IDENTIFICATION OF GENES
AND REGULATORS THAT ARE SHARED
ACROSS T CELL ASSOCIATED DISEASES

Danuta R. Gawel
Any knowledge that doesn’t lead to new questions quickly dies out: it fails to maintain the temperature required for sustaining life.

—Wisława Szymborska—
SUPERVISOR
Professor Mikael Benson
The Center for Personalised Medicine
Department of Clinical and Experimental Medicine
Linköping University, Sweden

Co-SUPERVISORS
Docent Mika Gustafsson
Department of Physics, Chemistry and Biology
Linköping University, Sweden
PhD Colm Nestor
Department of Clinical and Experimental Medicine
Linköping University, Sweden

OPPONENT
Docent Erik Larsson Lekholm
Department of Medical Biochemistry and Cell biology at Institute of Biomedicine
University of Gothenburg, Sweden
Table of Contents

0 Acknowledgements 7

1 Identifiering av gener och regulatorer som delas av flera olika T cells associerade sjukdomar 9

2 Abstract 11

3 List of original papers 13

4 Abbreviations 15

5 Introduction 19

6 Materials and Methods 31

6.1 Study subjects .......................... 31
6.2 Ethics statement .......................... 32
6.3 Definition of low and high responders to treatment ........... 33
6.4 GWAS data accession .......................... 33
6.5 Enrichment analyses .......................... 34
6.6 Public T-cell disease data .......................... 34
6.7 Gene expression microarray data processing .......................... 35
6.8 Analyses of sets of genes .......................... 36
6.9 Disease-associated network modules .......................... 36
6.10 Shared disease module validation in several human interactome databases .......................... 37
6.11 Classifications .......................... 37
6.12 Th1/Th2 regulatory region analysis .......................... 38
6.13 Construction of the Th1/Th2 GRN .......................... 38
6.14 TF-target relationship correlation analyses .......................... 39
6.15 ChIP-Seq analyses .......................... 40
6.16 Gene Regulatory Network and TFBS comparison using ChIP-Seq peak counts .......................... 40
6.17 Splice associated SNPs .......................... 40
6.18 Expression quantitative trait loci .......................... 41
6.19 siRNA mediated gene knock down .......................... 41
6.20 Sublingual immunotherapy (SLIT) .......................... 41
6.21 Quantitative PCR (qPCR) .......................... 41
6.22 MS gene expression in relapse and remission .......................... 42
6.23 GWAS analysis of multiple sclerosis .......................... 42
6.24 Construction of miR-mRNA regulatory network and module identification .......................... 42
7 Results and Discussion

7.1 Project I: Identification of shared disease genes ............ 45
7.2 Project II: Identification of the upstream regulators ....... 53
7.3 Project III: microRNAs in T helper cell differentiation and in T cell related diseases ...................... 65

8 Concluding Remarks ........................................ 73

9 Paper I ............................................................ 97

10 Paper II .......................................................... 111

11 Manuscript ....................................................... 123
Acknowledgements

“We are not lost. We’re locationally challenged.”
— John M. Ford —

In science there is no easy path that would lead straight to answers. In research, as in life, it is vital to be surrounded by great people sharing one goal. Here I would like to express my gratitude to everyone that joined me on my PhD journey.

To Mikael Benson, my supervisor, exceptional leader, who has welcomed me as PhD student in his group. His guidance and support broaden my horizons not only regarding scientific aspects. Thank you for the encouraging words that I needed so much, especially when I was becoming lost and confused (but always ‘at the higher level’).

To Mika Gustafsson, my co-supervisor, outstanding teacher, to whom I owe incredible development in the key field for me: bioinformatics. Thank you for recurring remainders of a big picture when I was getting too focused on details.

To Colm Nestor, my co-supervisor, for great assistance and always constructive, very precise and scientific critique, which never put me down but instead encouraged to search answers in a better way - which in my opinion is a very unique skill.

To my present and past co-workers for hours, and hours of great inspiring scientific discussions and lightening chats over the coffee. Special thanks to Huan Zhang, Jordi Serra Musach, Antonio Lentini, Cathrine Lagerwall, Sandra Lilja, Lina Mattson, and Robert Liljenström.

To Mattias Köpsén and Zainab Chaudry for putting up with me as your co-supervisor. It has been a great pleasure to work with you. You have been my teachers too.

To everyone at Cell Biology floor 13 for worm welcome, all the coffee and lunch breaks, and other activities, especially to Annelie Lindström, Björn Ingelsson, Birgitta Holmlund, and Ivana Tichá.

Last but not least I would like to thank my great family for never-ending, unquestionable support in every step I take. To my dearest husband thanks to whom even the darkest night isn’t scary. To my beloved son who with a
single smile brings things back to the right perspective. To my parents and sisters who always encourage me and stand by my side ‘no matter what’.

