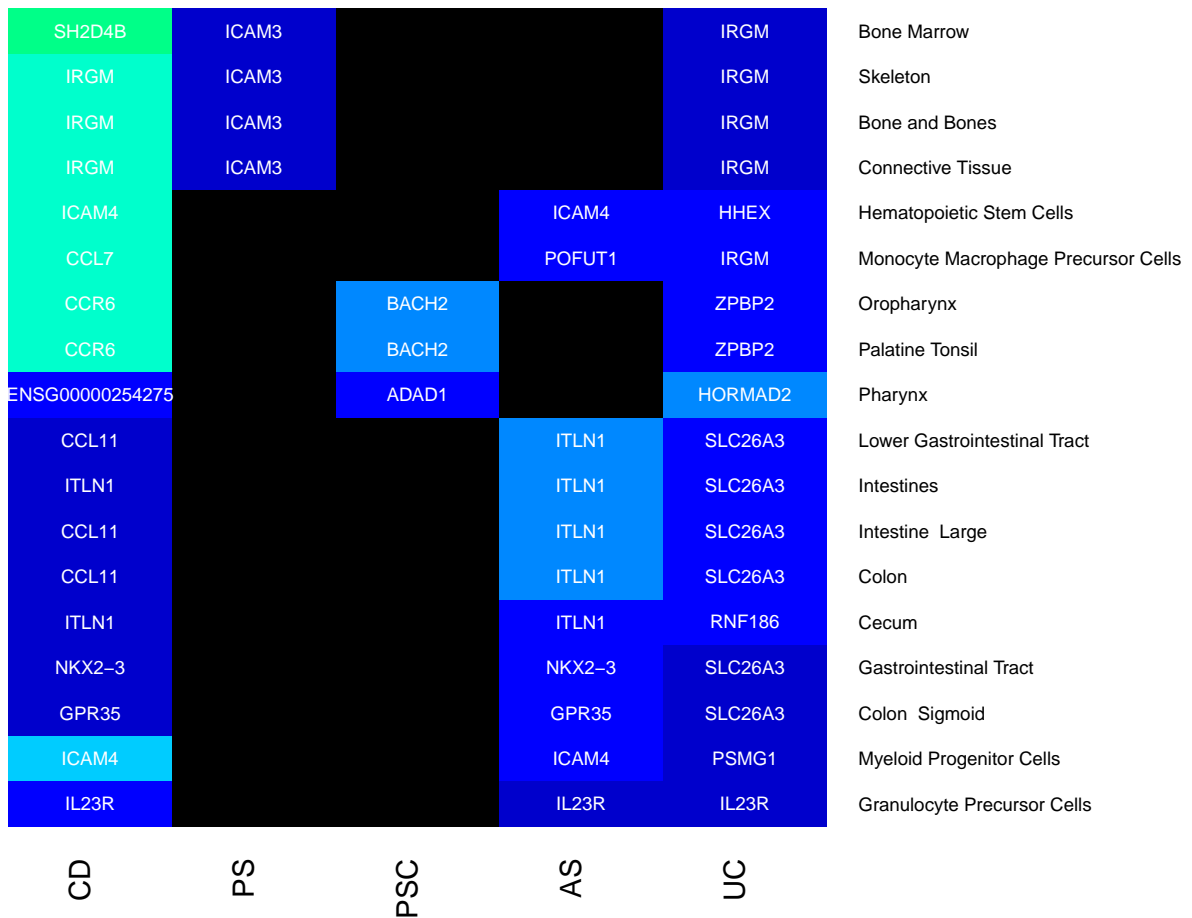


Figure 3.33.: Enriched tissues for every disease that are specific for a exactly 3 diseases. The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the strongest signal to the enrichment of the tissue in a disease.

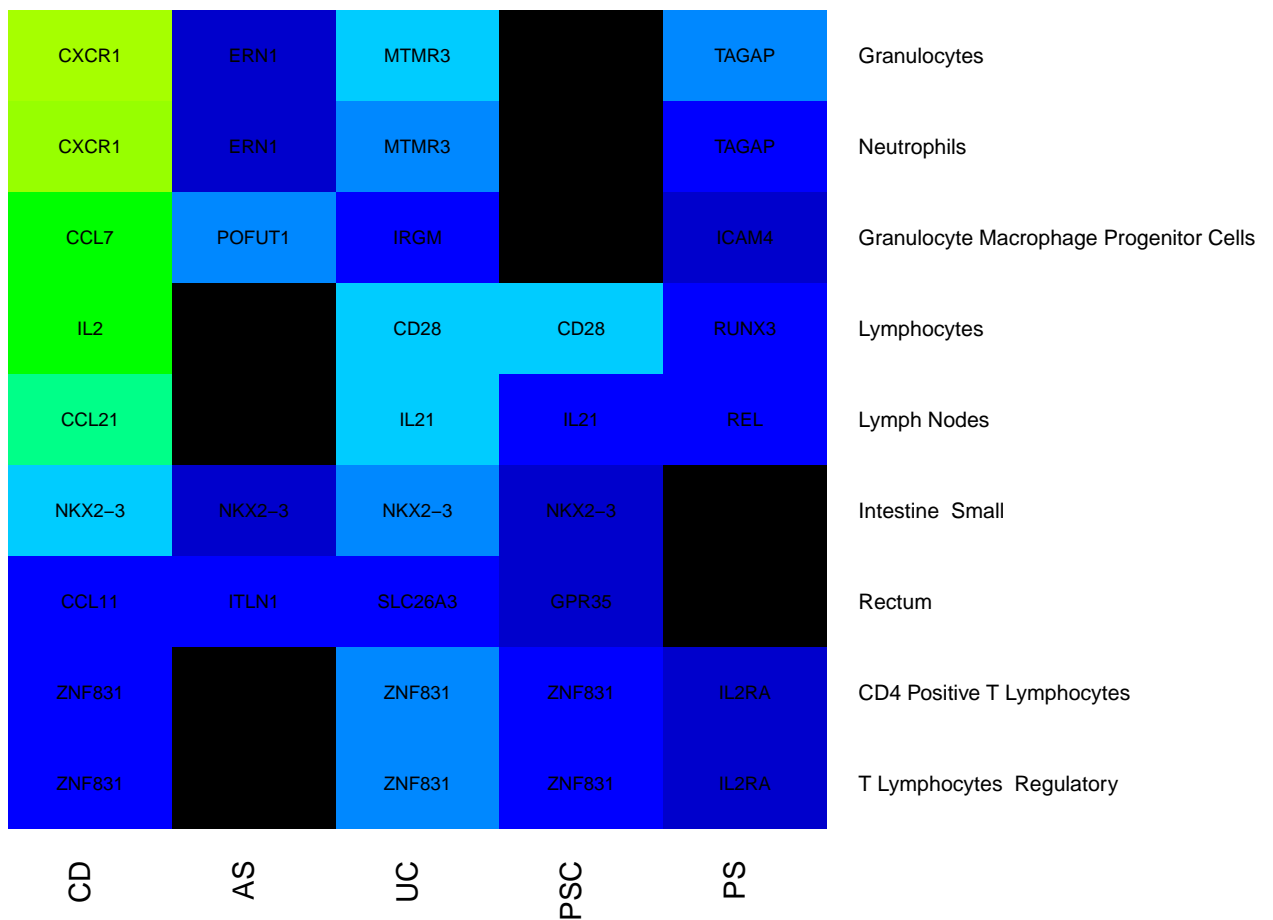


Figure 3.34.: Enriched tissues for every disease that are specific for a exactly 4 diseases. The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the strongest signal to the enrichment of the tissue in a disease.

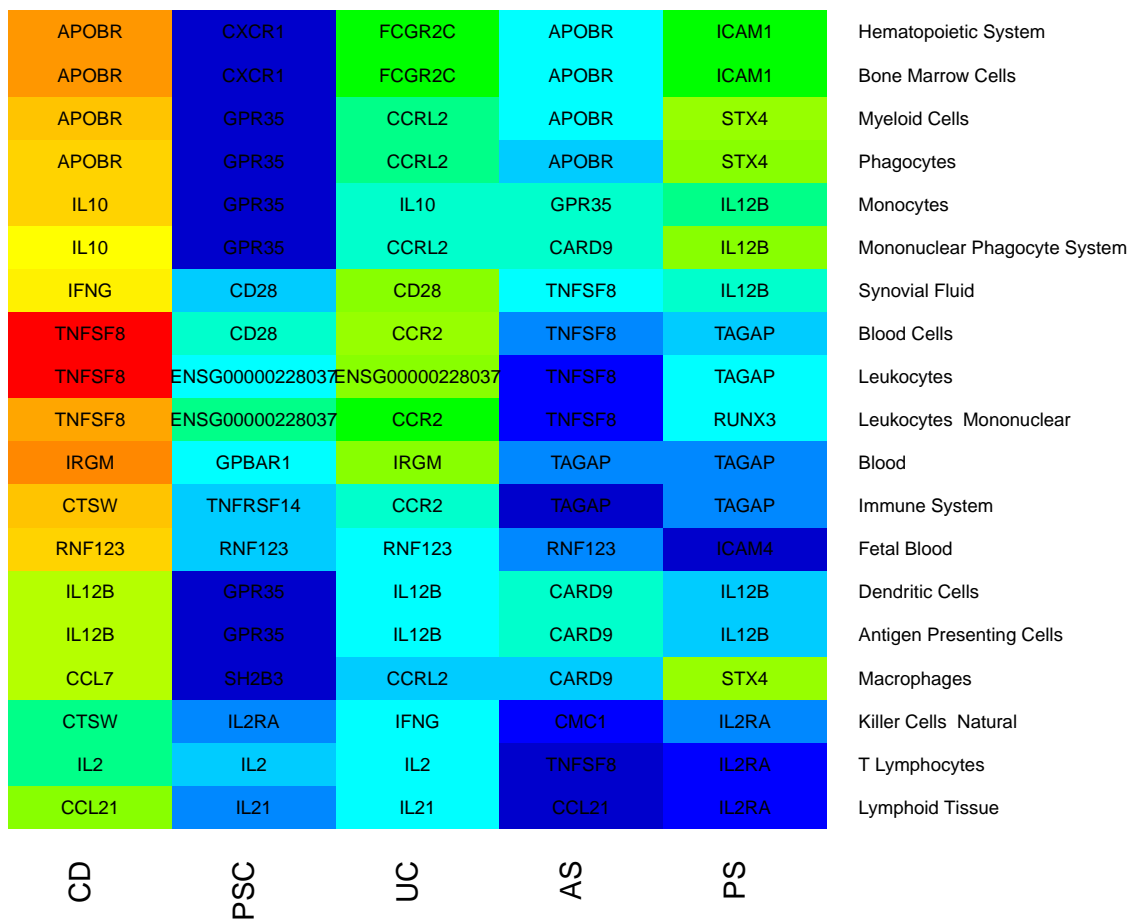


Figure 3.35.: Enriched tissues for every disease that are specific for a exactly 5 diseases. The colors correspond to the legend in figure 3.24. The heatmap cells are annotated with the gene that contributes the strongest signal to the enrichment of the tissue in a disease.

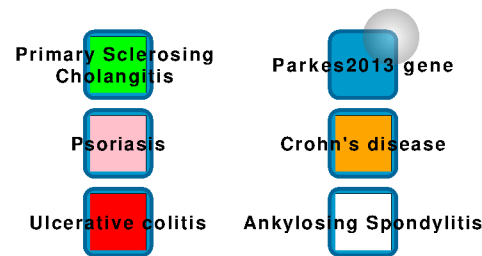


Figure 3.36.: Color legend for genes that are associated with different diseases. In addition, any gene that is listed in the paper by Parkes et al. from the year 2013¹⁰ has a grey bubble attached to the node

3.2.3. Geneset-based networks

Based on both, the enriched genesets and the enriched tissues, DEPICT compiles a list of genes that should reflect the genes that are affected by the input SNPs.

Based on the genes listed by DEPICT, network were created in Cytoscape based on the known interactions in the ConsensusPathDB version 30 database with each interaction requiring at least a confidence value of 95 percent. Genes without such high-confidence interactions are not included in the final networks. Overall it can be observed that the network based on the SBM p-values are very similar to the network based on disease-specific p-values: The SBM network contains seven additional nodes and the disease-specific p-value network contains three additional nodes. The largest connected components of both networks are identical (figure 3.37).

To evaluate how well the results from DEPICT align with the literature, the genelists from the paper by Parkes et al.¹⁰ were compared with the network. Parkes et al. describe genes that are associated with ankylosing spondylitis, Crohn's disease, ulcerative colitis, psoriasis and other diseases (but not PSC). These disease genes have not been determined in a systematic manner but were listed based on the author's knowledge. Overall 34 of the 37 disease genes from Parkes et al. have been prioritised by DEPICT. The only genes not predicted by DEPICT are IRF4, ICOS and FCGR2A. However, ICOSLG has been detected which is the ligand of ICOS⁸⁷. This observation is the same regardless of whether SBM p-values or disease-specific p-values are used. All three of the missing genes are related to immune system regulation.

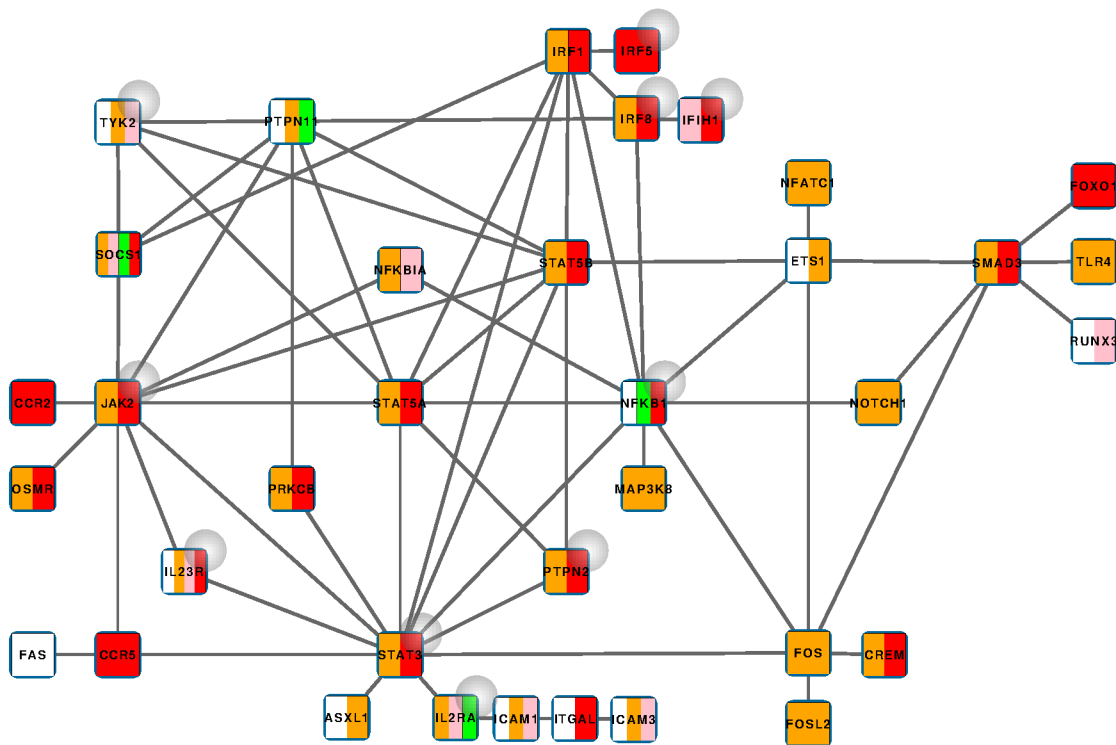


Figure 3.37.: Largest connected component of network based on DEPICT results (identical for disease-specific p-values and SBM p-values). The node colors correspond to the diseases associated with the genes. See figure 3.36 for the color coding.

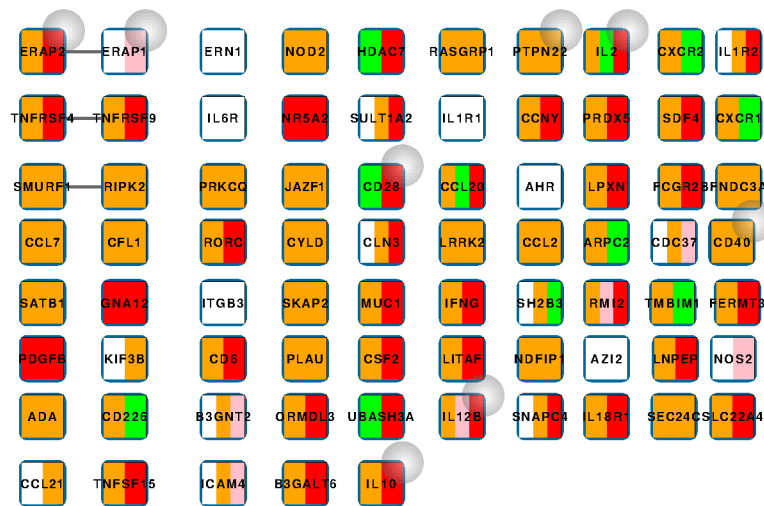


Figure 3.38.: Subnetwork of nodes that are not in the largest connected component. The network is based on the results of running DEPICT on the disease-specific p-values of the SNPs. In contrast to the results obtained from using the SBM p-values, this subnetwork contains the additional nodes LNPEP, PDGFB and ERAP2. The subnetwork based on the SBM p-values (not shown) contains the additional nodes EFNA1, ERFFI1, IL28RA, SETD1A, SLC9A8 STX4 and TNFAIP3. The node colors correspond to the diseases associated with the genes. See figure 3.36 for the color coding.

3.2.4. Enrichment Analyses

To better understand the networks on a biological level, the functional description of every protein in figure 3.37 was looked up on UniProt. It was observed that almost all of the proteins had something to do with regulation or signaling. Table 3.10 lists details about the proteins. A systematic analysis of all proteins in the network using DAVID also resulted in many enriched genesets that have something to do with regulation or signaling (details not shown).

Table 3.10.: Descriptions of proteins in the largest connected component in the DEPICT-based network (Figure 3.37). The functional descriptions were taken from uniprot.org

Protein	functional category
TYK2_HUMAN	Probably involved in intracellular signal transduction by being involved in the initiation of type I IFN signaling
ITAL_HUMAN	[...] receptor [...] It is involved in a variety of immune phenomena including leukocyte-endothelial cell interaction, cytotoxic T-cell mediated killing, and antibody dependent killing by granulocytes and monocytes.
NOTC1_HUMAN	regulate[s] cell-fate determination
IRF1_HUMAN	Transcriptional regulator which displays a remarkable functional diversity in the regulation of cellular responses.
OSMR_HUMAN	Capable of transducing OSM-specific signaling events.
IKBA_HUMAN	Inhibits the activity of dimeric NF κ B/REL complexes by trapping REL dimers

(continued on next page)

(continued from previous page)

FOXO1_HUMAN	Transcription factor that is the main target of insulin signaling and regulates metabolic homeostasis in response to oxidative stress.
IFIH1_HUMAN	Innate immune receptor [...] plays a major role in sensing viral infection and in the activation of a cascade of antiviral responses including the induction of type I interferons and proinflammatory cytokines .
CCR5_HUMAN	Receptor for a number of inflammatory CC-chemokines
NFAC1_HUMAN	Plays a role in the inducible expression of cytokine genes in T-cells
TNR6_HUMAN	Receptor for TNFSF6/FASLG. [...] The resulting death- inducing signaling complex (DISC)
CCR2_HUMAN	Receptor for the CCL2, CCL7 and CCL13 chemokines .
STAT3_HUMAN	Signal transducer and transcription activator
CREM_HUMAN	Transcriptional regulator that binds the cAMP response element (CRE), a sequence present in many viral and cellular promoters. Isoforms are either transcriptional activators or repressors.
PTN11_HUMAN	Acts downstream [...] to participate in the signal transduction

(continued on next page)

(continued from previous page)

M3K8_HUMAN	[...] activation of the MAPK/ERK pathway in macrophages, thus being critical for production of the proinflammatory cytokine TNF-alpha (TNF) during immune responses.
ETS1_HUMAN	Directly controls the expression of cytokine and chemokine genes in a wide variety of different cellular contexts
IRF8_HUMAN	Plays a negative regulatory role in cells of the immune system
KPCB_HUMAN	involved in various cellular processes such as regulation of the B-cell receptor (BCR) signalosome
JAK2_HUMAN	Mediates essential signaling events in both innate and adaptive immunity
TLR4_HUMAN	Cooperates with LY96 and CD14 to mediate the innate immune response [...] Acts via MYD88 [...], leading to NF κ B activation, cytokine secretion and the inflammatory response
STA5B_HUMAN	Carries out a dual function: signal transduction and activation of transcription
FOSL2_HUMAN	Controls osteoclast survival and size. As a dimer with JUN, activates LIF transcription

(continued on next page)

(continued from previous page)

ICAM1_HUMAN	ICAM1 engagement promotes the assembly of endothelial apical cups through ARHGEF26/SGEF and RHO G activation
ASXL1_HUMAN	involved in transcriptional regulation mediated by ligand -bound nuclear hormone receptors
STA5A_HUMAN	Carries out a dual function: signal transduction and activation of transcription
SOCS1_HUMAN	SOCS family proteins form part of a classical negative feedback system that regulates cytokine signal transduction
IRF5_HUMAN	involved in the induction of interferons IFNA and INFB and inflammatory cytokines upon virus infection
PTN2_HUMAN	Negatively regulates numerous signaling pathways and biological processes like hematopoiesis, inflammatory response, cell proliferation and differentiation, and glucose homeostasis.
SMAD3_HUMAN	Receptor-regulated SMAD (R-SMAD) that is an intracellular signal transducer and transcriptional modulator activated by TGF-beta (transforming growth factor) and activin type 1 receptor kinases.