Here I would like to also thank one more person that have inspired me to start this journey, great teacher – Krzysztof Fujarewicz.

Thank you!
Identifiering av gener och regulatorer som delas av flera olika T cells associerade sjukdomar

Genomvida associations studier (GVAS) gör det möjligt att studera miljontals små variationer i människans gener. Sådana studier har utförts på hundratals olika sjukdomar och miljontals patienter. Ett viktigt fynd är att samma gener kan vara associerade med olika sjukdomar, s.k. delade gener. Målen för min avhandling var att identifiera a) delade gener, b) vilka gener som hade uppströms, reglerande funktioner, samt c) hur samma gener kan vara associerade med olika sjukdomar.

Jag studerade en viss sorts vit blodkropp, s.k. CD4+ T celler, eftersom analys av alla GVAS gener visade att en signalväg som är uttryckt i denna celltyp var statistiskt mest signifikant. Jag studerade därför genaktiviteten, i form av mRNA uttryck, i denna celltyp vid flera olika sjukdomar. Detta gjordes med mikromatristeknik så att mRNA nivåerna för flertalet av människans gener kunde mätas. Genom att jämföra med mRNA nivåerna hos friska kunde tusentals sjukdomsassocierade gener och deras mRNA uttryck identifieras (s.k. transkript). För att få en överblick över dessa mappade jag, med datorns hjälp, transkripten på en nätverkskarta över människliga gener och deras interaktioner. Jag fann att många delade transkript fanns på samma ställe på kartan och därför var mycket hopkopplade. Bioinformatiska analyser visade att dessa transkript ingick i livsviktiga signalvägar som reglerade inflammation, ämnesomsättning och celltillväxt. Detta ledde till hypotesen att sjukdomsassocierade förändringar i någon av de delade transkripten skulle kunna spridas till mer än en signalväg och därigenom orsaka en eller flera sjukdomar. Hypotesen stöddes av att de gener som kodade för de delade transkripten var kraftigt berikade för genetiska varianter som identifierats av GVAS. Den andra frågan i min avhandling, vilka gener som hade reglerande funktioner, studerade jag genom att konstruera en nätverksmodell som över tid beskrev hur transkript aktiverades i CD4+ T celler. Jag fann då att tre s.k. transkriptions faktorer hade en övergripande roll vid många sjukdomar, och validerade detta genom studier av symptomfria stadier hos patienter med pollensnuva och multipel skleros. I en tredje studie analyserade jag en an-
nan sorts regelmolekyler, s.k. microRNA, men fann att de huvudsakligen var inblandade i senare stadier av olika sjukdomar. Avseende den tredje frågan i avhandlingen, hur samma gener kan vara inblandade i olika sjukdomar fann jag två förklaringar: 1) olika nivåer av samma transkript, samt 2) olika funktion av samma gen p.g.a. genetiska varianter.

Sammanfattningsvis fann jag att delade gener kan vara viktiga för uppkomsten av flera olika sjukdomar, och att detta kan bero på olika mRNA uttryck eller funktion hos samma gen.
Abstract

Genome-wide association studies (GWASs) of hundreds of diseases and millions of patients have led to the identification of genes that are associated with more than one disease. The aims of this PhD thesis were to a) identify a group of genes important in multiple diseases (shared disease genes), b) identify shared up-stream disease regulators, and c) determine how the same genes can be involved in the pathogenesis of different diseases. These aims have been tested on CD4+ T cells because they express the T helper cell differentiation pathway, which was the most enriched pathway in analyses of all disease associated genes identified with GWASs.

Combining information about known gene-gene interactions from the protein-protein interaction (PPI) network with gene expression changes in multiple T cell associated diseases led to the identification of a group of highly interconnected genes that were miss-expressed in many of those diseases – hereafter called ‘shared disease genes’. Those genes were further enriched for inflammatory, metabolic and proliferative pathways, genetic variants identified by all GWASs, as well as mutations in cancer studies and known diagnostic and therapeutic targets. Taken together, these findings supported the relevance of the shared disease genes.

Identification of the shared upstream disease regulators was addressed in the second project of this PhD thesis. The underlying hypothesis assumed that the determination of the shared upstream disease regulators is possible through a network model showing in which order genes activate each other. For that reason a transcription factor–gene regulatory network (TF-GRN) was created. The TF-GRN was based on the time-series gene expression profiling of the T helper cell type 1 (Th1), and T helper cell type 2 (Th2) differentiation from Naïve T-cells. Transcription factors (TFs) whose expression changed early during polarization and had many downstream predicted targets (hubs) that were enriched for disease associated single nucleotide polymorphisms (SNPs) were prioritised as the putative early disease regulators. These analyses identified three transcription factors: GATA3, MAF and MYB. Their predicted targets were validated by ChIP-Seq and siRNA mediated knockdown in primary human T-cells. CD4+ T cells isolated from seasonal allergic rhinitis (SAR) and multiple
sclerosis (MS) patients in their non-symptomatic stages were analysed in order to demonstrate predictive potential of those three TFs. We found that those three TFs were differentially expressed in symptom-free stages of the two diseases, while their TF-GRN–predicted targets were differentially expressed during symptomatic disease stages. Moreover, using RNA-Seq data we identified a disease associated SNP that correlated with differential splicing of GATA3.