(continued on next page)

(continued from previous page)

IL23R_HUMAN	Binds IL23 and mediates T-cells, NK cells and possibly certain macrophage/myeloid cells stimulation probably through activation of the Jak-Stat signaling cascade .
ICAM3_HUMAN	ICAM proteins are ligands for the leukocyte adhesion protein LFA-1 (integrin α -L/ β -2). ICAM3 is also a ligand for integrin α D/ β 2.
FOS_HUMAN	On TGF-beta activation, [...] to regulate TGF-beta- mediated signaling. [...] It is thought to have an important role in signal transduction .
RUNX3_HUMAN	binds to the core site [...] of a number of enhancers and promoters, including T-cell receptor enhancers
IL2RA_HUMAN	Receptor for interleukin-2. → IL2: this protein is required for T-cell proliferation and other activities crucial to regulation of the immune response
NFKB1_HUMAN	is the endpoint of a series of signal transduction events that are initiated by a vast array of stimuli related to many biological processes such as inflammation

3.2.5. Disease-specific subnetworks

To get a clearer picture of the parts of the network that are specific to a disease, the largest connected component was filtered down for every disease to only contain nodes that are associated with the disease in question and neighboring nodes that are directly linked to such disease-nodes. In the case of Crohn's disease and Ulcerative colitis no

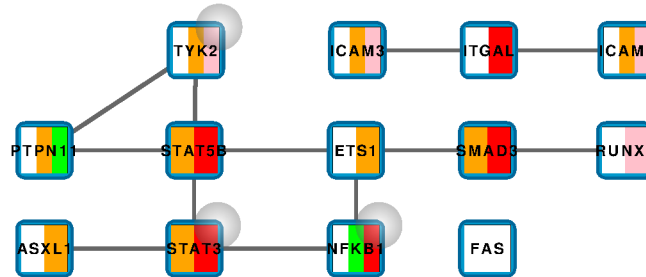


Figure 3.39.: Subnetwork for ankylosing spondylitis (with linker genes). AS-associated nodes have a white color component.

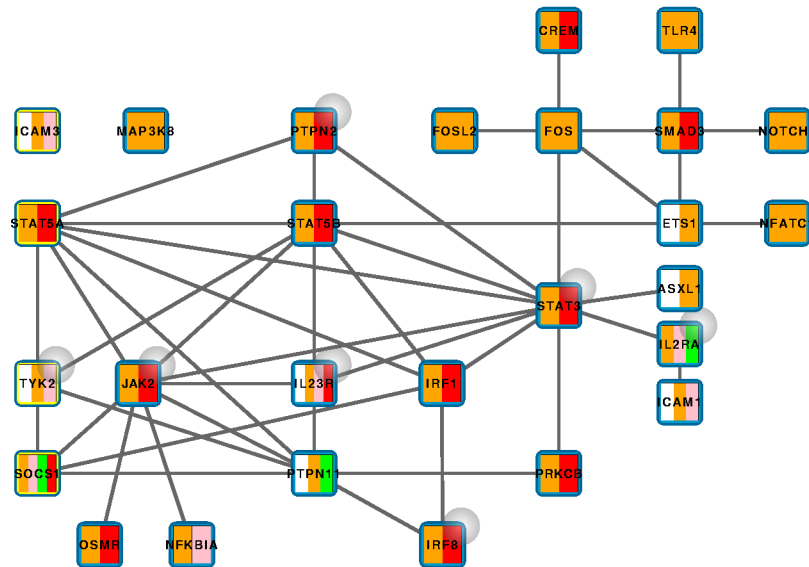


Figure 3.40.: Subnetwork for Crohn's disease (without linker genes).

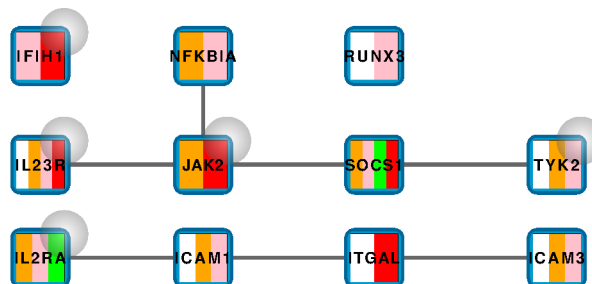


Figure 3.41.: Subnetwork for psoriasis (with linker genes). PS-associated nodes have a pink color component.

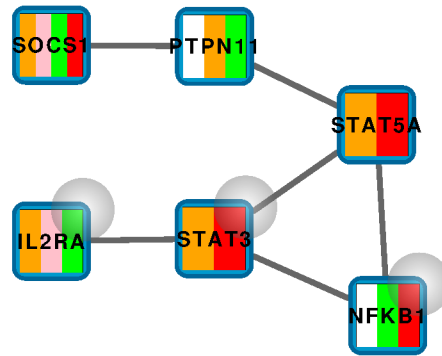


Figure 3.42.: Subnetwork for primary sclerosing cholangitis (with linker genes). PSC-associated nodes have a green color component.

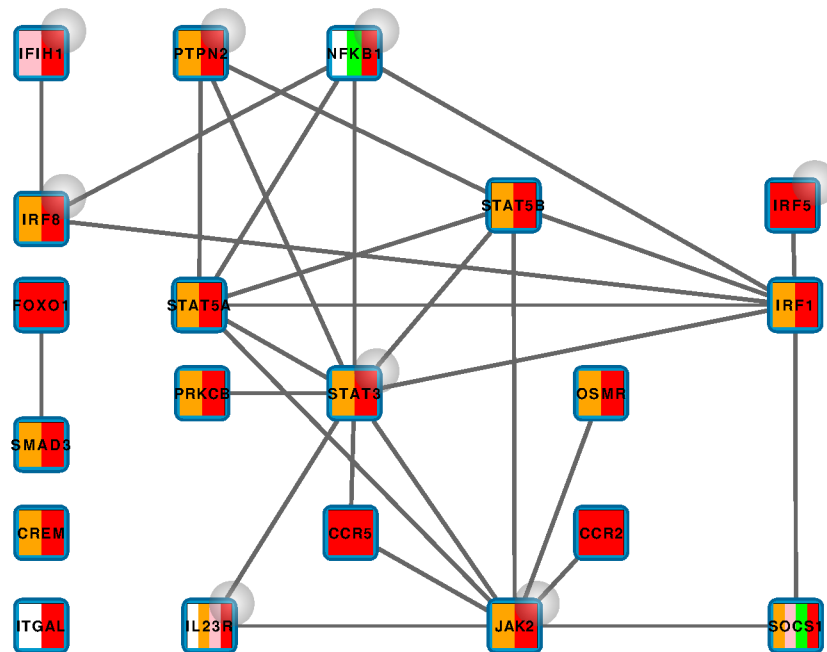


Figure 3.43.: Subnetwork for ulcerative colitis (without linker nodes).

Disease	Disease nodes	Neighbours	Total network size
ankylosing spondylitis	29	382	411
Crohn's disease	88	825	913
psoriasis	16	186	202
primary sclerosing cholangitis	15	209	224
ulcerative colitis	56	625	681

Table 3.11.: Number of disease-associated nodes and non-disease linker nodes for every disease (according to DEPICT). A non-disease linker node is a neighbour node of a disease node and may be associated with another disease.

neighbors were included because of the already great number of directly associated genes. The subnetworks can be seen in figures 3.39 to 3.43.

3.2.6. Linker nodes

In order to discover new potential disease genes, all nodes associated with a specific disease (according to DEPICT) were taken and the direct neighbours in the global protein interaction network (CPDB version 30) were taken as linker nodes to create new subnetworks. There are a great number of linker genes for every disease subnetwork. The sizes of these networks are listed in table 3.11.

In order to prioritise potential new disease genes, each of the linker node was given a score that is defined as the number of neighbours that are already associated with the disease in question, divided by the total number of neighbours. The division operation was performed to account for hub genes which have a great number of interactors and therefore a single interaction with a disease gene is not notable.

Tables 3.12 to 3.16 show the best-scoring non-disease genes/proteins. Most of the best-scoring linker genes were not associated with any of the five diseases which makes them interesting candidates for further analysis. However, only few of the nodes with a score of 0.5 or better had more than one disease-associated neighbour. Notable exceptions are IL23A (Crohn's disease, ulcerative colitis and psoriasis, with two of three neighbour genes) and IRF6 (ulcerative colitis, two out of two neighbouring genes).

Protein	disease neighbours/total neighbours (score)	Other diseases
IL1RA_HUMAN	1/1	
ERAP2_HUMAN	1/1	CD, UC
LEUK_HUMAN	1/1	
AFG32_HUMAN	1/2	
ICAM5_HUMAN	1/2	
CNTF_HUMAN	1/2	
TMM33_HUMAN	1/2	
ITAD_HUMAN	1/2	
RPIA_HUMAN	1/2	
DERL3_HUMAN	1/2	
AL1A1_HUMAN	1/2	
ARHGA_HUMAN	1/2	
IL23A_HUMAN	1/3	
IL6_HUMAN	1/3	
CXL13_HUMAN	1/3	
RHBT2_HUMAN	1/3	
UBA7_HUMAN	1/3	
THY1_HUMAN	1/3	
...		

Table 3.12.: Nodes linked to AS-associated genes. Each node has a score that describes the ratio of neighbours that are associated with AS to the total number of neighbours. Only the nodes with the best scores are shown.

Protein	disease neighbours/total neighbours (score)	Other diseases
IL1RA_HUMAN	1/1	
PAI2_HUMAN	1/1	
TNFL4_HUMAN	1/1	
SIA7B_HUMAN	1/1	
LIMK2_HUMAN	1/1	
STAC_HUMAN	1/1	
NDUF7_HUMAN	1/1	
TPC10_HUMAN	1/1	
KCNC4_HUMAN	1/1	
GALT4_HUMAN	1/1	
TAOK3_HUMAN	1/1	
MPC2_HUMAN	1/1	
ERAP1_HUMAN	1/1	AS, PS
DLL3_HUMAN	1/1	
AGAP2_HUMAN	1/1	
GLT15_HUMAN	1/1	
LEUK_HUMAN	1/1	
NOX4_HUMAN	1/1	
IL23A_HUMAN	2/3	
...		

Table 3.13.: Nodes linked to CD-associated genes. Each node has a score that describes the ratio of neighbours that are associated with CD to the total number of neighbours. Only the nodes with the best scores are shown.

Protein	disease neighbours/total neighbours (score)	Other diseases
ERAP2_HUMAN	1/1	CD, UC
LEUK_HUMAN	1/1	
IL23A_HUMAN	2/3	
ITAD_HUMAN	1/2	
AL1A1_HUMAN	1/2	
I12R1_HUMAN	2/5	
ITAL_HUMAN	2/6	AS, UC
IL12A_HUMAN	1/3	
RHBT2_HUMAN	1/3	
NLRC5_HUMAN	1/3	
C1GLC_HUMAN	1/4	
MATK_HUMAN	1/4	
ITPK1_HUMAN	1/4	
ITB2_HUMAN	3/18	
ITAM_HUMAN	2/12	
...		

Table 3.14.: Nodes linked to PS-associated genes. Each node has a score that describes the ratio of neighbours that are associated with PS to the total number of neighbours. Only the nodes with the best scores are shown.

Protein	disease neighbours/total neighbours (score)	Other diseases
IFNA1_HUMAN	1/2	
RPIA_HUMAN	1/2	
CD80_HUMAN	1/3	
CD86_HUMAN	1/3	
GNA14_HUMAN	2/8	
ADA1A_HUMAN	2/8	
PVRL2_HUMAN	1/5	
CBS_HUMAN	1/6	
PVR_HUMAN	1/6	
ATP5J_HUMAN	1/6	
NSF_HUMAN	1/6	
SOCS2_HUMAN	1/7	
TGFB3_HUMAN	1/7	
GNA15_HUMAN	2/15	
...		

Table 3.15.: Nodes linked to PSC-associated genes. Each node has a score that describes the ratio of neighbours that are associated with PSC to the total number of neighbours. Only the nodes with the best scores are shown.

3.3. Effect directions of Single Nucleotide Polymorphisms

Disease-associated SNPs can increase or decrease the risk of developing the disease. The risk itself is only a probability and there are cases of healthy individuals with high genetic risk scores and also cases of affected individuals with low genetic risk scores (Figure 1.2). The risk scores are based on the frequencies of the alleles in the study populations in healthy versus disease-affected individuals. The odds ratios obtained from these frequencies only provide an estimate on the effect size of a SNP. The underlying biological mechanisms are still unclear. But given that proteins interact with each other to perform their functions, there might be general principles that increase or decrease the quality of these interactions and these general principles might be observable on the network level. A special focus is given on protective variants because it is hypothesized that protective variants might induce buffering in the network against problematic signals.

As described in previous chapters, genetic variants have to be linked to genes to analyze them on the network level. For this part the linking was done on the basis of ENCODE *transcription factor binding sites* (TFBS) annotations. At this point it should be explained that every ENCODE TFBS is annotated with genes. It was originally assumed that these genes would be the transcriptional targets of the TFBS but while writing this thesis it was realised that these genes are the transcription factors or other proteins that bind to DNA at that specific location. Unfortunately these annotations do not contain information about which genes are controlled by the TFBS but they allow us to link SNPs to transcription factors and other DNA-binding proteins. And these proteins in turn might lead to enriched pathways. Because of the initial misunderstanding with the ENCODE annotation files, some of the results presented here might seem out of context. However, every transcription factor is still the product of a gene and therefore the principle of linking SNPs to genes still holds. All results have been corrected to reflect the fact that the ENCODE annotations contain transcription factors but not their target proteins. Even with this change of perspective, there are some interesting results.

The major advantage of these ENCODE TFBS annotations is that they are manually curated and that for every binding site it is always known which proteins bind to it. With the help of the program GoShifter all lead SNPs were mapped to LD SNPs and then the positional overlap of these LD SNPs with transcription factor binding sites was determined.

Protein	disease neighbours/total neighbours (score)	Other diseases
IL1RA_HUMAN	1/1	
RASA2_HUMAN	1/1	
TNFL4_HUMAN	1/1	
SIA7B_HUMAN	1/1	
GTPB1_HUMAN	1/1	
IRF6_HUMAN	2/2	
KCNC4_HUMAN	1/1	
GALT4_HUMAN	1/1	
ERAP1_HUMAN	1/1	AS, PS
AGAP2_HUMAN	1/1	
PCDA4_HUMAN	1/1	
GLT15_HUMAN	1/1	
IL23A_HUMAN	2/3	
AFG32_HUMAN	1/2	
ICAM5_HUMAN	1/2	
TMM33_HUMAN	1/2	
ST2A1_HUMAN	1/2	
TGIF2_HUMAN	1/2	
IFNA1_HUMAN	1/2	
GSC_HUMAN	1/2	
ISG20_HUMAN	1/2	
CC90B_HUMAN	1/2	
GALT1_HUMAN	1/2	
GLT10_HUMAN	1/2	
1A03_HUMAN	1/2	
I12R2_HUMAN	1/2	
INGR2_HUMAN	1/2	
I12R1_HUMAN	2/5	
FGFR3_HUMAN	2/5	
PRDC1_HUMAN	2/5	
...		

Table 3.16.: Nodes linked to UC-associated genes. Each node has a score that describes the ratio of neighbours that are associated with UC to the total number of neighbours. Only the nodes with the best scores are shown.

ENCODE annotation	GoShifter global p-value
Distal H3K4me1 annotations (cell type specific)	0.0033
Distal H3K27ac annotations (cell type specific)	0.0053
Distal H3K4me3 annotations (cell type specific)	0.0143
Distal TF binding sites	0.0145
Proximal TF binding sites	0.0179
Proximal H3K27ac annotations (cell type specific)	0.0465
Proximal H3K4me1 annotations (cell type specific)	0.0499
Proximal H3K4me3 annotations (cell type specific)	0.0634
Distal DNase peaks	0.1025
Proximal DNase peaks	0.1973
Proximal H3K9ac annotations (cell type specific)	0.2102
Distal H3K9 annotations (cell type specific)	0.3174

Table 3.17.: ENCODE annotations and the global enrichment p-values of using GoShifter with these annotations and the 244 cross-disease SNPs

The ENCODE project provides various annotations for genomic regions. Table 3.17 lists all annotations that have been downloaded and tested with GoShifter in combination with the 244 cross-disease SNPs. It could be observed that some histone markers (H3K4me1, H3K27ac, H3K4me3, H3K27ac, H3K4me1) and the proximal and distal transcription factor binding sites have significant enrichments given the 244 cross-disease SNPs. However, the histone marker annotations do not provide information which genes are regulated by the modifications. Therefore the histone markers have not been used for the mapping of SNPs to genes.

Some annotations may be the result of other annotations⁷². For example, DNase-I hypersensitive sites tend to be at the same positions as regulatory regions and exonic sites⁸⁸. In order to account for such colocating annotations, GoShifter also provides a stratified approach, to test annotations for dependence of one another. Figure 3.44 shows that all but one of the significant ENCODE annotations are independent of each other and can be regarded as primary signals for further analysis.

The classification of SNPs into protective and risk variants for every individual disease was taken from the cross-disease project². It should be noted that SNPs can be protective for one disease while at the same time they can be risk-inducing for another.

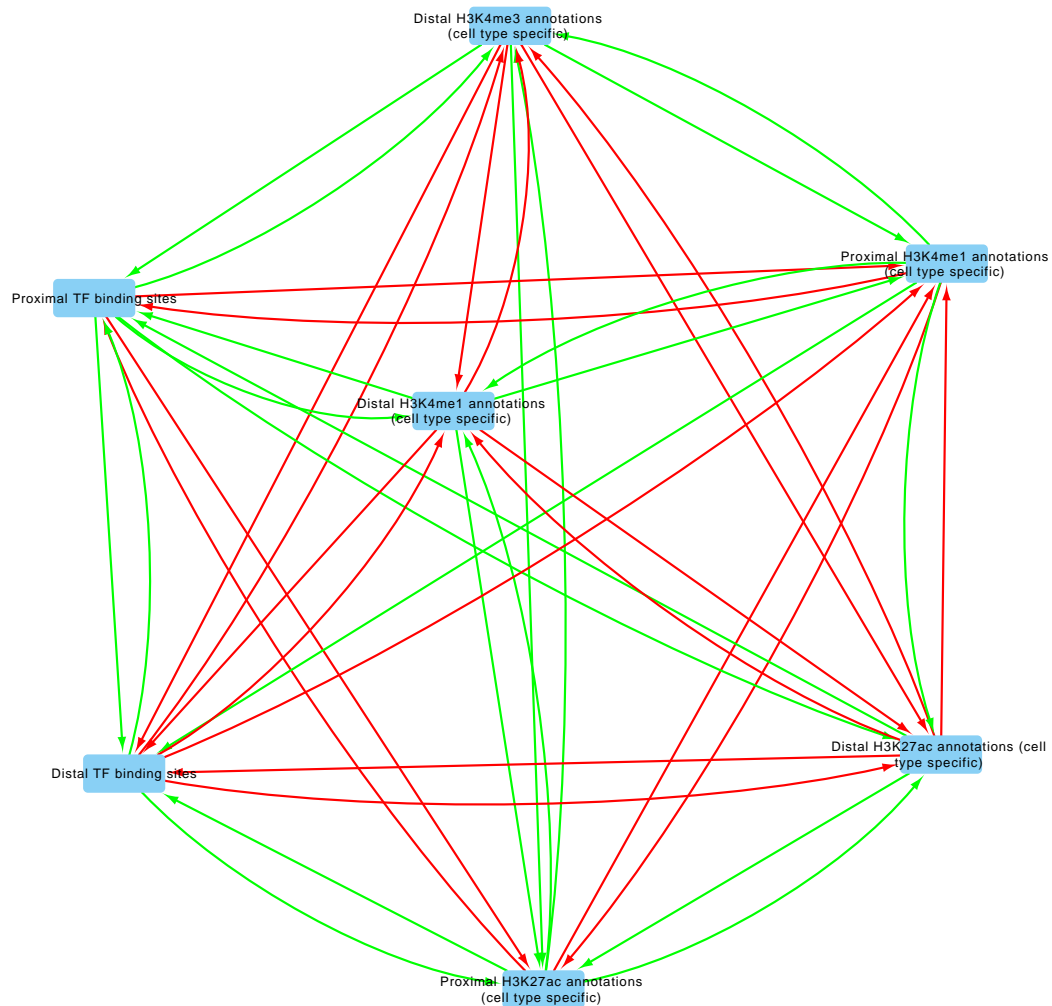


Figure 3.44.: Stratification tests of different annotations with GoShifter. Every annotation that was significant in the normal GoShifter run was tested with all other significant annotations to determine if one type of annotation is dependent on another. This was done for every pair of annotations in both directions. For most pairs of annotations no direction was superior to the other. But for the proximal H3K4me1 annotations there appears to be a dependence on distal transcription factor binding sites and also a dependence on distal H3K27ac modifications. Red arrows indicate dependence and green arrows indicate independence. Arrowheads indicate the primary annotation that was tested for independence.

annotation and risk class	AS	CD	PS	PSC	UC
Proximal TFBS, risk	84	105	58	18	94
Distal TFBS, risk	153	250	92	71	208
Proximal TFBS, protection	60	86	22	33	90
Distal TFBS, protection	100	214	41	63	138
Proximal TFBS, neutral	125	78	189	218	85
Distal TFBS, neutral	293	82	413	412	200

Table 3.18.: Counts of risk classes of LD SNPs overlapping annotations. Neutral LD SNPs are listed for completeness. They were not used in the following steps.

DNA-binding protein risk class	AS	CD	PS	PSC	UC
Bindings to only risk variants	19	4	30	24	11
Bindings to only protective variants	3	5	6	5	4
Bindings to both types of variants	121	139	110	106	136
Bindings to sites without effective variants	11	6	8	19	3

Table 3.19.: Risk classifications (based on the minor allele in the study populations from Ellinghaus et al.²) of DNA-binding proteins for every disease

3.3.1. Mapping of SNPs to genes/DNA binding elements

Because GoShifter does only report which LD SNPs overlap with an annotation but not the annotation itself, separate scripts have been written to perform the backmapping.

No SNP is located within a distal as well in a proximal TFBS. But according to the ENCODE annotations a single TFBS of size 150bp can be regulated by several DNA-binding elements at once. Every LD SNP has only one lead SNP linked to it. There are 73 shared lead SNPs between distal and proximal transcription factor binding sites. Based on the LD SNPs there are 149 different binding elements linked to proximal TFBS and 148 different binding elements linked to distal TFBS. The intersection of these aforementioned genes (binding elements) contains 143 genes and the total number of linked genes is thus 154.