A limitation of the above study is that it concentrated on TFs as main regulators in cells, excluding other potential regulators such as microRNAs. To this end, a microRNA–gene regulatory network (mGRN) of human CD4+ T cell differentiation was constructed. Within this network, we defined regulatory clusters (groups of microRNAs that are regulating groups of mRNAs). One regulatory cluster was differentially expressed in all of the tested diseases, and was highly enriched for GWAS SNPs. Although the microRNA processing machinery was dynamically upregulated during early T-cell activation, the majority of microRNA modules showed specialisation in later time-points.

In summary this PhD thesis shows the relevance of shared genes and upstream disease regulators. Putative mechanisms of why shared genes can be involved in pathogenesis of different diseases have also been demonstrated: a) differential gene expression in different diseases; b) alternative transcription factor splicing variants may affect different downstream gene target group; and c) SNPs might cause alternative splicing.
List of original papers

This thesis is based on the following original papers:

PROJECT I

PROJECT II

PROJECT III

†These authors contributed equally to this work and should be regarded as shared first or last authors, respectively.
## Abbreviations

_In the alphabetical order:_

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGO2</td>
<td>Argonaute 2, RISC catalytic component</td>
</tr>
<tr>
<td>ALL</td>
<td>Acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>AML</td>
<td>Acute myeloid leukemia</td>
</tr>
<tr>
<td>ATL</td>
<td>Adult T cell leukaemia/lymphoma</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under ROC curve</td>
</tr>
<tr>
<td>CEP290</td>
<td>Centrosomal protein 290</td>
</tr>
<tr>
<td>CFTR</td>
<td>Cystic fibrosis transmembrane conductance</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>Chromatin immunoprecipitation-sequencing</td>
</tr>
<tr>
<td>CLL</td>
<td>Chronic lymphoid leukaemia</td>
</tr>
<tr>
<td>DE</td>
<td>Differentially expressed</td>
</tr>
<tr>
<td>DGCR8</td>
<td>DGCR8, microprocessor complex subunit</td>
</tr>
<tr>
<td>DHS</td>
<td>DNase I hypersensitive sites</td>
</tr>
<tr>
<td>DICER</td>
<td>Dicer 1, ribonuclease III</td>
</tr>
<tr>
<td>DMD</td>
<td>Dystrophin</td>
</tr>
<tr>
<td>DROSHA</td>
<td>Drosha ribonuclease III</td>
</tr>
<tr>
<td>eQTL</td>
<td>Quantitative trait loci</td>
</tr>
<tr>
<td>ex-miRs</td>
<td>Extracellular miRs</td>
</tr>
<tr>
<td>FE</td>
<td>Fold enrichment</td>
</tr>
<tr>
<td>FET</td>
<td>Fisher exact test</td>
</tr>
<tr>
<td>FN1</td>
<td>Fibronectin 1</td>
</tr>
<tr>
<td>GATA3</td>
<td>GATA binding protein 3</td>
</tr>
<tr>
<td>GCs</td>
<td>Glucocorticoids</td>
</tr>
<tr>
<td>GN</td>
<td>Gene network</td>
</tr>
<tr>
<td>GO</td>
<td>Gene ontology</td>
</tr>
<tr>
<td>GRN</td>
<td>Gene regulatory network</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome wide association study</td>
</tr>
<tr>
<td>HES</td>
<td>Hypereosinophilic syndrome</td>
</tr>
<tr>
<td>HIV</td>
<td>Human immunodeficiency virus infection</td>
</tr>
<tr>
<td>HRs and LRs</td>
<td>High- and low-responders respectively</td>
</tr>
<tr>
<td>IKZF1</td>
<td>IKAROS family zinc finger 1</td>
</tr>
</tbody>
</table>
IL23R  Interleukin 23 receptor  
IPA  Ingenuity pathway analysis  
KEGG  Kyoto encyclopedia of genes and genomes  
LASSO  Least absolute shrinkage and selection operator  
LCC  Largest connected component  
LD  Linkage disequilibrium  
LEF1  Lymphoid enhancer binding factor 1  
lncRNAs  Long non-coding RNAs  
MAF  MAF bZIP transcription factor  
mGRN  MicroRNA-Gene Regulatory Network  
mIR  MicroRNA  
MS  Multiple sclerosis  
MYB  MYB proto-oncogene, transcription factor  
NOTCH4  Notch 4  
NT  Naïve T cell  
NZB  Humanized monoclonal antibody which binds to α4β1- integrin  
OMIM  Online mendelian inheritance in man  
OR  Odds ratio  
PAH  Phenylalanine hydroxylase  
pAID  Paediatric-age-of-onset autoimmune diseases  
PBMCs  Peripheral blood mononuclear cells  
PCC  Pearson’s correlation coefficient  
PPI  Protein-protein interactions  
PTPN22  Protein tyrosine phosphatase, non-receptor type 22  
RA  Rheumatoid arthritis  
RISC  RNA-induced silencing complex  
rSNP  Regulatory SNP  
SAR  Seasonal allergic rhinitis  
SeS  Sézary syndrome  
siRNA  Small interfering RNA  
SLE  Systemic lupus erythematosus  
SLIT  Sublingual immunotherapy  
SMN2  Survival of motor neuron 2, centromeric  
SNP  Single nucleotide polymorphism  
SRS  Splice regulatory sites  
TARBP2  TARBP2, RISC loading complex RNA binding subunit  
TERT  Telomerase reverse transcriptase
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>TFBS</td>
<td>Transcription factor binding sites predictions</td>
</tr>
<tr>
<td>TF-GRN</td>
<td>TF-Gene regulatory network</td>
</tr>
<tr>
<td>Th1</td>
<td>T helper cell type 1</td>
</tr>
<tr>
<td>Th2</td>
<td>T helper cell type 2</td>
</tr>
<tr>
<td>Th17</td>
<td>T helper cell type 17</td>
</tr>
<tr>
<td>Treg</td>
<td>T regulatory cell (Treg)</td>
</tr>
<tr>
<td>T1D</td>
<td>Type 1 diabetes</td>
</tr>
<tr>
<td>XPO5</td>
<td>Exportin 5</td>
</tr>
<tr>
<td>3'(5')UTR</td>
<td>3' (5') untranslated region</td>
</tr>
</tbody>
</table>
Introduction