Table 3.18 lists the numbers of SNPs that overlap a TFBS annotation together with their risk status. In any case the number of SNPs in distal TFBSs is greater than the number of SNPs in proximal TFBS. With the minor exception of 4 TFBSs, all TFBS

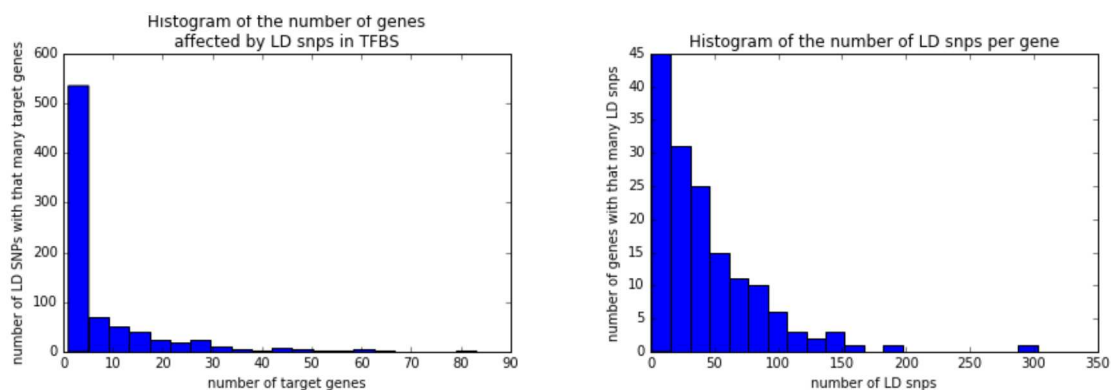


Figure 3.45.: Histograms of the number of mappings from LD SNPs to genes (left) and the number of the backmappings from genes to LD SNPs (right) based on the ENCODE TFBS annotations.

have a size of 150 bp (See section 4.2). The number of non-neutral SNPs in Crohn's disease and ulcerative colitis are the greatest which is probably due to the fact that they have been researched more intensively.

It has been observed that most gene TFBS contain variants that are both, protective and risk-inducing. Only very few genes have only protective variants (Table 3.19).

3.3.2. Relationships between transcriptional regulators and SNPs

Based on the cross-disease lead SNPs there are a total 812 LD SNPs that are linked to 154 different transcriptional regulators based on the ENCODE TFBS annotations. In total there are 6732 links from LD SNPs to genes with 3533 links from the distal TFBSs and 3199 from the proximal TFBSs. The degree distribution of genes to SNPs and the degree distribution of SNPs to genes are shown in figure 3.45.

3.3.3. Purely Protective Binding Factors

Table 3.20.: Best DAVID enrichment results for genes with only protective variants in TFBSs.

Term	term category	#genes	Corrected p-value
transcription regulation	SP_PIR_KEYWORDS	11	7e-08
Transcription	SP_PIR_KEYWORDS	11	4.4e-08
transcription regulator activity	GOTERM_MF_FAT	11	3.1e-07
regulation of transcription	GOTERM_BP_FAT	12	3.8e-06
regulation of transcription, DNA-dependent	GOTERM_BP_FAT	11	2.1e-06
regulation of RNA metabolic process	GOTERM_BP_FAT	11	1.7e-06
nucleus	SP_PIR_KEYWORDS	12	9.5e-07
transcription	GOTERM_BP_FAT	10	0.00015
DNA binding	GOTERM_MF_FAT	10	0.00017
sequence-specific DNA binding	GOTERM_MF_FAT	7	0.00013
dna-binding	SP_PIR_KEYWORDS	8	0.00016
transcription factor activity	GOTERM_MF_FAT	7	0.001
negative regulation of transcription, DNA-dependent	GOTERM_BP_FAT	5	0.0056
negative regulation of RNA metabolic process	GOTERM_BP_FAT	5	0.0052
negative regulation of transcription	GOTERM_BP_FAT	5	0.012
negative regulation of gene expression	GOTERM_BP_FAT	5	0.015
negative regulation of nucleobase,	GOTERM_BP_FAT	5	0.014

(continued on next page)

(continued from previous page)

nucleoside, nucleotide and nucleic acid metabolic process			
negative regulation of nitrogen compound metabolic process	GOTERM_BP_FAT	5	0.014
negative regulation of macromolecule biosynthetic process	GOTERM_BP_FAT	5	0.015
negative regulation of cellular biosynthetic process	GOTERM_BP_FAT	5	0.016
negative regulation of biosynthetic process	GOTERM_BP_FAT	5	0.016
negative regulation of macromolecule metabolic process	GOTERM_BP_FAT	5	0.033
cell fate commitment	GOTERM_BP_FAT	3	0.073
negative regulation of transcription from RNA polymerase II promoter	GOTERM_BP_FAT	3	0.16
positive regulation of gene expression	GOTERM_BP_FAT	7	0.00014
positive regulation of macromolecule metabolic process	GOTERM_BP_FAT	7	0.0011
regulation of transcription from RNA polymerase II promoter	GOTERM_BP_FAT	5	0.034
positive regulation of gene-specific transcription	GOTERM_BP_FAT	3	0.033
regulation of gene-specific transcription	GOTERM_BP_FAT	3	0.071

(continued on next page)

(continued from previous page)

positive regulation of transcription, DNA-dependent	GOTERM_BP_FAT	4	0.075
positive regulation of RNA metabolic process	GOTERM_BP_FAT	4	0.073
transcription factor binding	GOTERM_MF_FAT	4	0.068
positive regulation of transcription	GOTERM_BP_FAT	4	0.1
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GOTERM_BP_FAT	4	0.13
positive regulation of nitrogen compound metabolic process	GOTERM_BP_FAT	4	0.13
positive regulation of macromolecule biosynthetic process	GOTERM_BP_FAT	4	0.14
positive regulation of cellular biosynthetic process	GOTERM_BP_FAT	4	0.15
positive regulation of biosynthetic process	GOTERM_BP_FAT	4	0.15
positive regulation of transcription from RNA polymerase II promoter	GOTERM_BP_FAT	3	0.26
receptor	SP_PIR_KEYWORDS	4	0.17
zinc finger region:NR C4-type	UP_SEQ_FEATURE	3	0.024
DNA-binding region:Nuclear receptor	UP_SEQ_FEATURE	3	0.024
Zinc finger, nuclear hormone receptor-type	INTERPRO	3	0.018

(continued on next page)

(continued from previous page)

Steroid hormone receptor	INTERPRO	3	0.0097
Nuclear hormone receptor, ligand-binding, core	INTERPRO	3	0.0067
Nuclear hormone receptor, ligand-binding	INTERPRO	3	0.0067
Zinc finger, NHR/GATA-type	INTERPRO	3	0.0057
steroid hormone receptor activity	GOTERM_MF_FAT	3	0.01
ZnF_C4	SMART	3	0.012
HOLI	SMART	3	0.0067
promoter binding	GOTERM_MF_FAT	3	0.012
ligand-dependent nuclear receptor activity	GOTERM_MF_FAT	3	0.01
zinc finger	SP_PIR_KEYWORDS	3	0.02
zinc	SP_PIR_KEYWORDS	6	0.026
zinc-finger	SP_PIR_KEYWORDS	5	0.059
DNA binding	SP_PIR_KEYWORDS	3	0.065
metal-binding	SP_PIR_KEYWORDS	6	0.068
zinc ion binding	GOTERM_MF_FAT	6	0.22
receptor	SP_PIR_KEYWORDS	4	0.17
transition metal ion binding	GOTERM_MF_FAT	6	0.33
metal ion binding	GOTERM_MF_FAT	6	0.79
cation binding	GOTERM_MF_FAT	6	0.78
ion binding	GOTERM_MF_FAT	6	0.77

The number of transcription factors that only target protective variants is relatively low (Table 3.19). While keeping this in mind, an enrichment analysis with DAVID was done (Table 3.20). All twelve "protective" genes are annotated with "regulation of transcription" and with "nucleus" confirming their roles as regulators which is further supported by the other annotation terms in the first and second cluster which are all

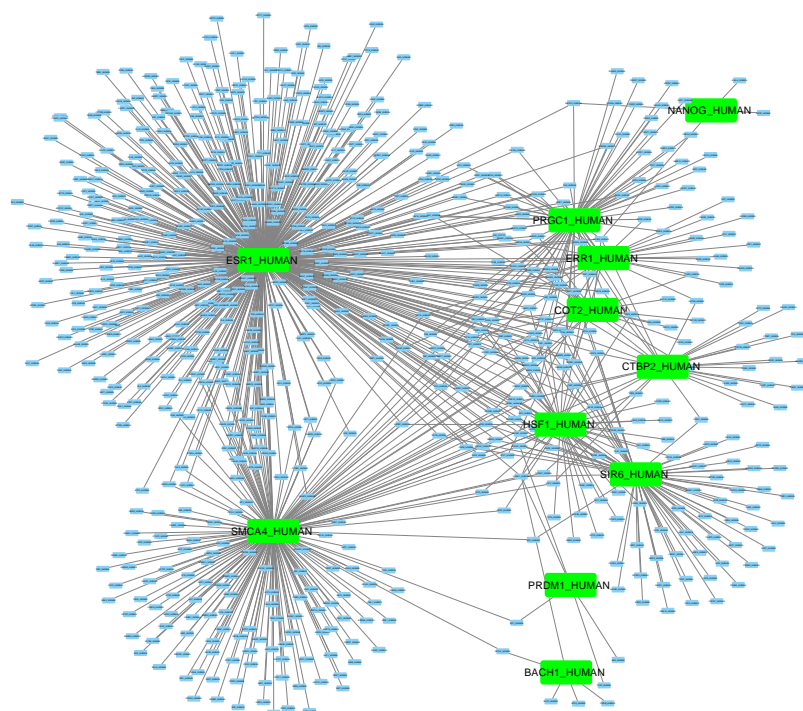


Figure 3.46.: Protein-Protein interactions between "protective genes" and their direct interaction partners. The protective genes/proteins are shown in green and the interaction partners are shown in blue. ESR1 is directly connected to ERR1, PRGC1 and SMCA4. ERR1 is furthermore connected to PRGC1 and HSF1 is connected to SMCA4.

related to DNA binding and transcription. Cluster 3 contains many terms that are related to negative regulation. The last two terms in cluster 3 are above the benjamini-hochberg significance threshold. Cluster 4 is somewhat the counterpart of Cluster 3 because it contains many terms for positive regulation. But most terms are above the significance threshold. Finally there is cluster 5 which consists of terms that are related to zinc-finger transcription factors and steroid hormone receptors.

In the following text the term "protective gene" will refer to transcription factors whose TFBSs contain protective but no risk variants for specific diseases. Conversely, "risk genes" will refer to transcription factors whose TFBSs contain risk variants but no protective variants. The term "mixed gene" refers to transcription factors with both, risk and protection variants within their TFBSs for a specific disease.

Eleven of the twelve protective genes/proteins have high-confidence interactions in the CPDB. They interact with 681 other proteins. Of these 681 proteins 138 interact directly with more than one of the protective genes. One of these protective genes is the estrogen

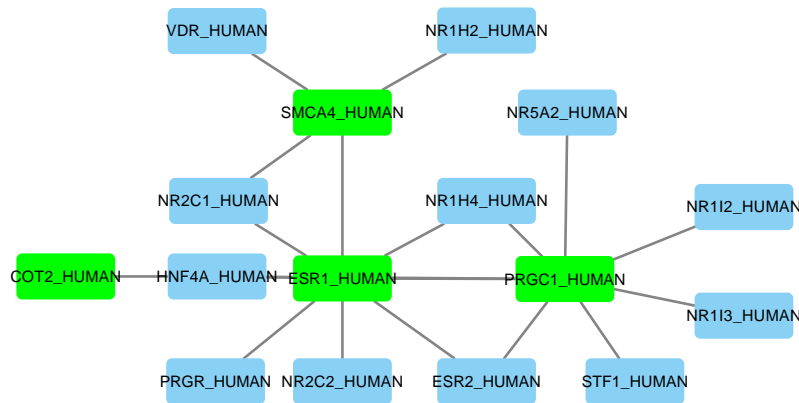


Figure 3.47.: Protective genes (green) and genes that are enriched for the vitamin D receptor (blue). The benjamini-hochberg-corrected DAVID enrichment p-value is 3^{-9} .

receptor (ESR1) and it should be noted that the estrogen receptor is a hub protein with 445 interaction partners. Figure 3.46 depicts the interactions of these protective genes. The non-protective interactors in turn have been tested for enrichment with DAVID. Notable terms include: (regulation of) transcription, nuclear lumen, positive/negative regulation of gene expression, transcription factor binding, chromatin regulator, Vitamin D receptor.

The vitamin D receptor is interesting, because Vitamin D deficiency has been linked to inflammatory bowel disease⁸⁹⁹⁰⁹¹. Figure 3.47 shows the twelve linker genes that enrich for the vitamin D receptor. It is known that dendritic cells respond to vitamin D stimuli and can induce pro- or anti-inflammatory responses²³.

3.3.4. Protective genes in detail

In the following section the twelve genes/proteins, for which only protective variants were detected, will be presented in detail.

BACH1

BACH1 is a protective gene for primary sclerosing cholangitis and a risk gene for psoriasis. In the other three diseases it is a mixed gene.

BACH1 serves as a transcriptional regulator. It can activate and repress transcription⁹².

BACH1 is similar to BACH2 which is already known to be related to IBD in the context of oxidative stress²³. BACH2 is also present in the heatmaps in the figures 3.31, 3.33 and 3.35.

COUP transcription factor 2 (COT2_HUMAN)

COT2 is a protective gene in psoriasis and a risk gene in PSC. In other diseases it is a mixed gene. The full name is COUP transcription factor 2.

CTBP2

CTBP2 is a protective gene in PSC. In all other diseases there are risk but also protective variants in the TFBSs. According to UniProt it is a "Corepressor targeting diverse transcription regulators". It is involved in the following human KEGG pathways:

- Wnt signaling pathway (hsa04310)
- Notch signaling pathway (hsa04330)
- Pathways in cancer (hsa05200)
- Chronic myeloid leukemia (hsa05220)

CTCF

This gene is protective in ankylosing spondylitis and ulcerative colitis. For psoriasis only risk variants are known and primary sclerosing cholangitis is a mixed case. It has no high-confidence interactions in the ConsensusPathDB. It is responsible for epigenetic reprogramming⁹³.

ESRRA (ERR1_HUMAN)

Protective in Crohn's disease and psoriasis. The full name is Steroid hormone receptor ERR1.

ESR1

ESR1 is protective in psoriasis while all other diseases have mixed types of risk factors. ESR1 is the estrogen receptor and is involved in the following KEGG pathways:

- Estrogen signaling pathway (hsa04915)
- Prolactin signaling pathway (hsa04917)
- Thyroid hormone signaling pathway (hsa04919)
- Endocrine and other factor-regulated calcium reabsorption (hsa04961)
- Proteoglycans in cancer (hsa05205)

HSF1

HSF1 is protective in the four diseases AS, PS, PSC and CD. It is mixed for ulcerative colitis. It is a transcriptional activator⁹⁴.

NANOG

NANOG is protective in ulcerative colitis. According to UniProt it is a "Transcription regulator involved in inner cell mass and embryonic stem (ES) cells proliferation and self-renewal. "

PRDM1

PRDM1 is protective in UC, PSC, CD and PS. According to UniProt it is a "Transcriptional repressor that binds specifically to the PRDI element in the promoter of the beta-interferon gene⁹⁵. [It] Drives the maturation of B-lymphocytes into Ig secreting cells⁹⁶."

PPARGC1A (PRGC1_HUMAN)

PPARGC1A is protective in Crohn's disease. On UniProt.org it has the following description:

Transcriptional coactivator for steroid receptors and nuclear receptors. [...] Plays an essential role in metabolic reprogramming in response to dietary availability through coordination of the expression of a wide array of genes involved in glucose and fatty acid metabolism. Induces the expression of PERM1 in the skeletal muscle in an ESRRA-dependent manner. [...]

There appear to be interesting connections to diet and estrogen-receptor-related issues. The protein is involved in the following KEGG pathways:

- AMPK signaling pathway (hsa04152)
- Longevity regulating pathway (hsa04211)
- Insulin signaling pathway (hsa04910)
- Adipocytokine signaling pathway (hsa04920)
- Glucagon signaling pathway (hsa04922)
- Insulin resistance (hsa04931)
- Huntington's disease (hsa05016)

SIRT6 (SIR6_HUMAN)

SIRT6 is protective in Crohn's disease while it has only known risk variants in AS and PS. Is involved in the KEGG pathway "Central carbon metabolism in cancer (hsa05230)". SIRT6 is involved in aging processes and NF κ B regulation through histone deacetylation⁹⁷.

SMARCA4 (SMCA4_HUMAN)

SMARCA4 has protective variants in all diseases. In Crohn's disease there are also risk variants. The full name is Transcription activator BRG1.

3.3.5. Comparison with geneset-based networks

A comparison with the genes obtained from using DEPICT in section 3.2 and all DNA binding elements detected in this section (irregardless of risk status) yielded only an overlap of nine genes: ETS1, FOSL2, FOS, IRF1, NFAC1, NFKB1, RUNX3, STA5A,

STAT3. For comparison: The largest connected component from the DEPICT-analysis (Figure 3.37) consists of 37 nodes and nine of these nodes overlap with the DNA binding elements from this section. There is no overlap with the mostly unconnected nodes (Figure 3.38) that were detected based on both, disease-specific p-values (76 mostly unconnected nodes) and SBM p-values (80 mostly unconnected nodes).

3.3.6. Network analysis

In order to study the importance of the genes with different risk status (protective, risk, mixed, neutral) in the network, the centrality of every node was determined. For this the whole CPDB network (version 31, interaction confidence ≥ 0.95) was taken and different centrality measures have been applied to it. The same centrality measures have been taken for the subnetwork of nodes that are associated with at least one of the five diseases (called Cross-disease subgraph). All centrality measures have been calculated with the networkx python package⁷⁹. The centralities are: Degree centrality, betweenness centrality, closeness centrality and eigenvector centrality. Figures 3.48 to 3.52 show a selection of the distribution of the centralities for the five diseases. The full set of distribution plots can be found in the supplementary figures (digital attachment).

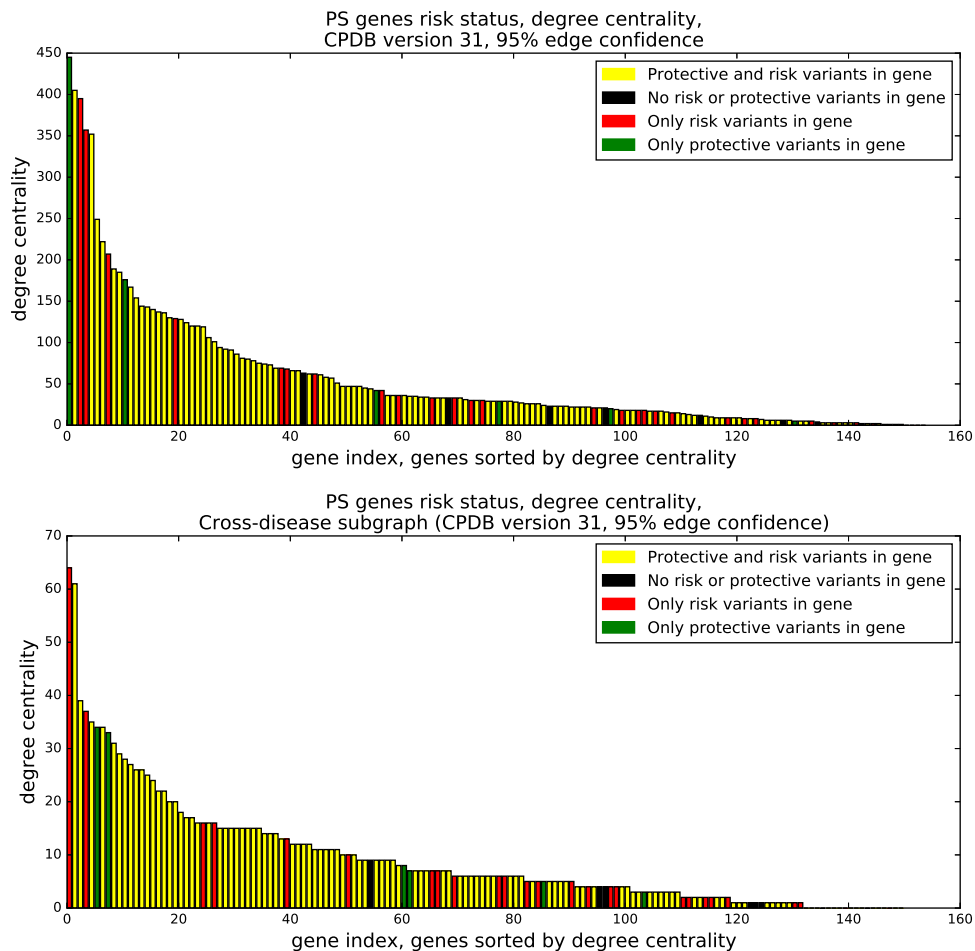


Figure 3.48.: **Top:** Distribution of node *degree* centrality in the complete ConsensusPathDB (minimum edge confidence 95 %) for nodes that are associated with at least one disease and **Bottom:** Distribution of node *degree* centrality in the subgraph of the ConsensusPathDB that consists only of nodes that are associated with at least one disease.

Both: The colors of the bars indicate the risk status of all minor variants in the study population² that a DNA-binding element interacts with.