Molecular cell biology is a very complex field, with many unknowns and uncertainties. For over a century many researches has been following the rule of ‘divide and conquer’ in order to gain insights into underlying mechanisms of biological processes. Following that rule much of twentieth-century has been an attempt to reduce biological phenotypes to individual cellular components like genes and proteins [1,2,3]. Following the assumption that a trait spread over families is synonymous with inheritance of an error in a single molecule, many monogenic disorders were found, including phenylketonuria (PAH) [4], cystic fibrosis (CFTR) [4], and many others [4,5]. Despite the huge success of reductionism in identification of monogenic genetic disorders, it has been later recognised that in general the link between genotype and phenotype is far too complex to be explained by a change in single gene [3]. This effect is especially noticeable in complex diseases which are an effect of multiple genetic factors, together with environmental and lifestyle components [4,6]. This gave rise to large-scale association studies which were believed to have a greater power in identification of disease genes even if those would have more modest effects [6]. Genome Wide Association Studies (GWASs) aim to identify genetic variants associated with a given trait in different individuals. Typically, GWASs focus on associations between single-nucleotide polymorphisms (SNPs) and human diseases but also other clinical conditions, like height or longevity. This genome-wide approach in human genetics gives an opportunity to identify genetic factors that contribute to human diseases.

Remarkably, among hundreds of studied diseases and millions of patients, GWASs identified many SNPs to be associated with more than one disease or trait. It has been reported that 4.6% of GWAS SNPs show pleiotropic effects [7]. In fact, more than half of the autoimmune GWAS SNPs are associated with at least two distinct autoimmune diseases [8,9]. Furthermore, linking GWAS variants to genes has given an insight into genes associated with more than one disease. A widely known example is PTPN22, which is associated with type 1 diabetes, rheumatoid arthritis and Crohn’s disease [10]. In ten paediatric-age-of-onset autoimmune diseases (pAIDs) [11] a genome wide study identified 27 loci significantly associated
with at least one pAID. Interestingly, they found 8 risk alleles shared by combined pAIDs, but that were also associated with protection against one of the pAIDs. Those findings led to the main question in this thesis: *How can the same genes contribute to pathogenesis of different diseases?*

Previous work has suggested some answers to this question. In a systematic analysis of 14 tumour types in [12] the authors searched for somatic mutations in regulatory regions that affect mRNA expression. One exceptional example was mutations in the promoter region of *TERT*, which were strongly associated with expression increase. However, those mutations and expression changes varied greatly between different tumours. This example shows how different changes of the same gene or its regulation can be associated with different disease manifestation – here expression change is putatively caused by a somatic mutation and/or copy number alterations and/or by somatic structural variants as shown in another study of 600 tumour and 18 cancer types [13]. However cancer is a very specific group of diseases, and therefore this example shows how one gene’s expression might vary in different cancers and what might cause those differences, but it doesn’t answer how the same gene might contribute to pathogenesis of different diseases.

In another study meta-analysis of expression profiling data from several different cells and tissues from inflammatory, malignant and metabolic diseases showed that transcripts that were shared by more than one disease tended to co-localise when mapped on the human protein-protein interaction (PPI) network [14]. By inference, they were more interconnected than expected by chance. This is at least partially consistent with previous research showing that genes associated with phenotypically similar diseases tend to be interconnected and also functionally related [15]. However, the diseases in the expression profiling meta-analysis were highly diverse. How can the same, highly interconnected and presumably functionally related transcripts be associated with different diseases? To answer this question, the transcripts were analysed for pathway enrichment. In other words, did the transcripts belong to any pathways? Indeed, the transcripts were highly enriched for inflammatory, metabolic and proliferative pathways. This suggested a hypothetical explanation for why the same genes could be associated with more than one disease: Those pathways have key roles, and because of their interconnectivity, malfunction in one could lead to malfunction in the others and thereby increase the risk for one or more dis-
eases. This hypothetical explanation was supported by meta-analysis of independent GWASs data of 145 other complex diseases: There was highly significant enrichment of genetic variants identified by those GWASs [14].

Later, it was shown that interplay between three endophenotypes (inflammation, fibrosis and thrombosis) and local organ-based environment may lead to disease-specific gene expression [16]. They identified that disease-genes associated with 52% of 299 studied diseases significantly overlapped with at least one of the 3 endophenotypes, whereas over 22% were enriched in all three endophenotypes. Taken together, these findings support that inflammation, fibrosis and thrombosis are common to many complex diseases, and that differential expression in different organ systems can explain why the same genes can contribute to pathogenesis of different diseases.

These studies also highlight an important concept underlying this thesis. The background to this concept is that in many studies pleiotropic SNPs/genes are analysed separately as independent incidents. This could be a limitation since many relationships between genotype and phenotype are difficult to explain by single molecules [3, 17]. In fact, many cellular processes are a result of multiple coordinated, often very complex, mechanisms. Therefore scaling-up from individual molecules to groups of functionally and/or psychically coordinated molecules is a potentially advantageous concept [1].

One way to identify such coordinated molecules is to identify gene modules. This concept requires introducing of a few definitions. First of all, a gene network, which can be seen as a graphical representation of gene to gene interactions (Fig.1). Nodes (circles) represent genes or proteins, and lines (edges) represent interactions. In such a representation highly connected genes (nodes connected with many edges) would form a module (Fig.1).

Here again both nodes and the edges (genes and gene interactions) have to be contemplated in the context of specific cell type and tissue [18,19,20], as they will differ in different cells depending on the cell function. In line with the ‘guilty-by-association’ or ‘local’ hypothesis, if a particular gene is involved in a process or disease, its direct interactors (neighbours) are also suspected to have some role in the same process [17]. Also in line with that hypothesis, the cellular components that are associated with a specific disease phenotype tend to cluster together in the same network neighbour-
Figure 1: Gene network (GN) – each node (disks and stars) represents a gene or protein, whereas each line represents interaction between genes/proteins. Black dashed circle marks a gene/protein module – a highly connected set of nodes. Stars represent nodes with many neighbours – hubs.

hood [14, 17, 21, 22]. Consequently, a group of tightly connected genes that contribute to a disease phenotype hereafter will be referred to as ‘disease module’. Disease modules tend to at least partially overlap with functional modules (modules of genes coordinated in the same function like a biological pathway) related to the disease in question. For example, inflammatory disease modules are likely to show similarities with a functional module enriched for inflammatory pathways [17]. By inference, functionally related diseases tend to have similar disease modules [23]. However, modules from different diseases may (as we show in this thesis) overlap. In [24], the authors state that a phenotype often correlates with the inability of a particular functional module to perform its tasks. Thus, different combinations of genes in the same module could result in the same disease phenotype. For example, Zellweger syndrome is caused by mutations in any of at least 11 genes, where all of those genes are associated with the same process – peroxisome biogenesis [24].

Modules in general are a widely observed network feature in biological networks [2, 25, 26]. For example modules have been reported in gene network of *Saccharomyces cerevisiae* [26], apoptosis network [27], PPI networks [2], metabolic networks [28], and many others. In order to identify modules one can choose between numerous methods [29, 30]. All of those
methods are based on a prior knowledge, which in biological contexts is often incomplete and/or noisy. This prior knowledge might be, for example, protein-protein interaction network (were both nodes and edges are not fully known [31, 32]), or experimentally measured gene expression data, or combination of those two. Gene expression measured with microarrays rise doubts about experiment reproducibility and comparability [33,34,35]. On the other hand, RNA-Seq is an alternative to microarrays. This technique was shown to be superior to microarrays in terms of detection of low abundance transcripts, and is not limited by the prior design (which genes and their fragments should be targeted in the experiment) [35]. Even though technical variability has been shown to be lower in RNA-Seq compared to microarray studies [36], it is still present. Also other sources of bias in the data cannot be fully eliminated. Furthermore key problem for all models is shortage of data – the number of measured genes is usually much larger than the number of measurements [37]. Therefore individual gene-gene interactions in the network can be doubtful, and there is a need for robust concepts that can be inferred despite data limitations. Wide abundance of modules and hubs (genes with many neighbours, Fig.1, discussed below) across various biological contexts might denote that those concepts are stable network features.