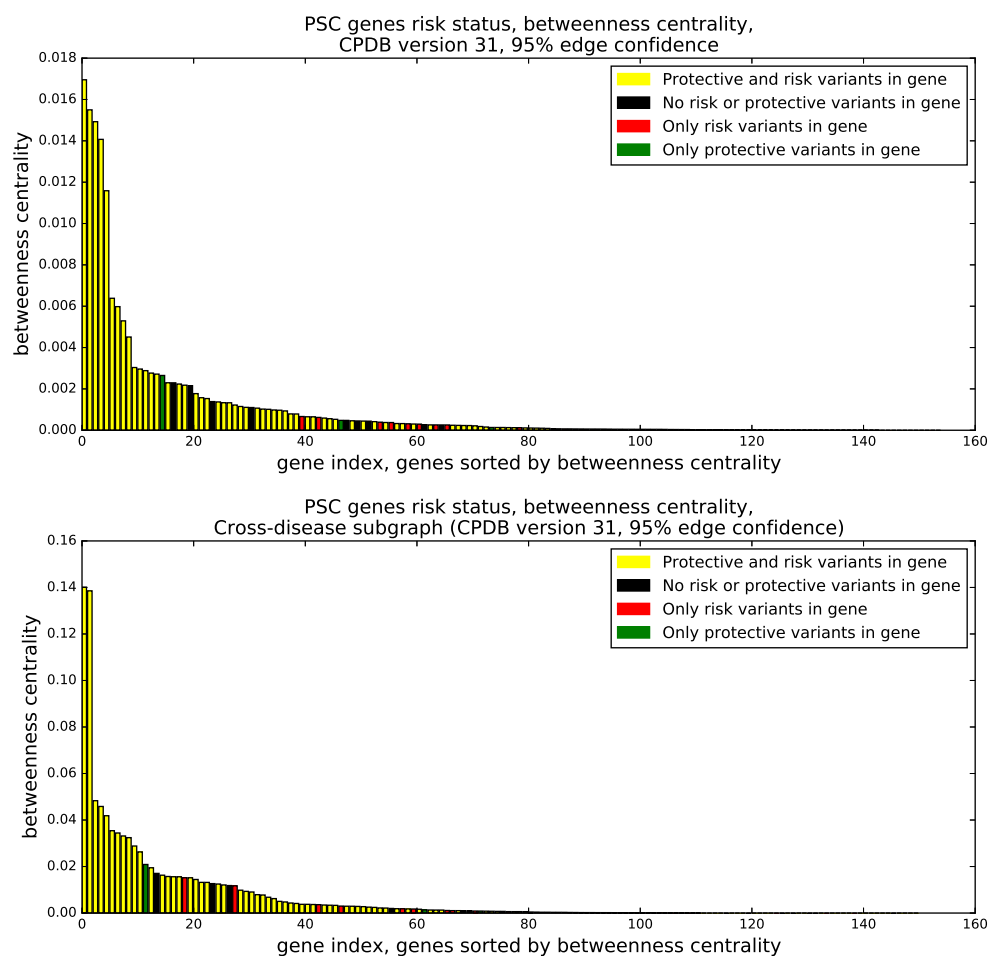


Figure 3.49.: **Top:** Distribution of node *betweenness* centrality in the complete Consensus-PathDB (minimum edge confidence 95%) for nodes that are associated with at least one disease and **Bottom:** Distribution of node *betweenness* centrality in the subgraph of the ConsensusPathDB that consists only of nodes that are associated with at least one disease.

Both: The colors of the bars indicate the risk status of all minor variants in the study population² that a DNA-binding element interacts with.

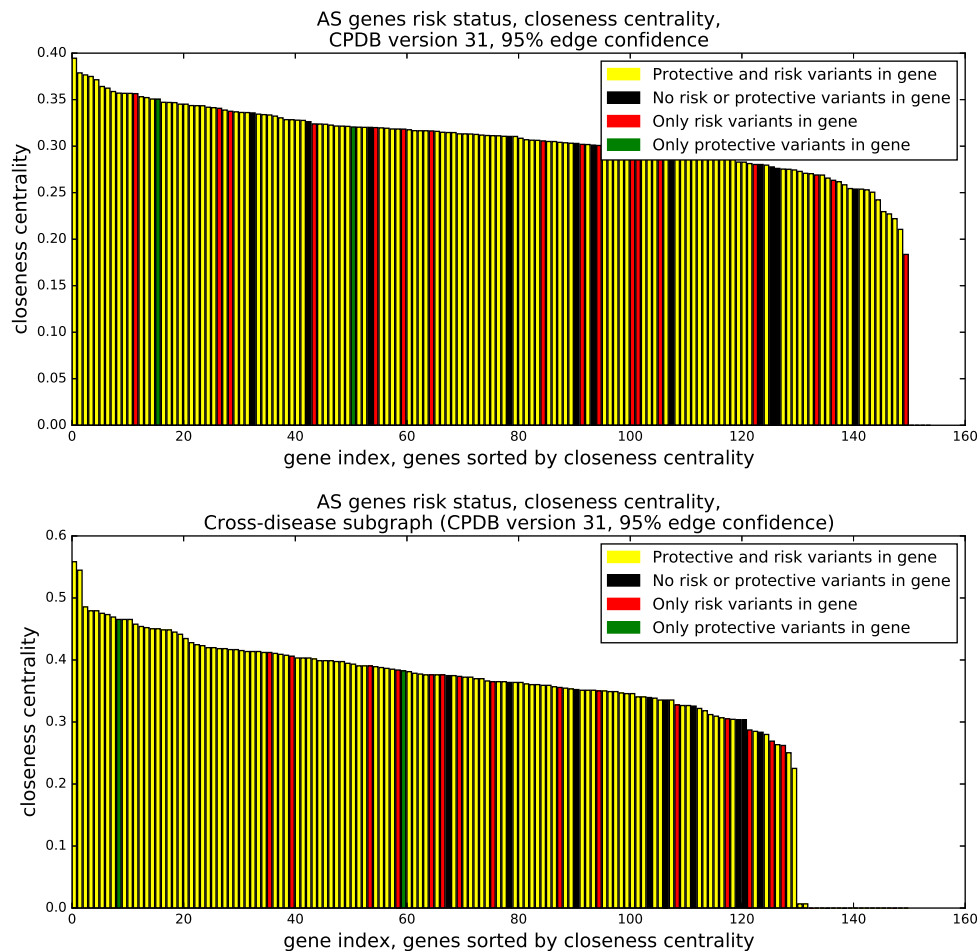


Figure 3.50.: **Top:** Distribution of node *closeness* centrality in the complete Consensus-PathDB (minimum edge confidence 95 %) for nodes that are associated with at least one disease and **Bottom:** Distribution of node *closeness* centrality in the subgraph of the ConsensusPathDB that consists only of nodes that are associated with at least one disease.

Both: The colors of the bars indicate the risk status of all minor variants in the study population² that a DNA-binding element interacts with.

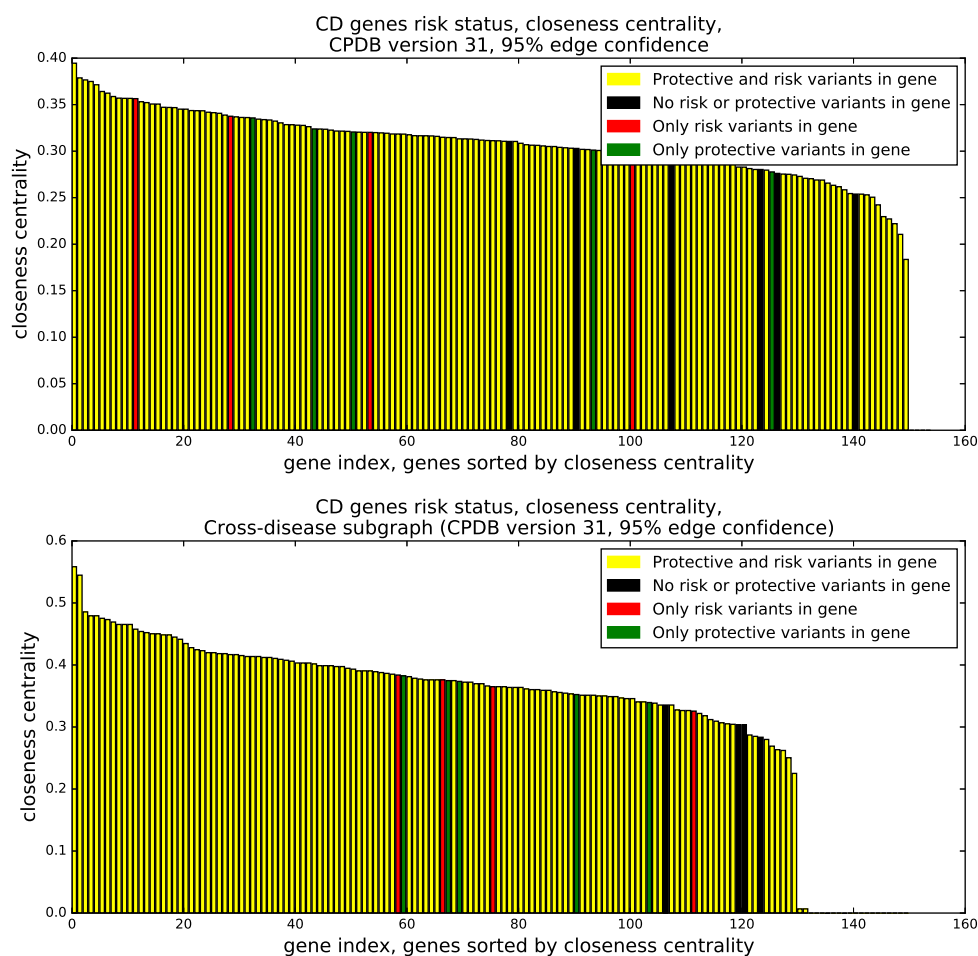


Figure 3.51.: **Top:** Distribution of node *closeness* centrality in the complete Consensus-PathDB (minimum edge confidence 95%) for nodes that are associated with at least one disease and **Bottom:** Distribution of node *closeness* centrality in the subgraph of the ConsensusPathDB that consists only of nodes that are associated with at least one disease.

Both: The colors of the bars indicate the risk status of all minor variants in the study population² that a DNA-binding element interacts with.

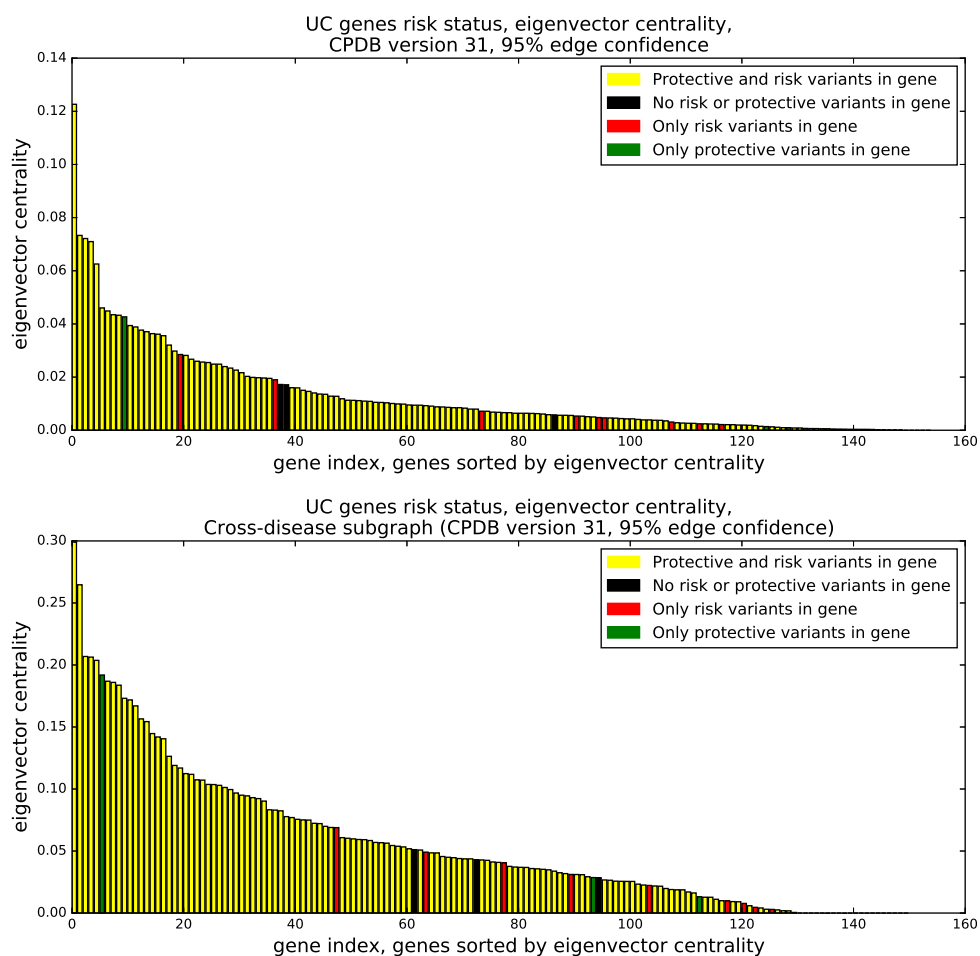


Figure 3.52.: **Top:** Distribution of node *eigenvector* centrality in the complete Consensus-PathDB (minimum edge confidence 95%) for nodes that are associated with at least one disease and **Bottom:** Distribution of node *eigenvector* centrality in the subgraph of the ConsensusPathDB that consists only of nodes that are associated with at least one disease.

Both: The colors of the bars indicate the risk status of all minor variants in the study population² that a DNA-binding element interacts with.

	AS	CD	PS	PSC	UC
CTCFL_HUMAN	protective	mixed	risk	neutral	protective
SIX5_HUMAN	mixed	mixed	risk	mixed	mixed
SP2_HUMAN	risk	mixed	risk	risk	mixed
SP4_HUMAN	risk	mixed	risk	risk	mixed

Table 3.21.: Risk status of proteins which do not have any high-confidence interactions in the ConsensusPathDB. The risk status is defined by the variants within TFBSs that are either risk-inducing, risk-reducing (protective), a combination of risk-inducing and risk-reducing (mixed) SNPs. If neither risk-inducing nor risk-reducing variants are known, the risk status is neutral.

The four disease-associated proteins CTCFL_HUMAN, SIX5_HUMAN, SP4_HUMAN, and SP2_HUMAN did not have any high-confidence interactions in the ConsensusPathDB (version 31). Table 3.21 shows the disease status of these four missing proteins.

In addition to the centralities, the assortativity of nodes of different risk status has been determined. In all cases the assortativity is close to zero and therefore does not indicate grouping or anti-grouping of nodes of the same risk classes.

Chapter 4.

Discussion

*“This was a triumph! I’m making a note here: Huge success!
It’s hard to overstate my satisfaction.”*

— Mad robot GLaDOS (Lyrics)

Networks are about connections. The interactors themselves are certainly relevant, but a far greater and more important challenge lies in determining which interactions play a role in a phenotype and what their effects are. Reference databases list plenty of interactions but the context in which these interactions have been observed is often not clearly documented or researched⁴⁸.

In the course of this thesis many networks have been created. But only a selection is presented. Especially networks of great complexity have been left out because they are too difficult to visualize and too difficult analyze directly. But even the networks presented in this work have many nodes and edges and therefore only a selection of nodes and edges will be discussed in detail.

4.1. Biology of Inflammatory Diseases

When considering the genes, networks and term enrichment results, it appears that regulation and signaling are the central aspects that matter in the diseases. Cellular communication is responsible for the coordination of many intra- and intercellular processes⁹⁸. It is plausible that this regulatory machinery is not reacting correctly to

external or internal stimuli. In the case of IBD there is accumulating evidence that the disease is a result of an inappropriate inflammatory response to intestinal microbes²².

There exist several mechanisms to downregulate immune responses, including downregulation of inflammation by IL-10²¹. T-cells have a PD-1 receptor that leads to their inactivation upon binding. Cancer cells exploit this receptor to neutralize T-cells⁹⁹. Some medications that interfere with this exploitation have side effects that lead to overreactions of the immune system and some cases were reported where patients developed psoriasis¹⁰⁰. This underlines the potential importance of signaling and regulation in inflammatory or immune-related diseases.

The enrichment results also clearly indicate that the immune system is a major factor in these diseases even when a background correction for the Immunochip is used.

Regulatory T- and B-cells²² but also macrophages and monocytes²¹ appear to play a major role in inflammatory diseases. These immune cells secrete and react to different cytokines. Macrophages are especially susceptible to regulatory signaling molecules and exhibit very diverse behaviours depending on the tissue and the cytokines they are exposed to²¹. This fact alone indicates that we need a better resolution on the tissue level when we want to understand inflammatory diseases better. This thesis does not try to take into account the different tissues and the different cells in the human body. Although DEPICT makes use of gene expression data and performs tissue enrichment. It detected many immune cells for all five diseases but it failed to detect the known affected organs of ankylosing spondylitis (spine), psoriasis (skin) and primary sclerosing cholangitis (liver/bile ducts).

Inflammatory diseases are not the only diseases that are linked to (dysfunctional) regulatory processes. Cancer is a category of diseases with aberrant signal processing⁹⁸. Some genes play a role in both types of diseases. For instance, JAK2 is a tyrosine kinase that regulates cellular growth processes but it is also involved in signaling for innate and adaptive immunity and many more signaling processes¹⁰¹. A similar common protein is NF κ B1 which participates in many biological processes¹⁸. Both proteins occur in several networks shown in the results section. It is already known that inflammation is a common observation in cancerous tissue⁹⁸. The same pathways might be active in inflammatory diseases and cancer. However, the enrichment listings in this thesis do not show any cancer-related results. But some of the protective transcription factors are known to be involved in cancer-related pathways: CTBP2, ESR1 and SIRT6. It is also known that patients with ulcerative colitis have an increased risk of colorectal cancer¹⁰².

Some studies have shown a connection between the nervous system and regulation of inflammation^{103;104}. Martin-Subero et al. report a high comorbidity of depression and inflammatory bowel disease¹⁰⁵. They state that depression is linked to immune-inflammatory, oxidative and nitrosative stress pathways which also includes gut-brain pathways. They also state that these pathways are relevant for IBD. However, the analyses presented in this thesis did not detect any obvious neurological pathways, mechanisms or genes.

Khor et al.²³ describe six IBD genes that are related to the epithelial barrier. But only one of these genes was detected in the context of this thesis: *ERRFI1*. This gene is present in the (mostly) unconnected DEPICT subnetwork when using SBM p-values. The enrichment results did not indicate that epithelial barriers plays an important role in IBD even though it obviously is important. An explanation could be that the clinical symptoms of the epithelial barrier are mainly determined by environmental factors and less by genetics.

Apoptosis is another common enrichment term. It is known that mucosal T-cells can have increased resistance against apoptosis¹⁰⁶. On a related note, Infliximab is a drug given as a treatment for several inflammatory diseases. This drug induces apoptosis in T-cells²⁵. However, the role of apoptosis in these diseases still seems to be unclear.

Further discoveries may change our perspective on these diseases. A common view is that these diseases are seronegative, that is, there are no autoantibodies involved in these diseases¹⁰. Very recently Quaden et al. describe the discovery of autoantibodies in AS. More interesting findings may follow.

4.1.1. Networks of Inflammatory Diseases

Most disease-associated SNPs only increase or decrease the risk for developing a disease by a low amount. The combined effects of several SNPs and environmental factors is probably required to cause the diseases¹. This thesis attempts to find connections between SNPs to see which SNPs affect genes and how these genes (or proteins) interact with each other.

Most of the used SNPs lie outside of coding regions. They might act as expression quantitative trait loci (eQTL) that affect the level of gene expression of genes which is a common phenomenon for disease genes¹⁰⁷. Different levels of expression may in turn change the strength of a regulatory signal and cause overreactions or insufficient

reactions²³. For instance, defective TGF β signaling is known to impair the transformation of pro-inflammatory monocytes to inflammation-anergic macrophages²¹.

A considerable overlap can be observed between the genes that are associated with the investigated diseases (Figures 3.18 and 3.37). This is only partly surprising because there is already a large overlap between the SNPs that are associated with the five diseases. A superficial comparison with multiple sclerosis shows that JAK2, STAT5A, ETS1, SOCS1, IL12B, CD40 and other genes are shared with some of the five diseases¹⁶. This indicates that there are more commonalities between inflammatory diseases. Goh et al. present the diseasome network which shows that many genes are relevant to several diseases⁴⁴.

The disease-specific subnetworks with less than five diseases (Figures 3.19 to 3.22) are mostly unconnected. Only the network that contains nodes that are associated with exactly one disease has one larger connected component that is specific to genes associated with Crohn's disease. One notable node of this component is EP300 which is responsible for histone acetyl transfer¹⁰⁸ and which is relevant for autophagy¹⁰⁸. Therefore EP300 fits well to Crohn's disease²⁹. However, the remaining nodes in this component are rather dissimilar and provide no obvious explanation why they might be relevant for CD.

The HLA/MHC region on chromosome 6 is highly associated with inflammatory diseases^{109;24}. However, due to high linkage disequilibrium and high variance in the population it is even more difficult to map significant variants to genes and to impute genotypes¹¹⁰. It was observed that the MHC region had a major influence on the genes that VEGAS choose to be significant (Table 3.2). The MHC genes are a crucial part for the interaction between host and microbes and a better understanding on how the MHC region affects genes is desirable to construct networks that represent the mechanisms of the diseases better¹.

Networks consist of nodes and edges. A causative SNP can either affect nodes or edges. That is, the SNP can either affect the *function* of a single gene/protein/RNA molecule or it can affect the *interaction* between two molecules. It is already difficult to determine the causative variant, but once it has been identified, it should be determined if the effect directly influences a node or if it influences an edge in the network. Making this distinction might offer a better understanding of the wider effect(s) that the SNP induces. But then it still has to be understood if the causative variant impairs or improves the biological process that it influences.

For instance, it is known that SOCS1 inhibits JAK2¹⁸. This interaction has been most notably observed in Figure 3.37 and it probably plays a role in CD and UC. In

the same figure there is also the Toll-like receptor 4 (TLR4) on the far right. It is involved in the innate immune response to bacteria and is therefore a good example for an immunological protein/gene that works at the interface between host and microbes. According to DEPICT it is only associated with CD but some publications indicate that it also plays a role in psoriasis¹¹¹ and ulcerative colitis¹¹² while others found no association for AS¹¹³ and for PSC¹¹⁴. Therefore DEPICT might have missed more disease associations. TLR4 is known to be involved in many other diseases¹⁰⁴.

Another well-known example of an immunological interface protein is NOD2 which was detected by the DEPICT-based workflow but which has no neighbours in the network (Figure 3.38, top row). NOD2 was correctly identified as a CD-associated gene¹¹⁵.

In the analyses of the effect directions of SNPs some exclusively "protective" genes were found for some diseases. However, these genes are DNA-binding elements, that bind close to or at the same position as a SNP. They are not the genes regulated by the TFBSs. These binding proteins might be a lot less disease-specific than the actual genes regulated by the TFBSs. But if a variant does indeed affect the binding of a factor, then this could be relevant for the development of diseases.

4.1.2. Representation of Biology in Networks

A common observation of protein-protein interaction networks is that there is one large connected component and several smaller components of relatively low size⁴⁶. This observation was also true for most of the networks that were observed in this study.

It is plausible that several different cell types are involved in the etiology of inflammatory diseases. We currently do not have a good resolution on the cell-type level to see which cell types are affected by which SNPs. The real biological networks might even span several different cell types which seems plausible given that signaling is a common theme in the networks. In addition, there are further dimensions to consider when looking at protein-protein interactions: Proteins can be in different states due to chemical modifications like phosphorylation and they can have different amino acid sequences due to alternate splicing. These factors can influence the binding and the activity of proteins⁶.

Various filtering methods (including subnetwork generation methods) were applied to the networks which resulted in singleton nodes without any connections. Under the assumption that these singletons are truly associated with a specific disease we either

lack interaction knowledge to connect these nodes to the rest of the network or the disease-causing role of such a node does not involve protein-protein interactions.

A central problem of network analyses is the biological interpretation of the connections. Especially in greater networks there are a lot of possible pathway flows that have to be considered. Semantic and causative annotations for interactions between nodes like those from the openBEL project¹¹⁶ could help to understand the networks as a system of dependent interactors.

For protein-protein interaction networks the direction of the edges are not known. An interaction denotes an observed binding of two proteins but many times it is unknown what the effect of the binding actually is and which protein is being regulated and which protein is a regulator. The Hippie PPI database tries to create directed PPI networks by taking a list of source nodes (starting nodes) and a list of sink nodes (target nodes) and determines the shortest paths between source and sinks nodes and assigns directionality along the shortest paths¹¹⁷. However, it is still the responsibility of the user to provide a correct list of source and target nodes for the network which is far from trivial.

A common goal of network analysis is finding new disease genes. This is usually achieved by selecting known neighbours of disease-associated nodes. This was done in the context of various analyses and in most cases the size of the networks drastically increased and it was difficult to get a clear picture of the situation. Enrichment analyses helped to characterise the neighbouring genes but overall too many genes were added to the network. Thus the principle guilt-by-association is problematic when dealing with too many neighbours. To compensate for the many linker genes, a prioritisation of the linker genes was attempted in section 3.2.6 by scoring linker genes higher that had a high ratio of disease neighbours to non-disease neighbours. Among these genes are IL23A, which is a subunit of the IL23 cytokine¹¹⁸ which in turn interacts with the highly disease-associated IL23 receptor¹¹⁸ (IL23R) (disease association with AS, CD, PS and UC). This makes it a clear candidate for being relevant for these diseases. IL12RRB1 (I12R1_HUMAN) is another interactor of the just mentioned IL23R¹¹⁸. ITGAL is a linker gene for PS and it is another receptor component that is involved in cell adhesion and immune system processes, including inflammation¹¹⁹. It is expressed in all leukocyte lineages¹¹⁹ and it is already associated with AS and UC. IRF6 is has only two connections, both which are UC-associated. It is involved in cell cycle regulation¹²⁰. FGFR3 is a tyrosine-protein kinase¹²¹ and it is linked to two UC genes. It is activated by the MAP kinase cascade and STATs¹²¹. FGFR3 activates the SOCS1 and SOCS3 (SOCS=suppressors of cytokine signaling)¹²¹. SOCS1 is linked to JAK2 in the DEPICT-based network (Figure 3.37)

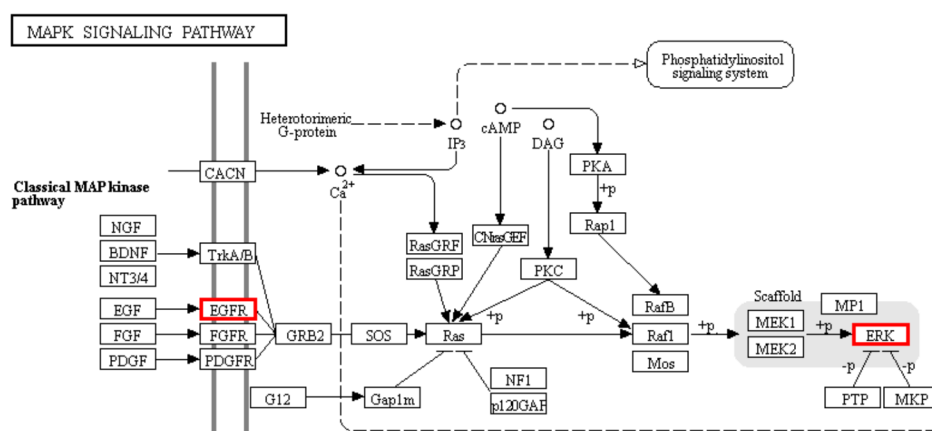


Figure 4.1.: The EGF/MAPK signaling pathway. This pathway consists of many individual steps that should be executed in a controlled manner because otherwise aberrant cellular growth might be the consequence. This image is based on the KEGG^{59;58} pathway diagram map04010.

and is known to inhibit JAK2¹⁸. It is also involved in the regulation of vitamin D metabolism¹²². Vitamin D deficiency is a common observation in IBD patients and Vitamin D is important for downregulation of inflammation and modulation of autophagy, gut barrier maintenance and immune system functions¹²³. All of this make FGFR3 a good candidate for further investigations. The gene PRDC1 is linked to two UC genes. It seems that not much about its function is known.

Because the human genome is diploid, all non-homozygous genes should actually be represented by two nodes in the network to account for alternative alleles. Or, in the case of non-homozygous regulatory regions, two edges should represent different regulatory effects. However, given that this thesis only worked with summary data (all known disease-associated SNPs across study populations), this modelling approach only makes limited sense. But it is conceivable that interactions between the same, albeit heterozygous genes could contribute to disease while homozygous alleles could be without a notable effect, especially when it comes to dimeric proteins like JAK2.

Finally there is the problem that the real pathways in a network might not be the shortest pathways. A typical pathway is the EGF pathway (Figure 4.1) which consists of many steps. But there are also alternative pathways that connect different parts of the EGF pathway and are actually shorter. A shortest path search between EGFR and ERK (MK03_HUMAN) shows that there are 43 proteins directly connected to both EGFR and ERK. Each of these intermediate proteins provides a shortest path between the two proteins that is only two steps long (based on CPDB version 31, 95% edge

confidence). One example for such a protein is the estrogen receptor (ESR1_HUMAN) which we have encountered previously in this thesis (section 3.3.3). While this agrees nicely with the small world phenomenon, it does not always do justice to the biological reality. In real biological systems certain conditions have to be met to make it possible for an interaction to take place. In the simplest case both interactors have to be in the same cellular compartment to interact. A challenge of network science is to understand these conditions and eliminate edges that are not relevant under the given conditions⁸².

4.2. Mapping of SNPs to genes

For most SNPs it is still unknown which genes they affect. In addition, linkage disequilibrium makes it difficult to determine the actual causal variant for diseases because unrelated variants are inherited together with the causal variant.

Various approaches exist to map SNPs to genes. VEGAS is a popular choice⁴⁸. However, the reliability of the tool is uncertain due to the fact that VEGAS turns off warnings in its code which might otherwise be helpful to discover problems. Furthermore VEGAS uses the `corpcor` R package to create positive definite matrices, but the package does not always succeed in creating proper positive definite matrices and VEGAS changes the resulting matrices on its own by overwriting entries along the diagonal with input-unspecific constants. Another strange observation is that VEGAS produces gene-wise p-values of magnitude zero during its calculations. The authors state that this is due to "computational reasons" which is not further elaborated. The smallest non-zero 64-bit floating point number that R can work with is 10^{-323} which is very precise. The smallest non-zero p-value that VEGAS produced in all analyses was 10^{-6} which is still very far away from the potential 10^{-323} that can be represented on a modern computer. This strange behaviour could therefore be the result of a software bug.

The VEGAS paper states that a p-value of 10^{-6} is sufficiently significant because it is below the Bonferroni-corrected threshold of $p < 2.8 \times 10^{-62}$ and at this level no additional attempts are made to reduce the gene-wise p-value even further. Still, this leads to a lot of genes having a p-value of zero and a lot of p-values having a p-value of 10^{-6} . The IMSCG relied on these p-values to prioritize genes to generate submodules with the Cytoscape plugin `jActiveModules`.

There are several successors to VEGAS. Among them are VEGAS 2, and PASCAL¹²⁴. VEGAS 2 was still in development when it was tested with our data. VEGAS 2 showed the following behaviour: All genes in the result set had the same p-value. When using VEGAS 1 with the same list of SNPs different gene p-values were reported. PASCAL is a newer software which has not been evaluated yet. According to the authors, PASCAL outperforms VEGAS and is also able to handle p-values smaller than 10^{-6} .

A simpler approach to gene-mapping is to determine the linkage region of a lead SNP and incorporate all genes that lie within that region for further analysis. This approach is simple and fast. Pers et al. determined that a linkage distance of $r^2 > \frac{1}{2}$ is a good choice for the size of a linkage region⁵¹. Further prioritisation of genes within each region can be performed. For instance, DEPICT tries to find combinations of genes from all regions that are as closely related as possible. The idea behind this seems plausible. All of these SNPs have something to do with a specific disease. Therefore all *genes* affected by these SNPs should also have something to do with a specific disease and share some commonalities.

However, the effects of SNPs can be remote. A good example are distal transcription factor binding sites which might change their binding affinity for transcription factors. This binding affinity can increase or decrease the rate of translation of a gene. That is why using annotations like those from the ENCODE project appear to be a good idea to determine which transcription factors (and, if it is known, which genes) could be affected by distant variants that lie far outside of proximal linkage regions. A considerable overlap between distal TFBS and SNPs was observed in this project indicating that distal TFBS might play a role in these diseases. However, GoShifter also tries to assess whether the overlap of all LD SNPs with an annotation is by chance. Unfortunately the GoShifter paper does not give a good guideline what threshold should be chosen for a good score and therefore no filtering was done in the work presented here. This might have led to more false positives. Another problem is that GoShifter only gives a score to the lead SNP but not the LD SNPs so that all LD SNPs have to be taken into account for further analysis.

An observation that was made throughout all analyses is that the number of SNPs that are known for a disease directly affects the number of genes that will be marked as potentially associated. Crohn's disease and ulcerative colitis are considerably more researched than the other three diseases. This in turn might lead to imbalances when it comes to reasoning about how these diseases manifest themselves on the network level. There might be more SNPs that could be relevant for the other diseases. Current

GWAS studies use imputation instead of complete sequencing and are unable to discover completely new variants²³.

On the flip side, less SNPs could make the analysis simpler. It is known for the five diseases that there are subtypes of each disease². Each subtype might have a different genetic profile and distributing the SNPs among the profiles could make understanding the etiology of the diseases easier. Given the multitude of genetic factors and yet to be discovered environmental factors, there is probably not a single etiology but several.

But even knowing a causal variant might not directly lead to an understanding of how that causal variant influences the development of disease because in the case of transcription factor binding sites there can be several transcription factors that bind to that site of 150 base pairs according to ENCODE. Almost all TFBS encountered in this work were of size 150bp. Four TFBS were 290bp or 270bp long.

Using ENCODE annotations that list transcription factors that bind to specific regions appears to be the most reliable way to link SNPs to genes/proteins, simply because the annotations are manually curated. The problem with this approach is that this only captures SNPs that potentially affect the binding of transcription factors. Therefore a combination of different mapping methods is advised to get the most realistic results.

Overall it was observed that many LD SNPs are indeed located at TFBS and might act as eQTLs that affect the efficiency of starting transcription. This would fit well with the small effect sizes because these SNPs could disturb the initialisation of transcription and make certain pathways less efficient without rendering them totally nonfunctional. But a combination of several genetic factors could create enough disturbances that it is more likely that functionality is critically impaired and disease develops.

It should also be noted that the calculation of odds ratios and thereby the risk classification of SNPs might not be reliable because evolutionary young causal variants have a different allele distribution in the population than evolutionary old causal variants¹⁴. Furthermore, in complex diseases, a combination of several specific SNPs might be required for the disease to develop so that many healthy individuals can carry individual risk genes but they do not get the disease because the required combination of genetic factors is missing¹⁴. These issues can lead to actual risk variants being classified as protective and vice-versa¹⁴. In addition, unknown environmental factors might further skew the association results.

The approach to map SNPs to DNA binding elements (transcription factors) revealed that there is only a small overlap of nine genes between the genes prioritised between DEPICT and the binding elements from ENCODE. This might simply be due to the fact that both approaches are inherently different. The DEPICT approach works with genomic proximity of SNPs to genes while the ENCODE annotations provide a list of proteins that bind to the region of the SNP. These proteins in turn can be located at any position within the genome, including different chromosomes. But still, it is a plausible explanation that SNPs influence the efficiency of transcription. Such an influence could be modelled as an edge in the network between a transcription factor and the gene that is regulated by it. Transcriptional regulation seems to be a neglected phenomenon in the field of inflammation-related networks.

In general there are many different ways a SNP could potentially affect biological function³⁷. There is probably not a single approach that does justice to all of these different ways. Future work should try to take into account known locations of non-mRNA coding regions because these might also play a role.

4.3. Construction of Networks

All analyses in this work used the ConsensusPathDB because it is a metadatabase that consolidates protein-protein interactions from several databases. There are alternative PPI metadatabases like the iRefIndex database which could also have been used but the ConsensusPathDB was chosen because of prior good experience with it and because it was recommended by external scientists.

Different versions of the CPDB have been used throughout this work because newer versions of the database were released over time. Initially an edge confidence threshold of 90 percent was used to have less false negative interactions but given the sheer size of interactions the confidence threshold was increased to 95 percent which still yielded a great number of interactions.

However, biology is much more complex than just protein-protein interactions. The vast majority of GWAS-network analyses use protein-protein interactions^{125;126;127} because they are freely available and quite extensive⁴⁸. But to fully capture molecular networks it is necessary to incorporate further types of interactions like transcription factor

binding interactions, microRNA-mRNA interactions and other types of regulatory RNA interactions.

It should also be noted that proteins can be in different states due to splicing⁶, post-translation modifications (glycosylation)⁶ and chemical modifications like phosphorylation⁶ which can influence how and if a protein interacts with other proteins¹²⁸. Therefore a better resolution about the states of proteins during binding is desirable.

4.3.1. Finding subnetworks and submodules

Only a fraction of the interactome is relevant for the diseases under investigation. Therefore the networks have been filtered down to contain only the genes/proteins that were determined by mapping the SNPs to genes. Given that all five diseases are genetically complex diseases, the subnetworks of these diseases were still too complex to directly analyse with the exception of the networks derived from the DEPICT genelists.

A popular approach is to filter subnetworks down to modules. This approach is implemented by the Cytoscape plugin jActiveModules. However, the algorithm behind jActiveModules was originally developed to detect modules of genes with differential gene expression and not modules of GWAS genes. These are two very different settings. Expression analyses experiments usually have a very limited set of stimuli that change the expression of some genes. In the case of GWAS lists a lot of genes from different contexts are part of the network. However, Ideker et al. also tested jActiveModules with 20 different simultaneous known perturbations in yeast and came up with a submodule of size 340 which they considered to be "extremely large" and to be difficult to analyse visually. They broke down this submodule even further by recursively applying jActiveModules on it. Figure 4.2 shows the submodules that Ideker et al. obtained from the large yeast interaction network. It can be observed that the submodules are almost tree-like, i.e. there are only few circles within the network. The submodules presented in this thesis have more connections in the larger connected components. The submodules obtained by the IMSGC also appear to have a different structure with more cycles than Ideker et al. (Figure 4.3). It is uncertain if these network structures are due to the fact that jActiveModules was originally developed for differential gene expression analyses. jActiveModules assumes that the nodes influence each other in a chain of interactions. This assumption does not hold true for GWAS genelists. Mapping SNPs to genes is difficult and it is likely that the genelists are incomplete and contain false positives. Furthermore, differential gene expression experiments try to measure causal relationships

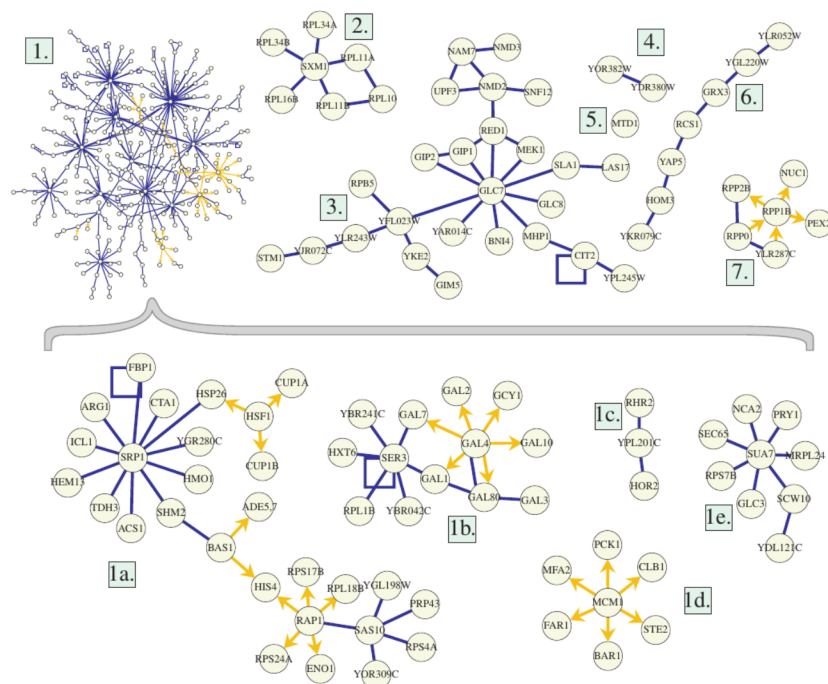


Figure 4.2.: Figure 5 from paper 'Discovering regulatory and signaling circuits in molecular interaction networks' by Ideker et al.⁵⁴. The depicted networks have been created with the jActiveModules algorithm. In contrast to the modules that were detected in this thesis and also in the paper by the IMSGC (see Figure 4.3), the modules are more star-shaped and less densely connected. Subnetwork 1 consists of 340 nodes and was broken down by applying the algorithm again on it which yielded the subnetworks in the lower row.

between genes i.e. only genes that are directly or indirectly influenced by a stimulus change expression. GWAS genes on the other hand can be associated to a disease for very distinct biological reasons even if they are connected in a network.

A different approach to reduce the sizes of the networks was presented in section 3.1.4. Edges are removed that connect nodes which have diseases associated with them that contradict each other. This approach removed many edges and created many unconnected nodes. This might indicate that more disease associations will be discovered in the future for genes that are already associated with some inflammatory diseases. But still, this approach is very radical and with the exception of Crohn's disease (Figure 3.22) there were no larger submodules (> 3) specific for single diseases that could be found.

A different potential approach to reduce the number of edges is to use gene ontology annotations and determine how similar the nodes are that are connected by an edge based on term annotation frequency⁶⁰. A network can then be simplified by subsequently removing the edges that connect the most dissimilar nodes up to the point where all

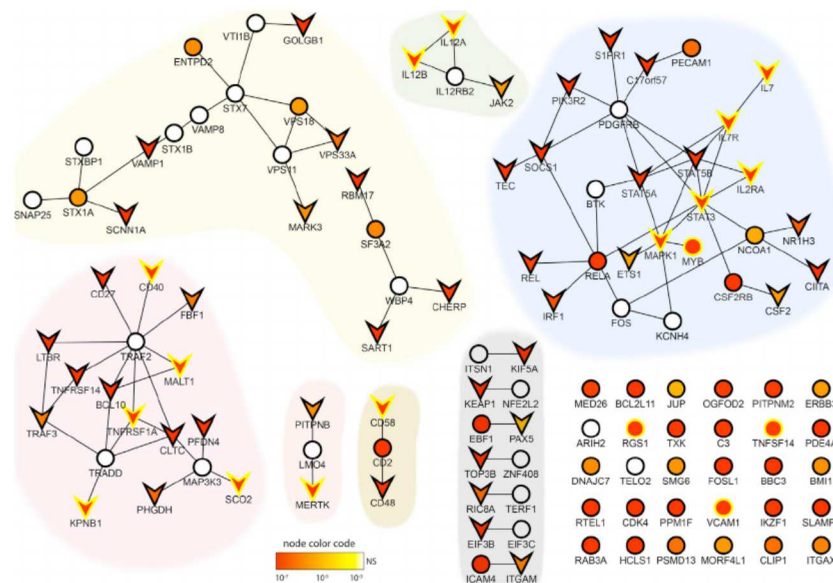


Figure 4.3.: Figure 3 from paper "Network-Based Multiple Sclerosis Pathway Analysis with GWAS Data from 15,000 Cases and 30,000 Controls" by the International Multiple Sclerosis Genetics Consortium. The displayed subnetworks have been determined by the jActiveModules Cytoscape plugin. In total 57 non-HLA loci with disease-association with multiple sclerosis were used in the analyses of that paper.

disease nodes are still connected. This makes it possible to prioritise edges for further investigation and also to make the visualization of the networks less convoluted because network layout algorithms tend to cluster densely connected modules together¹²⁹.

4.4. Enrichment Findings

There are several enrichment tools. DAVID appears to be one of the most popular services for enrichment which is probably due to the fact that it consolidates a lot of different gene annotation sources and can also handle different types of gene and protein IDs.

When doing enrichment analyses, biases of the data have to be taken into account. Because this work is based on SNPs which were largely obtained with the Immunochip, a special background with immune-system genes needs to be supplied to DAVID to compensate for this bias. But even with this background correction, the enrichment results contained many immune-system related terms which in turn indicates that it is worthwhile to focus mainly on the immune system for further research².