However, the question in this thesis is how the same genes can be involved in the pathogenesis of different diseases. As stated above, one explanation is that same biological process (like pathway) might be affected by different gene malfunctions and therefore can lead to one or more diseases. Those malfunctions could be quantitative or qualitative. Examples of the former include differential mRNA or protein expression levels, which can be due to both genetic and epigenetic causes. Examples of the latter include genetic variants affecting protein composition, including differential splicing, which could affect protein-protein interactions [38].

Because of this complexity, disease-associated genes can be defined in multiple ways. One definition is genes associated with at least one disease-associated genetic variant. However, differentially expressed genes at the protein or mRNA level, may also be referred to as disease-associated. Since mRNA changes may reflect both genetic and epigenetic disease mechanisms, this thesis is focused on first identifying shared disease genes based on mRNA profiling, and secondly, mechanisms that may explain pleiotropy.

Although all cells in an individual have the same DNA, there are hun-
dreds of different cell types. This heterogeneity is achieved by epigenetic variations. Furthermore expression quantitative trait loci (eQTL) – genomic positions that affect gene expression, have been shown to be cell- or tissue specific [39]. Moreover GWAS perturbations of functional groups of genes (modules) often point to known disease-relevant cell types and tissues [40]. Further, disease genes often exhibit tissue-specific expression patterns [18, 21, 41]. Therefore, in order to identify and functionally study shared disease-associated genes it would be ideal to obtain patient samples from a tissue associated with many diseases. In this thesis project, we searched for a cell type that was both associated with many diseases and accessible for clinical studies. In order to identify such a cell type, the first project in my thesis was based on pathway enrichment analyses of all published GWAS genes. This showed that T cell differentiation was the most enriched pathway. This pathway is expressed in CD4+ T cells, which is the focus of my thesis.

This focus is supported by several known factors: T cells constantly patrol all parts of the body in order to detect, heal or combat disease processes. Therefore, any alterations in its regulation may affect other systems. For example although autoimmune diseases often affect specific organs, they might also have more systematic manifestations [42]. Examples include coeliac disease which primarily damages intestinal mucosa, but may also have neurological, psychiatric, or thyroid manifestations [42]. Moreover in [42] authors claim that most if not all of dysregulated processes involved in autoimmune diseases will in fact at some point result in T-cell activation. Indeed, T cells are, either primarily or secondarily, involved in allergic, autoimmune, infectious, and malignant diseases [43]. This further supports that the immune system might be a common denominator for many diseases. As an example, many shared autoimmune and inflammatory disease associated variants are specifically enriched in the active enhancer elements in stimulated T-cell subpopulations [44, 45, 46]. Another advantage of focusing on T cells is that their activation and differentiation are well defined and can be induced in vitro in primary cells from healthy blood donors. This allows functional studies of the effects of shared genetic or epigenetic variants on T cell activation or differentiation (Fig.2).

Taken together these findings support the relevance of T cells to study shared disease genes. In the first project of my thesis we have identified disease modules in expression profiling of published data from T cells from
Figure 2: Conceptual model of how a naïve T cell can differentiate into different subsets, T regulatory (Treg), and T helper (Th) cells of types 17, 2 and 1.

eight inflammatory and malignant diseases. We found that those modules in fact do overlap. These overlapping genes were also highly interconnected and formed a new module, hereafter referred to as a shared disease module [47]. We also found that the shared disease module was enriched for key cellular pathways regulating metabolism, proliferation and inflammation. In support of its disease relevance the shared module was highly enriched for GWAS genes, primary drug targets and diagnostic markers. Variable enrichment of GWAS genes pointed to an explanation for why the same genes can be associated with different diseases. Also, different SNPs mapped to the same gene might have presumably different effects and hereby contribute to different disease phenotypes. For example SNP rs11209026 has been associated with inflammatory bowel disease [48], and rs7517847 with Crohn’s disease [49]. Both those SNPs map to the same gene - \( IL23R \), which is known to have an important role in T cell functions and also proposed as a therapeutic target in inflammatory diseases [50]. rs11209026 is located in the exon of \( IL23R \) (missense SNP), whereas rs7517847 is located in intron [51], therefore provided that those are causing variants they have different effects on the \( IL23R \) gene.