It was hoped that the enrichment analyses would also turn up known pathways. DAVID supports enrichment for KEGG pathways but these never showed up in the results. However, gene ontology terms for the JAK-STAT pathway were enriched in several cases.

DEPICT also does enrichment to find the best consensus among genes. It detected several genesets which have names of the form "*G* PPI subnetwork" where "*G*" is some gene. These subnetwork-genesets are increasingly common the fewer diseases are specifically associated with them: Figure 3.25 lists many such subnetwork genesets while figure 3.29 lists only one: The subnetwork that is common for all five diseases is the IL2 subnetwork. IL2 is involved in the regulation of the T-cell based immune response¹³⁰ and it is plausible that it is relevant for all of these inflammatory diseases. Other subnetworks detected by DEPICT are also relevant for immune system function. The KIT PPI subnetwork is a diverse signaling network which also influences STAT signaling. Some surprising terms are the MRPL44 and MRPS14 subnetworks because MRPL44 (RM44_HUMAN) and MRPL44 (RT14_HUMAN) are components of the mitochondrial ribosomes¹³¹. The MRPL44 term is exclusively enriched for the diseases AS, CD and UC while the MRPS14 term is exclusive for UC¹³². Other subnetworks include the BID subnetwork that is relevant for apoptosis, which is also a common enrichment term detected by DAVID. The connection between inflammatory diseases and apoptosis is well known¹³³.

When dealing with a list of genes, enrichment analyses are a convenient approach to get a quick overview on common functionality of these genes. However, a closer inspection of the genes is still needed to understand their role in disease⁶¹.

DEPICT also performed a tissue enrichment based on expression profiles from existing background data. The results do not appear to fit the diseases well. For ankylosing spondylitis there should be enrichment for bone tissue, for psoriasis there should be skin tissue and for primary sclerosing cholangitis there should be liver or bile tissue but these were not observed for these diseases. However, this is not a new observation. In the case of AS it has been previously noted that genes determined through GWAS are not known to be involved in the ankylosing process of the spine²⁴.

In contrast the tissues enriched for Crohn's disease and ulcerative colitis included parts of the gastrointestinal tract. All diseases had enrichment for blood and immune-system related tissues which makes sense for inflammatory diseases. But overall there are many

mismatches between the organs known to be affected by a disease and the enriched tissues.

Li et al. report a pathway association of IBD with the EGF pathway⁶⁹ based on enrichment analyses. They used VEGAS to map SNP p-values to gene p-values. They chose a significance threshold of 0.05 to distinguish between "significant genes" and "nonsignificant genes". However, they also state that only few of the 18 genes from the EGF pathway are highly significant which makes an enrichment less strong.

DEPICT detected "EGFR PPI subnetwork" as an enrichment term that was only associated with Crohn's disease (Figure 3.25) but it was a weak signal and several other subnetwork terms were also detected. Apart from this finding there is no considerable signal for the EGF pathway in the results.

The genes that DEPICT determined were enriched for terms that are related to signaling and regulation. It should be kept in mind that DEPICT tries to find genes that are similar when mapping SNPs to genes. This could in principle cause a bias in the enrichment results. However, the categories "signaling" and "regulation" were determined as meta-enrichment terms based on manual inspection of the various enrichment results. Different terms like transcription factors and protein receptors describe different mechanisms of regulation.

DEPICT scored the genes and the highest-scoring genes for each significant reconstituted geneset can be seen in figures 3.25 to 3.29. The gene with the highest overall score is GPBAR1 within the genesets *cell activation in immune response* and *leukocyte activation involved in immune response*. GPBAR1 is a receptor for bile acids. It has the highest score for Crohn's disease while the score for PSC is rather moderate even though bile acid is much more fitting for PSC than for CD. It is hypothesized that it is involved in the suppression of macrophage function by bile acids¹³⁴ which makes it a good candidate for a disease gene in inflammatory diseases.

GoShifter is another form of enrichment which tries to assess whether an overlap of a set with SNPs in a linkage region with genomic annotations is by chance or a real signal. Downsides of this approach are that it is not clearly defined where to put the score cutoff (the corresponding paper⁷² presents a hit with a score of 10%). Furthermore, the score is only assigned to the lead SNP even though it is often the LD SNPs that have an overlap with the annotations. GoShifter produces a global p-value for all p-values and all annotations together. This global p-value can be significant (Table 3.17) even if many

scores of the individual lead SNPs are only moderate. This makes the interpretation of the results challenging.

4.5. Methodological Considerations

Going from GWAS SNPs to network submodules requires several steps. Each of these steps has uncertainties that harbour the danger of failing to transfer important information to the next step but also to transferring irrelevant information and thereby increasing the noise in the results.

And once a result has been obtained it is often impractical to validate it directly because it would require expensive experiments^{135;14}. To date there is no complex disease that has been sufficiently well understood so that it is possible to create analysis pipelines that are validated with the disease. But even with such a validated pipeline there would probably still be the need for exploratory analysis.

In their review Jia et al.⁴⁸ enumerate various categories of network analyses (see section 1.5.3). But they do not explain in which situations one type of analysis should be preferred over another. Based on the independent categories they describe, a total of $2^4 \times 3 = 48$ different types of approaches to analyse networks are possible. And each of these approaches has further parameters and data sources that need to be chosen. Each choice has the potential to change the results.

In wet labs experiments there are usually positive and negative controls to determine if an experimental procedure worked correctly. This is something that is still mostly missing in network science of inflammatory diseases. We know about certain pathways that are involved in IBD, like JAK-STAT signaling¹⁰¹. We also find JAK-STAT signaling in our data which to some extent serves as a positive control. But with so many genetic factors more positive controls would be desirable and negative controls to detect noise would also be very helpful.

There is currently no gold standard for mapping SNPs to genes and genesets¹² and it is still an open question how to do this properly. For coding SNPs the answer is trivial, but even then it is still conceivable that such a SNP may have additional effects on RNA-related regulation.

With these problems in mind, network workflows can still serve as a method to generate hypotheses in the context of the genetics of inflammatory diseases.

4.5.1. Automation, Reproducibility and Repeatability

A common method to "validate" results is to create random results and to determine the similarity of the real results to the distribution of random results. If the real result is significantly different from most of the random results it can be assumed that these results did not just occur by chance. However, the biological relevance of the results is still not proven.

When generating such random results it would be desirable to store the random results in a file so that it is possible to reproduce the significance of the real results in the future deterministically instead of randomly. Even if converging random processes like Monte Carlo simulations are used, there will still be minor differences in the results¹³⁶ that can later lead to uncertainty about successful reproduction.

GoShifter uses randomness internally and when it is running it prints the random seed used to generate the random numbers. But there is no way to provide such a random seed to GoShifter to make it run deterministically. Other tools used for this thesis that make use of randomness include VEGAS, jActiveModules and PINBPA and they provide no facilities to control randomness.

In scientific work it is helpful to automate as much as possible. This makes it possible to improve details in a data processing pipeline and to automatically perform all the steps that are needed to compute an updated result. Graphical tools like Cytoscape have the fundamental problem that they are not fully scriptable and therefore not all workflows can be automated. Automation scripts also implicitly document the steps that were performed with the data. Even if complete reproducibility cannot be guaranteed, automation provides a way to ensure repeatability¹³⁷ of the steps that were performed and to inspect every step for errors.

Given that exploratory data analysis is required for network analyses, the Jupyter notebook proved to be a valuable tool to directly execute data transformations but also write a rationale for every step directly alongside the transformative code. This strikes a good balance between exploratory analysis and full automation while still documenting each individual step and the thought process so that it is later possible to follow the reasoning of the individual analysis steps and potentially find flaws. The relevant notebook sessions are attached to this thesis as supplementary data.

A central problem of reproducibility is software dependencies. A software called dmGWAS¹³⁸ appeared to be a promising tool to work with GWAS data and networks. But

dmGWAS depends on the igraph library¹³⁹. After the official release of dmGWAS, this library changed its behaviour regarding indexing of arrays (zero-based versus one-based). This made dmGWAS unusable because the original assumptions in the dmGWAS code regarding indexing were no longer true. A remedy to this problem is to package up scientific software together with all dependencies. The Docker virtualisation infrastructure¹⁴⁰ appears to be a good solution for this because it makes it possible to bundle all needed software dependencies at a specific version together with the main software as one package of moderate size. Docker has been used for some parts of this work (GoShifter, DEPICT, Jupyter Notebook) but it will not be discussed in detail because it did not affect the scientific results themselves and is also unlikely to affect the speed of the computations¹⁴⁰.

4.6. Future work

Complex diseases are the result of many different factors working together. Unfortunately, we only have a limited resolution on the genetic level when it comes to causative variants. Furthermore, environmental factors also play a major role that should be taken into account if possible.

Given the great number of factors it is unlikely that all factors are involved in every case when a disease develops. In fact, not every patient carries all disease variants. The etiology should therefore rather be inspected on the individual level: Which risk factors does a patient actually have? Where do these factors occur on the network level? It is very likely that there are several different etiologies for these diseases and it is therefore difficult to consider them all at once.

It is well known that there are different subtypes of these diseases^{141;26}. A better stratification between patients with different subtypes could lead to a drastically simpler picture. But then again, the knowledge about the individual subtypes of patients might simply not have been recorded when collecting data for the studies.

When working with the genetic profiles of individuals, the actual SNPs can be mapped into a gene/protein interaction network. This network might be sparser with disease genes than the networks presented in this thesis. It is likely that several factors act together to cause disease. In logical terms this can be expressed as an AND conjunction. One approach to predict such conjunctions would be to determine the nearest disease node neighbour(s)

of every disease node. Then a combined risk score could be calculated to see if the combination of these neighbouring SNPs is significantly more or less common in people who have the disease in contrast to those who do not have it. Other factors are independent of one another and they could be expressed in a logical OR disjunction. Predicting such disjunctions could be done by overlaying the networks of individual patients and controls: Every disease node in one patient tries to find the nearest neighbour(s) that are not disease nodes in the current network but are marked as disease nodes in one of the other patient networks.

Another approach to simplify the analysis would be to cluster patients based on their genotypes. This way it might be possible to stratify for different disease subtypes on the genetic level.

We know that the odds ratios or risk scores of SNPs are only providing a tendency for developing the disease. The just described approach does not try to express the chance of developing a disease in boolean terms. But it tries to capture the assumption that genetic factors act together or can act independently while still causing similar phenotypes. The relationships between these connected genes could then be more closely investigated because they might provide additional understanding how the diseases work. Special attention should be given to genes that act at the interface between host and microbes and genes that are known to be affected by smoking. The connections between genetic factors and environmental factors should be more closely investigated.

When mapping SNPs to genes, a combination of different approaches/tools should be used because different approaches are able to capture different genomic structures better. For every single SNP it should be decided which approach yields the most plausible biological association and then use that association for the mapping to genes. In the case of transcription factor binding sites, it should be attempted to determine both, the gene and the transcription factors that bind to this site.

It could also be attempted to use disease-associated chromatin and histone marks to determine further genes that may be regulated by them. And it should also be investigated if there are any overlaps of SNPs with areas that code for long non-coding RNAs or other types of RNAs because it has already been shown that microRNAs play a role in Crohn's disease¹⁴².

Future network analyses should in general focus much more on the edges between genes and less on the nodes themselves. Every true edge describes a biological process or function. There is the problem of false positive edges that describe non-existing

interactions or interactions of biological irrelevance (i.e. interactions that do not have any notable effects on the capability of the proteins to execute their functions). It is therefore important to develop methods that prioritise edges and remove edges of less importance from the network. One such approach could be to assign a score to each edge depending on how many GO terms the connected nodes share normalized by the background distribution of these GO terms. Removing edges has also major advantages for visualizing networks because well-connected networks tend to form complex clusters that are visually hard to dissect.

There are more inflammatory diseases that could be investigated in relation to the five diseases to get a better understanding of what is really typical for inflammatory diseases. The five diseases studied in this thesis have been chosen because they are very similar to each other on the genetic level¹. A comparison with more distantly "related" diseases could provide further insights on common and different disease mechanisms.

This thesis also presented some interesting candidates for new disease genes based on the principle of guilt-by-association (section 4.1.2). A closer investigation of these genes could be of interest.

And finally a combination of different reference networks should be used to create networks that include not only protein-protein interactions but also transcriptional regulators and translational regulators.

Chapter 5.

Conclusions

“We do what we must because we can. For the good of all of us. Except the ones who are dead.”

— Mad robot GLaDOS (Lyrics)

When constructing networks, we try to establish a digital approximation of what biology is like. Establishing a good approximation is a real challenge, because we often do not have a good enough resolution on the molecular level to understand which components interact and how they interact. But further research will continuously add more observations and might provide a better understanding of the mechanisms behind the diseases.

The observations made in this thesis indicate that the biology of the diseases is dependent on signaling and regulation to a large degree. This conforms well with the common assumption that - at least in the case of inflammatory bowel disease - the disease is caused by an aberrant immune response to environmental triggers¹⁴³. It is plausible that the cellular coordination in patients is not adjusted well and therefore the immune system overreacts.

It is still unclear how to best link associated loci to genes. And this is the foundation for all further network analyses. More research is needed to understand DNA-based regulatory mechanisms better in general.

PPI networks tend to be noisy and contain many edges that are probably not relevant for the biological phenomenon in question. A better understanding on how genes/proteins

work together would be helpful, to prioritise putative relevant edges to get a clearer picture on the network level.

Therefore it can be concluded that the overarching primary aim with the networks is to get a good signal-to-noise ratio. After that, the next challenge is to understand the biology in the networks. Depending on how much is already known about the interactions, this might be a feasible endeavor. But there are still a lot of gaps in our knowledge about molecular biology that we need to fill in.

Further research needed.

Appendix A.

Zusammenfassung

Die Genetik von fünf chronischen Entzündungserkrankungen wurde mithilfe von Netzwerk- und Enrichmentmethoden untersucht. Viele Ansätze wurden ausprobiert um Netzwerke bestehend aus krankheitsassoziierten Genen zu konstruieren. Große Netzwerke wurden in Teilnetzwerke aufgeteilt mit dem Ziel Module zu finden, die spezifische funktionale zelluläre Funktionen repräsentieren. Ein klares Bild gab es in der Regel nicht. Generell konnte aber beobachtet werden, dass Regulation und Signalübertragung wesentliche Konzepte der krankheitsassoziierten Gene sind und weitere Forschung daher sich darauf fokussieren sollte diese Signal- und Regulationsprozesse besser zu verstehen, um potentiell herauszufinden, wie die Gene Entzündungsprozesse und Interaktionen mit Umweltfaktoren beeinflussen.

Mehrere nicht-Krankheitsgene wurden gefunden die direkt mit Krankheitsgenen interagieren und daher genauer untersucht werden sollten, weil sie nach dem Prinzip *guilt-by-association* gute Kandidaten für noch unbekannte Krankheitsgene sind.

Die fünf Erkrankungen, die in dieser Arbeit untersucht wurden, sind sich auf genetischer Ebene sehr ähnlich. Diese Ähnlichkeit war auch auf der Netzwerkebene sichtbar. Viele Gene sind möglicherweise relevant bei mehreren Krankheiten gleichzeitig und die krankheitsspezifischen Teilnetzwerke überlappen sich. Allerdings wurde nur im Falle von Morbus Crohn ein größeres Teilnetzwerk gefunden, dass exklusiv nur für diese Erkrankung spezifisch ist.

Wenn man die Effektrichtung der SNPs für die einzelnen Erkrankungen berücksichtigt, so wurden keine klaren Tendenzen im Zusammenhang mit den Lokus-assoziierten DNA-Bindeelementen (z.B. Transkriptionsfaktoren) auf der Protein-Protein Interaktionsebene gefunden.

Die Beobachtungen in dieser Arbeit bestätigen die gängige Sichtweise, dass das Immunsystem eine wichtige Rolle bei der Entstehung der Krankheiten spielt, weil viele Immunsystem-relevante Gene in den Analysen gefunden wurden und weil die Enrichment-Ergebnisse dies ebenfalls bestätigen.

Appendix B.

Summary

The genetics of five chronic inflammatory diseases have been investigated with network and enrichment methods. Various approaches have been tried out to construct networks of disease-associated genes. Networks of great size have been split up into subnetworks with the aim to find modules that represent specific cellular functions. However, a clear picture could usually not be obtained. But in general it could be observed that regulation and signaling are major themes of disease-associated genes and further research should therefore focus on understanding these signaling and regulation processes better to potentially find out how they influence inflammation or interactions with environmental factors.

Several non-disease-associated genes have been found to be directly connected to disease genes and should therefore be investigated more closely as they are good candidates for being yet unknown disease genes based on the principle of *guilt-by-association*.

The five diseases investigated in this thesis are very similar on the genetic level. This similarity was also visible on the network level. Many genes are putatively shared among the diseases and disease-specific subnetworks overlap each other. However, only in the case of Crohn's disease there was a greater subnetwork that was specific to this disease. All other diseases had no greater (> 3 nodes) exclusively disease-specific subnetworks.

When trying to take into account the effect direction of SNPs for individual diseases, no clear tendencies were observed when investigating the locus-associated DNA-binding proteins (e.g. transcription factors) on the PPI-network level.

The findings in this thesis confirm the common view that the immune system plays an important role in the etiology of the diseases because many immune system-relevant genes were found in the networks and the enrichment results also confirmed their relevance.

Appendix C.

Supplementary Chapters

C.1. A few notes on the Microbiome

The microbiome is not a focus of this thesis. But it is still a very important biological factor of several inflammatory diseases and is therefore worthy of a few notes to provide a more complete picture of the relevant factors of the diseases presented in this thesis.

There are three large interfaces where human cells and microbes meet: The skin, the respiratory tract and the gastrointestinal tract¹⁸. Microbes inhabit or come in contact with these body regions and therefore every human contains a huge number of microbes in its body. It is often said that the ratio between microbes and human cells is ten to one¹⁴⁴, but recently Sender et al. argued that these numbers might not be accurate and suggest that the ratio should be one to one¹⁴⁵. In any case, the human body harbours a large number of microbes that can be commensal or pathogenic.

Many microbes have a symbiotic relationship with their human host. In the simplest cases, their purpose is just to take up a niche so that no other potentially harmful bacteria can colonize on the tissue¹⁴⁶. Some bacteria also have immunomodulating abilities. They regulate immune tolerance¹⁴⁷. Then there are also bacteria that actively fight pathogens that could harm the host. In a sense, they are an extension of the human immune system¹⁴⁸. They also play major roles in the development and training of the immune system¹⁹.

The human body consists of many different organs which serve different functions. Organs that are open to microbes tend to have specific microbiome populations i.e. the

bacterial species found in the large intestine differ from the small intestine because each of these organs is a different ecological niche²¹.

It is very likely that humans need a healthy microbiome to be healthy. It is thought that disruption or even extinction of microbial communities can have severe effects on the human body⁵. In patients with inflammatory bowel disease a common observation is that the diversity of the microbiome is severely reduced¹⁴⁷. Many microbes probably can not survive the chronic inflammation in their host while other microbial strains are able to deal with it and rise in numbers¹⁴⁷.

It is unclear whether the lack of diversity is a cause or a consequence of inflammatory diseases. Due to lifestyle it is conceivable that the immune system of humans in more sterile environments has a lot less exposure to microbes in early years of life and is therefore not properly trained to distinguish harmless microbes from bad microbes¹⁴⁹.

However, it should also be noted that in inflammatory bowel disease the mucosal barriers are repeatedly broken and microbes travel into tissue²¹. This in turn also leads to inflammatory responses to kill off the invading microbes²¹.

C.2. Modifications to VEGAS

VEGAS⁶² was downloaded and installed locally. In order to run with our data it had to be patched to avoid crashes. A reference to a user's home directory had to be removed from the source code. Furthermore, VEGAS makes use of the `corpcor` R package to correct for non-positive definite correlation matrices. However, apparently this correction does not always succeed because several parts of the VEGAS source code make additional adjustments to the matrices:

```
library ( corpcor )

reps <- $_[0]

if ( is . positive . definite ( co ) == F ) {
  co <- make . positive . definite ( co )
}
if ( is . positive . definite ( co ) == F ) {
```

```
matrix(scan('plink.ld', quiet=T), nc=numsnp) -> co
for(i in 1:numsnp){
  co[i, i] <- 1.0001
}
}
if(is.positive.definite(co)==F){
  for(i in 1:numsnp){
    co[i, i] <- 1.001
  }
}
if(is.positive.definite(co)==F){
  for(i in 1:numsnp){
    co[i, i] <- 1.01
  }
}
}
```

And there is no final check for success. VEGAS crashed on our dataset. Adding the following code removed the crash:

```
if(is.positive.definite(co)==F){
  for(i in 1:numsnp){
    co[i, i] <- 1.1
  }
}
}
```

While these flaws in software design were some reason for concern, the mathematics behind the algorithm could not be challenged and the tool is still widely used in the research community^{68;69;16}. For this reason it was still used for parts of this thesis.