However, this study was conducted based on the gene expression measured during the active disease, which may result from early causal genes. In the second project we have aimed to identify early putatively causal genes. Our hypotheses were that a) early regulators of T cell associated diseases could be found by defining upstream transcription factors (TFs)
in T helper cell differentiation TF-gene regulatory network (TF-GRN); and b) such early regulators can be further prioritized by the GWAS SNP enrichment in order to identify early putatively causal T cell master regulators.

Here, yet again we focus on T cells, and specifically on T helper cell differentiation, for the same reason presented above – genes involved in T helper cell differentiation pathway, are the most enriched for GWAS genes. Hypothetically, a T cell based GRN should be activated early in order to combat disease, and therefore abnormalities in first steps of T cell differentiation might have cascading effect on other downstream T cell genes.

We have therefore performed time-series profiling of gene expression in four major CD4+ T-cell subpopulations differentiating from Naive towards mature cells. Specifically, we have measured the transcriptomes of T helper cell type 1 (Th1), T helper cell type 2 (Th2), T helper cell type 17 (Th17), and T regulatory cells (Treg) at six time points. We found several genes that were differentially expressed during early Th1 versus Th2 differentiation\(^1\). In further support of the relevance of TF-GRN we found that early Th1/Th2 transcription factors (TFs) were the class of genes most enriched for known GWAS genes. Therefore, we proceed to create a Th1/Th2 TF-GRN. The time aspect is a crucial component to our experimental approach, as we aimed to identify upstream regulators of T helper cell differentiation. Those would intuitively change their expression very early during the cell differentiation process. Furthermore we have prioritized early TFs with many downstream targets – so called hub TFs.

The background is that previous data support that human beings can cope with a relatively big percentage of gene mutations if such mutations are located in the peripheral parts of the interactome, where the affected genes have fewer connections [22]. Peripheral genes tend to be non-essential, whereas genes located towards network centre tend to have more essential roles like tumour suppressors or oncogenes [22] and their mutations more frequently contribute to pathological consequences [22].

The importance of hubs (Fig.1) for maintaining the network topology is supported by many observations, and is in line with discussed above concept that even though individual edges in the network are doubtful hubs seem to be a robust network feature. Most networks, including cellular networks, have skewed node degree distribution (often described as scale-free) [52,53] – such networks are dominated by few key hub nodes – nodes with very

\(^1\)i.e. differentially expressed at early time points of Th1/Th2 differentiation.
many neighbours. The existence of hubs by chance is unprobable but it is reasonable from biological point of view [37]. In general, such networks are quite robust upon random node removal – which is in line with what we observe, that healthy individuals may harbour many mutations with no disease manifestation. However when hub-genes are removed it may have catastrophic effect on the network topology [2]. This is consistent with the effects of such genes being likely to flux over the entire network causing error propagation [17]. Moreover the importance of hub-genes is supported by the fact that they are conserved and tend to evolve slower than the non-hub genes [17], and that viruses evolved to target hub-genes [54,55]. Taken together these findings led us to prioritize hub TFs in the TF-GRN.

In summary, we have hypothesized that putative shared disease regulators can be inferred from a TF-GRN of T helper cell differentiation, where early differentially expressed hub-TFs enriched for GWAS-identified SNPs are prioritized. We identified three such hub-TFs: GATA3, MAF and MYB. This project gave us an insight into how the same ‘driver’ genes may be involved in multiple diseases:

Firstly - differential gene expression: we found that the three hub-TFs were differentially expressed at different levels, in different combinations and directions (up- or down-downregulation) in six out of eight studied diseases. Which might be a consequence of the SNP effect (as discussed in the next point).

Secondly - all three hub-TFs were enriched for SNPs that were associated with multiple different diseases, which means that the same gene might be affected by different SNPs, which in turn, may lead to different phenotypic effects.

Thirdly- differential splicing. In this project we have identified one probable reason why a gene gets differentially spliced in disease. We have shown an association between disease SNP and exon inclusion in GATA3 [43].

In principal, the first and last points might both be a consequence of SNPs (second point), however other factors might contribute to differential gene expression and splicing such as differential gene methylation.

For the validation of the upstream role of the three hub-TFs one would need to analyse samples from the pre-symptomatic stages of the disease. Such samples are difficult to obtain as patients generally present to the health care when they are already symptomatic. Therefore, a prospective study would be required or an access to a biobank. In the absence of the
samples from pre-symptomatic disease stages we have hypothesized that relapsing diseases in their symptom-free phases can serve as a model of non-symptomatic disease stages. Here we concentrated on two such diseases, seasonal allergic rhinitis (SAR) and multiple sclerosis (MS). SAR is an ideal disease model as symptoms occur at the same time during the year caused by known external trigger – pollen [56]. Therefore, we hypothesized that CD4+ T cells collected from SAR patients during winter time can then serve as a proxy of the early disease stages, whereas same cells in vitro allergen challenged would then represent active disease. Another disease we have used is MS. Here we hypothesized that remission can be seen as a proxy of symptom-free early disease stage.