List of figures

1.1. Risk allele distribution in IBD	3
1.2. Distribution of risk odds ratios in the population	9
1.3. Shared loci among the five diseases studied in this thesis. Image taken from Ellinghaus et al ² , supplementary figure 2	10
1.4. Exemplary power law	13
2.1. Generalized workflow for GWAS and networks	21
2.2. Cross-Disease SNP association counts	21
3.1. Linker nodes of cross-disease genes	39
3.2. Permutation tests without MHC region	42
3.3. Permutation tests with MHC region	42
3.4. Best ankylosing spondylitisjActiveModule	43
3.5. Best jActiveModule for Crohn’s disease	43
3.6. Best jActiveModule for psoriasis	43
3.7. Second-best jActiveModule for psoriasis	44
3.8. Best jActiveModule for primary sclerosing cholangitis	45
3.9. Best jActiveModule for ulcerative colitis	45
3.10. Best jActiveModule for ankylosing spondylitis with MHC region	46
3.11. Best jActiveModule for Crohn’s disease with MHC region	47

3.12. Best jActiveModule for psoriasis with MHC region	47
3.13. Best jActiveModule for primary sclerosing cholangitis with MHC region .	48
3.14. Best jActiveModule for ulcerative colitis with MHC	48
3.15. Node legend for number of diseases	49
3.16. Nodes with different numbers of associated disease and their interactions (circle layout)	50
3.17. Nodes with different numbers of associated disease and their interactions (pentagram layout)	51
3.18. Nodes associated with all five diseases	52
3.19. Nodes associated with exactly four diseases	53
3.20. Nodes associated with exactly three diseases	53
3.21. Nodes associated with exactly two diseases	54
3.22. Nodes associated with exactly one disease	54
3.23. Overlap of significant geneset terms between different diseases according to DEPICT.	56
3.24. Heatmap color key	56
3.25. Enriched genesets for every disease that are specific for a single disease (according to DEPICT)	57
3.26. Enriched genesets for every disease that are specific for exactly 2 diseases (according to DEPICT)	58
3.27. Enriched genesets for every disease that are specific for exactly 3 diseases (according to DEPICT)	59
3.28. Enriched genesets for every disease that are specific for exactly 4 diseases (according to DEPICT)	60
3.29. Enriched genesets for every disease that are specific for exactly 5 diseases (according to DEPICT)	61
3.30. Overlap of significant tissues between different diseases according to DEPICT.	62

3.31. Enriched tissues for every disease that are specific for a single disease. No significant tissues were detected that are exclusively specific for psoriasis or PSC	63
3.32. Enriched tissues for every disease that are specific for a exactly 2 diseases	64
3.33. Enriched tissues for every disease that are specific for a exactly 3 diseases	65
3.34. Enriched tissues for every disease that are specific for a exactly 4 diseases	66
3.35. Enriched tissues for every disease that are specific for a exactly 5 diseases	67
3.36. Color legend for genes associated with different diseases	68
3.37. Largest connected component (DEPICT-based)	69
3.38. Nodes outside of largest connected component (DEPICT-based)	70
3.39. Subnetwork for ankylosing spondylitis (with linker genes). AS-associated nodes have a white color component.	76
3.40. Subnetwork for Crohn's disease (without linker genes).	76
3.41. Subnetwork for psoriasis (with linker genes). PS-associated nodes have a pink color component.	76
3.42. Subnetwork for primary sclerosing cholangitis (with linker genes). PSC-associated nodes have a green color component.	77
3.43. Subnetwork for ulcerative colitis (without linker nodes).	77
3.44. Stratification tests of ENCODE annotations with GoShifter	86
3.45. Histogram of LD SNPs and their linked genes (in both directions)	88
3.46. Protein-Protein interactions between "protective genes" and their direct interaction partners	93
3.47. Protective genes (green) and genes that are enriched for the vitamin D receptor (blue)	94
3.48. degree centrality of PS nodes	99
3.49. betweenness centrality of PSC nodes	100
3.50. closeness centrality of AS nodes	101

3.51. closeness centrality of CD nodes	102
3.52. eigenvector centrality of UC nodes	103
4.1. The EGF/MAPK signaling pathway	111
4.2. Submodules produced by jActiveModules (from Ideker et al. ⁵⁴)	117
4.3. MS Submodules produced by jActiveModules (from the IMSGC ¹⁶)	118

List of tables

1.	List of abbreviations	v
1.1.	Inflamed Organs in different inflammatory diseases	8
2.1.	Sizes of ConsensusPathDB reference networks	22
2.2.	Important options that were used when running DEPICT. The list of SNPs differed between each run.	25
3.1.	Number of genes reported by VEGAS with and without MHC region for different diseases	37
3.2.	Numbers of significant genes detected by VEGAS for different diseases	37
3.3.	Counts of VEGAS p-values that are exactly zero	37
3.4.	Sizes of disease-specific subnetworks	38
3.5.	Percentiles of random networks (edge randomisation)	41
3.6.	Scores from jActiveModules subnetworks	41
3.7.	Number of edges between different classes of disease nodes	50
3.8.	Numbers of enriched genesets	55
3.9.	Numbers of enriched tissues	62
3.10.	Descriptions of proteins from largest connected component.	71
3.11.	Number of disease-associated nodes and linker nodes	78
3.12.	Nodes associated to AS-associated genes	79
3.13.	Nodes associated to CD-associated genes	80

3.14. Nodes associated to PS-associated genes	81
3.15. Nodes associated to PSC-associated genes	82
3.16. Nodes associated to UC-associated genes	84
3.17. ENCODE annotations and GoShifter p-values	85
3.18. LD SNP risk class counts for ENCODE annotations	87
3.19. Risk classifications (based on the minor allele in the study populations from Ellinghaus et al. ²) of DNA-binding proteins for every disease	87
3.20. Best DAVID enrichment results for genes with only protective variants in TFBSs.	89
3.21. Risk status of proteins without high-confidence interactions	104

Bibliography

- [1] D. Ellinghaus, J. Bethune, B.-S. Petersen, and A. Franke, “The genetics of Crohn’s disease and ulcerative colitis – status quo and beyond,” *Scandinavian Journal of Gastroenterology*, vol. 50, pp. 13–23, Jan. 2015.
- [2] D. Ellinghaus, L. Jostins, S. L. Spain, A. Cortes, J. Bethune, B. Han, Y. R. Park, S. Raychaudhuri, J. G. Pouget, M. Hübenthal, T. Folseraas, Y. Wang, T. Esko, A. Metspalu, H.-J. Westra, L. Franke, T. H. Pers, R. K. Weersma, V. Collij, M. D’Amato, J. Halfvarson, A. B. Jensen, W. Lieb, F. Degenhardt, A. J. Forstner, A. Hofmann, The International IBD Genetics Consortium (iIBDGC), International Genetics of Ankylosing Spondylitis Consortium (iGAS), International PSC Study Group (iPSCSG), Genetic Analysis of Psoriasis Consortium (GAPC), Psoriasis Association Genetics Extension (PAGE), S. Schreiber, U. Mrowietz, B. D. Juran, K. N. Lazaridis, S. Brunak, A. M. Dale, R. C. Trembath, S. Weidinger, M. Weichen-
thal, E. Ellinghaus, J. T. Elder, J. N. W. N. Barker, O. A. Andreassen, D. P. McGovern, T. H. Karlsen, J. C. Barrett, M. Parkes, M. A. Brown, and A. Franke, “Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci,” *Nature Genetics*, vol. advance online publication, Mar. 2016.
- [3] C. Aranow, “Vitamin D and the Immune System,” *Journal of Investigative Medicine : the official publication of the American Federation for Clinical Research*, vol. 59, pp. 881–886, Aug. 2011.
- [4] M. I. Qadir, “Review - Skin cancer: Etiology and management,” *Pakistan Journal of Pharmaceutical Sciences*, vol. 29, pp. 999–1003, May 2016.
- [5] D. R. Littman and E. G. Pamer, “Role of the commensal microbiota in normal and pathogenic host immune responses,” *Cell Host & Microbe*, vol. 10, pp. 311–323, Oct. 2011.

- [6] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell, Fourth Edition*. Garland Science, 4 ed., 2004. Published: Hardcover.
- [7] L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J.-P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Büning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskis, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, International IBD Genetics Consortium (IIBDGC), M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes, S. Vermeire, J. C. Barrett, and J. H. Cho, "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease," *Nature*, vol. 491, pp. 119–124, Nov. 2012.
- [8] A. Ziegler, I. R. König, and F. Pahlke, *A statistical approach to genetic epidemiology*. Weinheim: Wiley-VCH, 2010. OCLC: 898722762.
- [9] S. Bhattacharjee, P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, P. Hartge, M. Yeager, C. C. Chung, S. J. Chanock, and N. Chatterjee, "A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits," *The American Journal of Human Genetics*, vol. 90, pp. 821–835, May 2012.
- [10] M. Parkes, A. Cortes, D. A. van Heel, and M. A. Brown, "Genetic insights into common pathways and complex relationships among immune-mediated diseases," *Nature reviews. Genetics*, vol. 14, pp. 661–673, Sept. 2013.

- [11] B. M. Neale, ed., *Statistical genetics: gene mapping through linkage and association*. New York: Taylor & Francis Group, 2008.
- [12] A. Al-Chalabi and L. Almasy, eds., *Genetics of complex human diseases: a laboratory manual*. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press, 2009.
- [13] I. Pe'er, R. Yelensky, D. Altshuler, and M. J. Daly, "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants," *Genetic Epidemiology*, vol. 32, pp. 381–385, May 2008.
- [14] S. Siegert, A. Wolf, D. N. Cooper, M. Krawczak, and M. Nothnagel, "Mutations Causing Complex Disease May under Certain Circumstances Be Protective in an Epidemiological Sense," *PLOS ONE*, vol. 10, p. e0132150, July 2015.
- [15] N. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher, "Pitfalls of predicting complex traits from SNPs," *Nature Reviews. Genetics*, vol. 14, pp. 507–515, July 2013.
- [16] International Multiple Sclerosis Genetics Consortium, "Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls," *American journal of human genetics*, vol. 92, pp. 854–865, June 2013.
- [17] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Research*, vol. 42, pp. D1001–1006, Jan. 2014.
- [18] A. K. A. MBBS, A. H. H. L. M. PhD, and S. P. M. PhD, *Cellular and Molecular Immunology*, 8e. Philadelphia, PA: Saunders, 8 edition ed., Aug. 2014.
- [19] J. Chow, S. M. Lee, Y. Shen, A. Khosravi, and S. K. Mazmanian, "Host–Bacterial Symbiosis in Health and Disease," *Advances in immunology*, vol. 107, pp. 243–274, 2010.
- [20] J. N. Fullerton and D. W. Gilroy, "Resolution of inflammation: a new therapeutic frontier," *Nature Reviews Drug Discovery*, vol. advance online publication, Mar. 2016.
- [21] A. A. Kühn, U. Erben, L. I. Kredel, and B. Siegmund, "Diversity of Intestinal Macrophages in Inflammatory Bowel Diseases," *Frontiers in Immunology*, vol. 6,

- p. 613, 2015.
- [22] G. Fonseca-Camarillo and J. K. Yamamoto-Furusho, “Immunoregulatory Pathways Involved in Inflammatory Bowel Disease,” *Inflammatory Bowel Diseases*, June 2015.
- [23] B. Khor, A. Gardet, and R. J. Xavier, “Genetics and pathogenesis of inflammatory bowel disease,” *Nature*, vol. 474, pp. 307–317, June 2011.
- [24] F. W. Tsui, H. W. Tsui, A. Akram, N. Haroon, and R. D. Inman, “The genetic basis of ankylosing spondylitis: new insights into disease pathogenesis,” *The Application of Clinical Genetics*, vol. 7, pp. 105–115, 2014.
- [25] M. D’Amato and J. D. Rioux, eds., *Molecular genetics of inflammatory bowel disease*. New York: Springer, 2013. OCLC: ocn874092920.
- [26] J. E. Eaton, J. A. Talwalkar, K. N. Lazaridis, G. J. Gores, and K. D. Lindor, “Pathogenesis of primary sclerosing cholangitis and advances in diagnosis and management,” *Gastroenterology*, vol. 145, pp. 521–536, Sept. 2013.
- [27] M. Underner, J. Perriot, J. Cosnes, P. Beau, G. Peiffer, and J.-C. Meurice, “Smoking, smoking cessation and Crohn’s disease,” *Presse Medicale (Paris, France: 1983)*, Mar. 2016.
- [28] M. M. Monick, L. S. Powers, K. Walters, N. Lovan, M. Zhang, A. Gerke, S. Hansdottir, and G. W. Hunninghake, “Identification of an autophagy defect in smokers’ alveolar macrophages,” *Journal of Immunology (Baltimore, Md.: 1950)*, vol. 185, pp. 5425–5435, Nov. 2010.
- [29] J. Hampe, A. Franke, P. Rosenstiel, A. Till, M. Teuber, K. Huse, M. Albrecht, G. Mayr, F. M. De La Vega, J. Briggs, S. Günther, N. J. Prescott, C. M. Onnie, R. Häslér, B. Sipos, U. R. Fölsch, T. Lengauer, M. Platzer, C. G. Mathew, M. Krawczak, and S. Schreiber, “A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG1611,” *Nature Genetics*, vol. 39, pp. 207–211, Feb. 2007.
- [30] J.-F. Bach, “The effect of infections on susceptibility to autoimmune and allergic diseases,” *The New England Journal of Medicine*, vol. 347, pp. 911–920, Sept. 2002.
- [31] J. A. Hamerman, J. Pottle, M. Ni, Y. He, Z.-Y. Zhang, and J. H. Buckner, “Negative regulation of TLR signaling in myeloid cells-implications for autoimmune diseases,”

- Immunological Reviews*, vol. 269, pp. 212–227, Jan. 2016.
- [32] H. H. Uhlig, “Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease,” *Gut*, vol. 62, pp. 1795–1805, Dec. 2013.
- [33] K. Szarc vel Szic, M. N. Ndlovu, G. Haegeman, and W. Vanden Berghe, “Nature or nurture: let food be your epigenetic medicine in chronic inflammatory disorders,” *Biochemical Pharmacology*, vol. 80, pp. 1816–1832, Dec. 2010.
- [34] K. J. Tracey, “Physiology and immunology of the cholinergic antiinflammatory pathway,” *The Journal of Clinical Investigation*, vol. 117, pp. 289–296, Feb. 2007.
- [35] J. F. Cryan, “Stress and the Microbiota-Gut-Brain Axis: An Evolving Concept in Psychiatry,” *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, vol. 61, pp. 201–203, Apr. 2016.
- [36] B. Maher, “Personal genomes: The case of the missing heritability,” *Nature News*, vol. 456, pp. 18–21, Nov. 2008.
- [37] Y. I. Li, B. v. d. Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, and J. K. Pritchard, “RNA splicing is a primary link between genetic variation and disease,” *Science*, vol. 352, pp. 600–604, Apr. 2016.
- [38] Y. Liu and M. R. Chance, “Pathway analyses and understanding disease associations,” *Current genetic medicine reports*, vol. 1, Dec. 2013.
- [39] L. He and G. J. Hannon, “MicroRNAs: small RNAs with a big role in gene regulation,” *Nature Reviews. Genetics*, vol. 5, pp. 522–531, July 2004.
- [40] J. M. Perkel, “Visiting "noncodarnia",” *BioTechniques*, vol. 54, pp. 301, 303–304, June 2013.
- [41] S. Oliver, “Proteomics: Guilt-by-association goes global,” *Nature*, vol. 403, pp. 601–603, Feb. 2000.
- [42] Y. Guo, X. Wei, J. Das, A. Grimson, S. M. Lipkin, A. G. Clark, and H. Yu, “Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle,” *American journal of human genetics*, vol. 93, pp. 78–89, July 2013.

- [43] S. Navlakha, A. Gitter, and Z. Bar-Joseph, “A network-based approach for predicting missing pathway interactions,” *PLoS computational biology*, vol. 8, no. 8, p. e1002640, 2012.
- [44] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, “The human disease network,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 8685–8690, May 2007.
- [45] L. Franke, H. v. Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga, “Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes,” *The American Journal of Human Genetics*, vol. 78, pp. 1011–1025, June 2006.
- [46] M. Newman, *Networks: An Introduction*. Oxford ; New York: Oxford University Press, 1 edition ed., May 2010.
- [47] T. Dagan, “Phylogenomic networks,” *Trends in Microbiology*, vol. 19, pp. 483–491, Oct. 2011.
- [48] P. Jia and Z. Zhao, “Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives,” *Human Genetics*, vol. 133, pp. 125–138, Feb. 2014.
- [49] S. Raychaudhuri, R. M. Plenge, E. J. Rossin, A. C. Y. Ng, I. S. Consortium, S. M. Purcell, P. Sklar, E. M. Scolnick, R. J. Xavier, D. Altshuler, and M. J. Daly, “Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions,” *PLOS Genet*, vol. 5, p. e1000534, June 2009.
- [50] M. Ahmadi, R. Jafari, S.-A. Marashi, and A. Farazmand, “Evidence for the relationship between the regulatory effects of microRNAs and attack robustness of biological networks,” *Computers in Biology and Medicine*, vol. 63, pp. 83–91, May 2015.
- [51] T. H. Pers, J. M. Karjalainen, Y. Chan, H.-J. Westra, A. R. Wood, J. Yang, J. C. Lui, S. Vedantam, S. Gustafsson, T. Esko, T. Frayling, E. K. Speliotes, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, M. Boehnke, S. Raychaudhuri, R. S. N. Fehrmann, J. N. Hirschhorn, and L. Franke, “Biological interpretation of genome-wide association studies using predicted gene functions,” *Nature Communications*, vol. 6, p. 5890, 2015.

- [52] G. Fehring, G. Liu, L. Briollais, P. Brennan, C. I. Amos, M. R. Spitz, H. Bickelböller, H. E. Wichmann, A. Risch, and R. J. Hung, “Comparison of pathway analysis approaches using lung cancer GWAS data sets,” *PloS One*, vol. 7, no. 2, p. e31816, 2012.
- [53] Q. Wang, P. Jia, K. T. Cuenco, Z. Zeng, E. Feingold, M. L. Marazita, L. Wang, and Z. Zhao, “Association signals unveiled by a comprehensive gene set enrichment analysis of dental caries genome-wide association studies,” *PloS One*, vol. 8, no. 8, p. e72653, 2013.
- [54] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, “Discovering regulatory and signalling circuits in molecular interaction networks.,” *Bioinformatics (Oxford, England)*, vol. 18 Suppl 1, pp. S233–S240, July 2002.
- [55] H. Huang and Q. Wu, “CRISPR Double Cutting through the Labyrinthine Architecture of 3d Genomes,” *Journal of Genetics and Genomics*, vol. 43, pp. 273–288, May 2016.
- [56] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene Ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, pp. 25–29, May 2000.
- [57] T. G. O. Consortium, “Gene Ontology Consortium: going forward,” *Nucleic Acids Research*, vol. 43, pp. D1049–D1056, Jan. 2015.
- [58] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 28, pp. 27–30, Jan. 2000.
- [59] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic Acids Research*, vol. 44, pp. D457–462, Jan. 2016.
- [60] P. Resnik, “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language,” *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1999.
- [61] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,” *Nucleic Acids Research*, vol. 37, pp. 1–13, Jan. 2009.

- [62] J. Z. Liu, A. F. McRae, D. R. Nyholt, S. E. Medland, N. R. Wray, K. M. Brown, AMFS Investigators, N. K. Hayward, G. W. Montgomery, P. M. Visscher, N. G. Martin, and S. Macgregor, “A versatile gene-based test for genome-wide association studies.” *American journal of human genetics*, vol. 87, pp. 139–145, July 2010.
- [63] L. Sachs and J. Hedderich, *Angewandte Statistik: Methodensammlung mit R ; mit 180 Tabellen*. Berlin: Springer, 12., vollst. neu bearb. Aufl. ed., 2006. OCLC: 180941317.
- [64] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig, “ConsensusPathDB—a database for integrating human functional interaction networks,” *Nucleic Acids Research*, vol. 37, pp. D623–D628, Jan. 2009.
- [65] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig, “ConsensusPathDB: toward a more complete picture of cell biology,” *Nucleic Acids Research*, vol. 39, pp. D712–D717, Jan. 2011.
- [66] A. Kamburov, U. Stelzl, and R. Herwig, “IntScore: a web tool for confidence scoring of biological interactions,” *Nucleic Acids Research*, vol. 40, pp. W140–146, July 2012.
- [67] S. Razick, G. Magklaras, and I. M. Donaldson, “iRefIndex: a consolidated protein interaction database with provenance,” *BMC bioinformatics*, vol. 9, p. 405, 2008.
- [68] N. R. Wray, M. L. Pergadia, D. H. R. Blackwood, B. W. J. H. Penninx, S. D. Gordon, D. R. Nyholt, S. Ripke, D. J. MacIntyre, K. A. McGhee, A. W. Maclean, J. H. Smit, J. J. Hottenga, G. Willemsen, C. M. Middeldorp, E. J. C. de Geus, C. M. Lewis, P. McGuffin, I. B. Hickie, E. J. C. G. van den Oord, J. Z. Liu, S. Macgregor, B. P. McEvoy, E. M. Byrne, S. E. Medland, D. J. Statham, A. K. Henders, A. C. Heath, G. W. Montgomery, N. G. Martin, D. I. Boomsma, P. a. F. Madden, and P. F. Sullivan, “Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned,” *Molecular Psychiatry*, vol. 17, pp. 36–48, Jan. 2012.
- [69] J. Li, Z. Wei, X. Chang, C. J. Cardinale, C. E. Kim, R. N. Baldassano, H. Hakonarson, and the International IBD Genetics Consortium, “Pathway-based Genome-wide Association Studies Reveal the Association Between Growth Factor Activity and Inflammatory Bowel Disease,” *Inflammatory Bowel Diseases*, Apr. 2016.

- [70] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, “PLINK: a tool set for whole-genome association and population-based linkage analyses,” *American Journal of Human Genetics*, vol. 81, pp. 559–575, Sept. 2007.
- [71] K. Lage, E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak, “A human phenome-interactome network of protein complexes implicated in genetic disorders,” *Nature Biotechnology*, vol. 25, pp. 309–316, Mar. 2007.
- [72] G. Trynka, H.-J. Westra, K. Slowikowski, X. Hu, H. Xu, B. Stranger, R. Klein, B. Han, and S. Raychaudhuri, “Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci,” *The American Journal of Human Genetics*, vol. 97, pp. 139–152, July 2015.
- [73] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik, “BigWig and BigBed: enabling browsing of large distributed datasets,” *Bioinformatics (Oxford, England)*, vol. 26, pp. 2204–2207, Sept. 2010.
- [74] C. Zuo, S. Shin, and S. Keleş, “atSNP: transcription factor binding affinity testing for regulatory SNP detection,” *Bioinformatics (Oxford, England)*, vol. 31, pp. 3353–3355, Oct. 2015.
- [75] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape: a software environment for integrated models of biomolecular interaction networks,” *Genome Research*, vol. 13, pp. 2498–2504, Nov. 2003.
- [76] J. H. Morris, A. Kuchinsky, T. E. Ferrin, and A. R. Pico, “enhancedGraphics: a Cytoscape app for enhanced node graphics,” *F1000Research*, vol. 3, p. 147, 2014.
- [77] K. Ono, T. Muetze, G. Kolishovski, P. Shannon, and B. Demchak, “CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API,” *F1000Research*, Aug. 2015.
- [78] R. Hickey, “The Clojure Programming Language,” in *Proceedings of the 2008 Symposium on Dynamic Languages, DLS '08*, (New York, NY, USA), pp. 1:1–1:1, ACM, 2008.
- [79] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring Network Structure, Dynamics, and Function using NetworkX,” in *Proceedings of the 7th Python in*

- Science Conference* (G. Varoquaux, T. Vaught, and J. Millman, eds.), (Pasadena, CA USA), pp. 11 – 15, 2008.
- [80] F. Pérez and B. E. Granger, “IPython: a System for Interactive Scientific Computing,” *Computing in Science and Engineering*, vol. 9, pp. 21–29, May 2007.
- [81] A. Alexandrescu, *The D Programming Language*. Upper Saddle River, NJ: Addison-Wesley Professional, 1 edition ed., June 2010.
- [82] N. Dianati, “Unwinding the hairball graph: Pruning algorithms for weighted complex networks,” *Physical Review. E*, vol. 93, p. 012304, Jan. 2016.
- [83] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, “Prioritizing candidate disease genes by network-based boosting of genome-wide association data,” *Genome Research*, vol. 21, pp. 1109–1121, July 2011.
- [84] C. R. Farber, “Systems-Level Analysis of Genome-Wide Association Data,” *G3: Genes/Genomes/Genetics*, vol. 3, pp. 119–129, Jan. 2013.
- [85] P. J. Lupardus and K. C. Garcia, “The structure of interleukin-23 reveals the molecular basis of p40 subunit sharing with interleukin-12,” *Journal of Molecular Biology*, vol. 382, pp. 931–941, Oct. 2008.
- [86] L. Wang, T. Matsushita, L. Madireddy, P. Mousavi, and S. Baranzini, “PINBPA: Cytoscape app for network analysis of GWAS data.,” *Bioinformatics (Oxford, England)*, Sept. 2014.
- [87] S. K. Yoshinaga, M. Zhang, J. Pistillo, T. Horan, S. D. Khare, K. Miner, M. Sonnenberg, T. Boone, D. Brankow, T. Dai, J. Delaney, H. Han, A. Hui, T. Kohno, R. Manoukian, J. S. Whoriskey, and M. A. Coccia, “Characterization of a new human B7-related protein: B7rp-1 is the ligand to the co-stimulatory protein ICOS,” *International Immunology*, vol. 12, pp. 1439–1447, Oct. 2000.
- [88] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, “High-resolution mapping and characterization of open chromatin across the genome,” *Cell*, vol. 132, pp. 311–322, Jan. 2008.
- [89] Y. Abreu-Delgado, R. A. Isidro, E. A. Torres, A. González, M. L. Cruz, A. A. Isidro, C. I. González-Keelan, P. Medero, and C. B. Appleyard, “Serum vitamin D and colonic vitamin D receptor in inflammatory bowel disease,” *World Journal of Gastroenterology*, vol. 22, pp. 3581–3591, Apr. 2016.

- [90] T. A. Kabbani, I. E. Koutroubakis, R. E. Schoen, C. Ramos-Rivers, N. Shah, J. Swoger, M. Regueiro, A. Barrie, M. Schwartz, J. G. Hashash, L. Baidoo, M. A. Dunn, and D. G. Binion, "Association of Vitamin D Level With Clinical Status in Inflammatory Bowel Disease: A 5-Year Longitudinal Study," *The American Journal of Gastroenterology*, Mar. 2016.
- [91] V. Bruzzese, A. Zullo, A. Piacchianti Diamanti, L. Ridola, R. Lorenzetti, C. Marrese, P. Scolieri, V. De Francesco, C. Hassan, A. Migliore, and B. Laganà, "Vitamin D deficiency in patients with either rheumatic diseases or inflammatory bowel diseases on biologic therapy," *Internal and Emergency Medicine*, Mar. 2016.
- [92] M. Ohira, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, and O. Ohara, "Characterization of a human homolog (BACH1) of the mouse Bach1 gene encoding a BTB-basic leucine zipper transcription factor and its mapping to chromosome 21q22.1," *Genomics*, vol. 47, pp. 300–306, Jan. 1998.
- [93] D. I. Loukinov, E. Pugacheva, S. Vatolin, S. D. Pack, H. Moon, I. Chernukhin, P. Mannan, E. Larsson, C. Kanduri, A. A. Vostrov, H. Cui, E. L. Niemitz, J. E. J. Rasko, F. M. Docquier, M. Kistler, J. J. Breen, Z. Zhuang, W. W. Quitschke, R. Renkawitz, E. M. Klenova, A. P. Feinberg, R. Ohlsson, H. C. Morse, and V. V. Lobanenko, "BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 6806–6811, May 2002.
- [94] C. I. Holmberg, V. Hietakangas, A. Mikhailov, J. O. Rantanen, M. Kallio, A. Meinander, J. Hellman, N. Morrice, C. MacKintosh, R. I. Morimoto, J. E. Eriksson, and L. Sistonen, "Phosphorylation of serine 230 promotes inducible transcriptional activity of heat shock factor 1," *The EMBO journal*, vol. 20, pp. 3800–3810, July 2001.
- [95] I. Györy, G. Fejér, N. Ghosh, E. Seto, and K. L. Wright, "Identification of a functionally impaired positive regulatory domain I binding factor 1 transcription repressor in myeloma cell lines," *Journal of Immunology (Baltimore, Md.: 1950)*, vol. 170, pp. 3125–3133, Mar. 2003.
- [96] A. D. Keller and T. Maniatis, "Identification and characterization of a novel repressor of beta-interferon gene expression," *Genes & Development*, vol. 5, pp. 868–879,

- May 1991.
- [97] T. L. A. Kawahara, E. Michishita, A. S. Adler, M. Damian, E. Berber, M. Lin, R. A. McCord, K. C. L. Ongaigui, L. D. Boxer, H. Y. Chang, and K. F. Chua, “SIRT6 links histone H3 lysine 9 deacetylation to NF-kappaB-dependent gene expression and organismal life span,” *Cell*, vol. 136, pp. 62–74, Jan. 2009.
- [98] R. A. Weinberg and R. A. Weinberg, *The Biology of Cancer, 2nd Edition*. New York: Garland Science, 2nd edition ed., May 2013.
- [99] D. Jäger, N. Halama, I. Zörnig, P. Klug, J. Krauss, and G.-M. Haag, “Immunotherapy of Colorectal Cancer,” *Oncology Research and Treatment*, vol. 39, no. 6, pp. 346–350, 2016.
- [100] V. Sibaud, N. Meyer, L. Lamant, E. Vigarios, J. Mazieres, and J. P. Delord, “Dermatologic complications of anti-PD-1/PD-L1 immune checkpoint antibodies,” *Current Opinion in Oncology*, vol. 28, pp. 254–263, July 2016.
- [101] R. Galien, “Janus kinases in inflammatory bowel disease: Four kinases for multiple purposes,” *Pharmacological reports: PR*, Apr. 2016.
- [102] R. B. Gupta, N. Harpaz, S. Itzkowitz, S. Hossain, S. Matula, A. Kornbluth, C. Bodian, and T. Ullman, “Histologic Inflammation Is a Risk Factor for Progression to Colorectal Neoplasia in Ulcerative Colitis: A Cohort Study,” *Gastroenterology*, vol. 133, pp. 1099–1105, Oct. 2007.
- [103] The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium, “Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways,” *Nature Neuroscience*, vol. 18, pp. 199–209, Feb. 2015.
- [104] K. Lucas and M. Maes, “Role of the Toll Like receptor (TLR) radical cycle in chronic inflammation: possible treatments targeting the TLR4 pathway,” *Molecular Neurobiology*, vol. 48, pp. 190–204, Aug. 2013.
- [105] M. Martin-Subero, G. Anderson, B. Kanchanatawan, M. Berk, and M. Maes, “Comorbidity between depression and inflammatory bowel disease explained by immune-inflammatory, oxidative, and nitrosative stress; tryptophan catabolite; and gut-brain pathways,” *CNS spectrums*, pp. 1–15, Aug. 2015.

- [106] R. Atreya, J. Mudter, S. Finotto, J. Müllberg, T. Jostock, S. Wirtz, M. Schütz, B. Bartsch, M. Holtmann, C. Becker, D. Strand, J. Czaja, J. F. Schlaak, H. A. Lehr, F. Autschbach, G. Schürmann, N. Nishimoto, K. Yoshizaki, H. Ito, T. Kishimoto, P. R. Galle, S. Rose-John, and M. F. Neurath, “Blockade of interleukin 6 trans signaling suppresses T-cell resistance against apoptosis in chronic intestinal inflammation: evidence in crohn disease and experimental colitis in vivo,” *Nature Medicine*, vol. 6, pp. 583–588, May 2000.
- [107] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyaavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos, “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA,” *Science*, vol. 337, pp. 1190–1195, Sept. 2012.
- [108] F. Pietrocola, S. Lachkar, D. P. Enot, M. Niso-Santano, J. M. Bravo-San Pedro, V. Sica, V. Izzo, M. C. Maiuri, F. Madeo, G. Mariño, and G. Kroemer, “Spermidine induces autophagy by inhibiting the acetyltransferase EP300,” *Cell Death and Differentiation*, vol. 22, pp. 509–516, Mar. 2015.
- [109] X. Yin, H. Q. Low, L. Wang, Y. Li, E. Ellinghaus, J. Han, X. Estivill, L. Sun, X. Zuo, C. Shen, C. Zhu, A. Zhang, F. Sanchez, L. Padyukov, J. J. Catanese, G. G. Krueger, K. C. Duffin, S. Mucha, M. Weichenthal, S. Weidinger, W. Lieb, J. N. Foo, Y. Li, K. Sim, H. Liany, I. Irwan, Y. Teo, C. T. S. Theng, R. Gupta, A. Bowcock, P. L. De Jager, A. A. Qureshi, P. I. W. de Bakker, M. Seielstad, W. Liao, M. Stähle, A. Franke, X. Zhang, and J. Liu, “Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility,” *Nature Communications*, vol. 6, p. 6916, 2015.
- [110] S.-S. Khor, W. Yang, M. Kawashima, S. Kamitsuji, X. Zheng, N. Nishida, H. Sawai, H. Toyoda, T. Miyagawa, M. Honda, N. Kamatani, and K. Tokunaga, “High-accuracy imputation for HLA class I and II genes based on high-resolution SNP data of population-specific references,” *The Pharmacogenomics Journal*, vol. 15, pp. 530–537, Dec. 2015.
- [111] R. Panzer, C. Blobel, R. Fölster-Holst, and E. Proksch, “TLR2 and TLR4 expression in atopic dermatitis, contact dermatitis and psoriasis,” *Experimental Dermatology*, vol. 23, pp. 364–366, May 2014.

- [112] Y. Fan and B. Liu, “Expression of Toll-like receptors in the mucosa of patients with ulcerative colitis,” *Experimental and Therapeutic Medicine*, vol. 9, pp. 1455–1459, Apr. 2015.
- [113] W.-D. Xu, S.-S. Liu, H.-F. Pan, and D.-Q. Ye, “Lack of association of TLR4 polymorphisms with susceptibility to rheumatoid arthritis and ankylosing spondylitis: a meta-analysis,” *Joint, Bone, Spine: Revue Du Rhumatisme*, vol. 79, pp. 566–569, Dec. 2012.
- [114] H. Matsushita, Y. Miyake, A. Takaki, T. Yasunaka, K. Koike, F. Ikeda, H. Shiraha, K. Nouse, and K. Yamamoto, “TLR4, TLR9, and NLRP3 in biliary epithelial cells of primary sclerosing cholangitis: relationship with clinical characteristics,” *Journal of Gastroenterology and Hepatology*, vol. 30, pp. 600–608, Mar. 2015.
- [115] A. Kaser, S. Zeissig, and R. S. Blumberg, “Inflammatory bowel disease,” *Annual Review of Immunology*, vol. 28, pp. 573–621, 2010.
- [116] T. Slater, “Recent advances in modeling languages for pathway maps and computable biological networks,” *Drug Discovery Today*, vol. 19, pp. 193–198, Feb. 2014.
- [117] M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro, “HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores,” *PLOS ONE*, vol. 7, p. e31826, Feb. 2012.
- [118] C. Parham, M. Chirica, J. Timans, E. Vaisberg, M. Travis, J. Cheung, S. Pflanz, R. Zhang, K. P. Singh, F. Vega, W. To, J. Wagner, A.-M. O’Farrell, T. McClanahan, S. Zurawski, C. Hannum, D. Gorman, D. M. Rennick, R. A. Kastelein, R. de Waal Malefyt, and K. W. Moore, “A receptor for the heterodimeric cytokine IL-23 is composed of IL-12rbeta1 and a novel cytokine receptor subunit, IL-23r,” *Journal of Immunology (Baltimore, Md.: 1950)*, vol. 168, pp. 5699–5708, June 2002.
- [119] A. Nueda, M. López-Cabrera, A. Vara, and A. L. Corbí, “Characterization of the CD11a (alpha L, LFA-1 alpha) integrin gene promoter,” *The Journal of Biological Chemistry*, vol. 268, pp. 19305–19311, Sept. 1993.
- [120] C. M. Bailey, D. E. Abbott, N. V. Margaryan, Z. Khalkhali-Ellis, and M. J. C. Hendrix, “Interferon regulatory factor 6 promotes cell cycle arrest and is regulated by the proteasome in a cell cycle-dependent manner,” *Molecular and Cellular*

- Biology*, vol. 28, pp. 2235–2243, Apr. 2008.
- [121] T. Ben-Zvi, A. Yayon, A. Gertler, and E. Monsonogo-Ornan, “Suppressors of cytokine signaling (SOCS) 1 and SOCS3 interact with and modulate fibroblast growth factor receptor signaling,” *Journal of Cell Science*, vol. 119, pp. 380–387, Jan. 2006.
- [122] C. D. Touchberry, T. M. Green, V. Tchikrizov, J. E. Mannix, T. F. Mao, B. W. Carney, M. Girgis, R. J. Vincent, L. A. Wetmore, B. Dawn, L. F. Bonewald, J. R. Stubbs, and M. J. Wacker, “FGF23 is a novel regulator of intracellular calcium and cardiac contractility in addition to cardiac hypertrophy,” *American Journal of Physiology. Endocrinology and Metabolism*, vol. 304, pp. E863–873, Apr. 2013.
- [123] R. Del Pinto, D. Pietropaoli, A. K. Chandar, C. Ferri, and F. Cominelli, “Association Between Inflammatory Bowel Disease and Vitamin D Deficiency: A Systematic Review and Meta-analysis,” *Inflammatory Bowel Diseases*, vol. 21, pp. 2708–2717, Nov. 2015.
- [124] D. Lamparter, D. Marbach, R. Rueedi, Z. Kutalik, and S. Bergmann, “Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics,” *PLoS computational biology*, vol. 12, p. e1004714, Jan. 2016.
- [125] G. Ragnedda, G. Disanto, G. Giovannoni, G. C. Ebers, S. Sotgiu, and S. V. Ramagopalan, “Protein-protein interaction analysis highlights additional loci of interest for multiple sclerosis,” *PloS One*, vol. 7, no. 10, p. e46730, 2012.
- [126] T. Raj, J. M. Shulman, B. T. Keenan, L. B. Chibnik, D. A. Evans, D. A. Bennett, B. E. Stranger, and P. L. De Jager, “Alzheimer disease susceptibility loci: evidence for a protein network under natural selection,” *American Journal of Human Genetics*, vol. 90, pp. 720–726, Apr. 2012.
- [127] L. García-Alonso, R. Alonso, E. Vidal, A. Amadoz, A. d. María, P. Minguez, I. Medina, and J. Dopazo, “Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments,” *Nucleic Acids Research*, vol. 40, pp. e158–e158, Nov. 2012.
- [128] T. Moriguchi, N. Kuroyanagi, K. Yamaguchi, Y. Gotoh, K. Irie, T. Kano, K. Shirakabe, Y. Muro, H. Shibuya, K. Matsumoto, E. Nishida, and M. Hagiwara, “A novel kinase cascade mediated by mitogen-activated protein kinase kinase 6 and MKK3,” *The Journal of Biological Chemistry*, vol. 271, pp. 13675–13679, June

- 1996.
- [129] T. Kamada and S. Kawai, “An algorithm for drawing general undirected graphs,” *Information Processing Letters*, vol. 31, pp. 7–15, Apr. 1989.
- [130] W. Liao, J.-X. Lin, and W. J. Leonard, “IL-2 Family Cytokines: New Insights into the Complex Roles of IL-2 as a Broad Regulator of T helper Cell Differentiation,” *Current Opinion in Immunology*, vol. 23, pp. 598–604, Oct. 2011.
- [131] C. J. Carroll, P. Isohanni, R. Pöyhönen, L. Euro, U. Richter, V. Brilhante, A. Götz, T. Lahtinen, A. Paetau, H. Pihko, B. J. Battersby, H. Tyynismaa, and A. Suomalainen, “Whole-exome sequencing identifies a mutation in the mitochondrial ribosome protein MRPL44 to underlie mitochondrial infantile cardiomyopathy,” *Journal of Medical Genetics*, vol. 50, pp. 151–159, Mar. 2013.
- [132] E. C. Koc, W. Burkhart, K. Blackburn, A. Moseley, H. Koc, and L. L. Spremulli, “A proteomics approach to the identification of mammalian mitochondrial small subunit ribosomal proteins,” *The Journal of Biological Chemistry*, vol. 275, pp. 32585–32591, Oct. 2000.
- [133] T. Nunes, C. Bernardazzi, and H. S. de Souza, “Cell Death and Inflammatory Bowel Diseases: Apoptosis, Necrosis, and Autophagy in the Intestinal Epithelium,” *BioMed research international*, vol. 2014, 2014.
- [134] Y. Kawamata, R. Fujii, M. Hosoya, M. Harada, H. Yoshida, M. Miwa, S. Fukusumi, Y. Habata, T. Itoh, Y. Shintani, S. Hinuma, Y. Fujisawa, and M. Fujino, “A G protein-coupled receptor responsive to bile acids,” *The Journal of Biological Chemistry*, vol. 278, pp. 9435–9440, Mar. 2003.
- [135] S. E. Baranzini, N. W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, W. Wu, B. M. J. Uitdehaag, L. Kappos, G. Consortium, C. H. Polman, P. M. Matthews, S. L. Hauser, R. A. Gibson, J. R. Oksenberg, and M. R. Barnes, “Pathway and network-based analysis of genome-wide association studies in multiple sclerosis,” *Human Molecular Genetics*, vol. 18, pp. 2078–2090, June 2009.
- [136] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, 3rd Edition*. Cambridge, Mass: The MIT Press, 3rd edition ed., July 2009.
- [137] S. Krishnamurthi and J. Vitek, “The Real Software Crisis: Repeatability As a Core Value,” *Commun. ACM*, vol. 58, pp. 34–36, Feb. 2015.

- [138] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao, “dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks,” *Bioinformatics*, vol. 27, pp. 95–102, Jan. 2011.
- [139] G. Csardi and T. Nepusz, “The igraph software package for complex network research,” *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.
- [140] P. Di Tommaso, E. Palumbo, M. Chatzou, P. Prieto, M. L. Heuer, and C. Notredame, “The impact of Docker containers on the performance of genomic pipelines,” *PeerJ*, vol. 3, p. e1273, Sept. 2015.
- [141] N. Garzorz, L. Krause, F. Lauffer, A. Atenhan, J. Thomas, S. P. Stark, R. Franz, S. Weidinger, A. Balato, N. S. Mueller, F. J. Theis, J. Ring, C. B. Schmidt-Weber, T. Biedermann, S. Eyerich, and K. Eyerich, “A novel molecular disease classifier for psoriasis and eczema,” *Experimental Dermatology*, May 2016.
- [142] P. Brest, P. Lapaquette, M. Souidi, K. Lebrigand, A. Cesaro, V. Vouret-Craviari, B. Mari, P. Barbry, J.-F. Mosnier, X. Hébuterne, A. Harel-Bellan, B. Mograbi, A. Darfeuille-Michaud, and P. Hofman, “A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn’s disease,” *Nature Genetics*, vol. 43, pp. 242–245, Mar. 2011.
- [143] D. K. Podolsky, “The current future understanding of inflammatory bowel disease,” *Best Practice & Research. Clinical Gastroenterology*, vol. 16, pp. 933–943, Dec. 2002.
- [144] D. C. Savage, “Microbial Ecology of the Gastrointestinal Tract,” *Annual Review of Microbiology*, vol. 31, no. 1, pp. 107–133, 1977.
- [145] R. Sender, S. Fuchs, and R. Milo, “Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans,” *Cell*, vol. 164, pp. 337–340, Jan. 2016.
- [146] W. A. A. de Steenhuijsen Piters and D. Bogaert, “Unraveling the Molecular Mechanisms Underlying the Nasopharyngeal Bacterial Community Structure,” *mBio*, vol. 7, no. 1, pp. e00009–00016, 2016.
- [147] T. Kanai, Y. Mikami, and A. Hayashi, “A breakthrough in probiotics: *Clostridium butyricum* regulates gut homeostasis and anti-inflammatory response in inflammatory bowel disease,” *Journal of Gastroenterology*, pp. 1–12, May 2015.

-
- [148] B. Zhang, B. Chassaing, Z. Shi, R. Uchiyama, Z. Zhang, T. L. Denning, S. E. Crawford, A. J. Pruijssers, J. A. Iskarpatyoti, M. K. Estes, T. S. Dermody, W. Ouyang, I. R. Williams, M. Vijay-Kumar, and A. T. Gewirtz, “Prevention and cure of rotavirus infection via TLR5/NLRC4-mediated production of IL-22 and IL-18,” *Science*, vol. 346, pp. 861–865, Nov. 2014.
- [149] L. Saidel-Odes and S. Odes, “Hygiene hypothesis in inflammatory bowel disease,” *Annals of Gastroenterology : Quarterly Publication of the Hellenic Society of Gastroenterology*, vol. 27, no. 3, pp. 189–190, 2014.