Our results have shown that the three TFs or their splicing variants were differentially expressed in the symptom-free phases of those two diseases, whereas TF-targets were miss-expressed in the active disease.

In this thesis to this point we have concentrated on the transcription factors as the main cell regulators, although there are others like for example epigenetic modifications or noncoding RNAs. Hyper-methylation of some genes has been reported at early event of tumorigenesis [57] as well as tumour-type specific methylation landscape has been reported [57]. Furthermore DNA methylation changes have been shown to separate allergic patients from healthy controls [58]. However as pointed before in this thesis we concentrate on mRNA expression patterns as it is believed to reflect both genetics and epigenetics. Another example of a class of molecules that might affect gene expression are noncoding RNA molecules. For example, in cancer key mutational drivers have been associated with long non-coding RNAs (lncRNAs): [59]. However lncRNAs, are thought to be downstream effectors [59]. Another example of noncoding RNA molecules includes microRNAs (miRs). Those are short (~22nt long) noncoding RNAs that post-transcriptionally negatively regulate gene expression. It is estimated that microRNAs regulate around 60% of protein coding genes [60]. Canonically, microRNAs fine-tune mRNA expression by binding to the transcripts 3’UTR region and inducing either mRNA cleavage or causing inhibition of translation [60]. For the interaction to be present it is commonly accepted that the miR ‘seed’ region has to be perfectly or nearly perfectly complementary to the mRNA sequence [60,61]. Usually ‘seed’ denotes nucleotides at the 2nd to 7th position at the 5’ end of the microRNA [61] and has been shown to be most conserved among the miR sequence [61]. It
is commonly accepted that conservation may imply sequence importance because of evolutionary pressure [61]. However there is no single model describing all possible microRNA-mRNA interactions because of their heterogeneity [61].

miRs have been reported to have important functions in cell proliferation and development, differentiation, programmed cell death and stress response [62]. Furthermore, miRs are thought to be key mediators in the host response to infections, they are involved in innate and adaptive immune pathways [60]. Since miRs seem to play an important role in mediating host response to infection, their deregulation may in turn lead to susceptibility towards disease [60]. For example, *Mycobacterium tuberculosis* has been shown to supress miR-let-7f production leading to increased survival of this bacteria [63].

microRNA profiles in cells isolated from patients suffering from viral, bacterial or fungal infections differed from healthy controls [60]. This tendency has been also shown in vitro infected cells [60]. The crucial role of miRs in the immune system was also reported previously. They are involved in cell differentiation, development [64,65], activation [60], proliferation and apoptosis [65], cell lineage stage fate determination [60], maturation, and maintenance of immune homeostasis and regular function [66]. Changes in miR expression has been reported in a variety of human diseases, including autoimmunity, especially in T cell related autoimmunity [66]. For example silencing or reduction of *DICER* (an enzyme involved in the miRNA processing machinery) lead to reduced number of T cells in the thymus and periphery lymphoid tissues [66]. Furthermore, SNPs and mutations have been seen in some miRNA targeted genes and miRNA target sites, which are linked to causation and progress of some inflammatory autoimmune diseases [66].

One of the major advantages of miRs in clinical contexts is that they can be isolated from most biological fluids (extracellular miRs - ex-miRs); [60] and since it has already been proven that the ex-miRs signatures are highly predictive of an infection [60], they can serve as a possible disease markers.

Recognizing the importance of miRs, we have aimed to infer if microRNAs might have an upstream role in disease pathogenesis. It has been reported that miRs’ repressive effects on GRNs are cumulative [67]. On the system level miRs can be seen as stabilizers of GRN, which makes sense as it can be shown that the wider the spread of repressions over different nodes,
the more stable the network becomes [67]. One microRNA might regulate many mRNAs and one mRNA might be regulated by many microRNAs. Moreover, individual miRs are often seen to target proteins in the same or connected pathways [60], and are expected to act in combinations on their target genes [68]. Because of that we have decided to identify groups of miRs that cumulatively regulated the same targets – in other words we have aimed to identify miR-mRNA modules. Since targets of miRs can differ between cells [60] here again we have created cell type, and process specific microRNA-gene regulatory network (mGRN). We have taken into consideration not only sequence based miR-gene target predictions, but also miR-gene anticorrelation\(^2\) in T-cell differentiation. Furthermore we have created miR-mRNA modules – i.e. groups of miRs that are predicted to synergistically regulate same group of gene targets. We have tested disease relevance of those miR-mRNA modules in total of 11 CD4+ T-cell related diseases finding one module miss-expressed in all 15 microarray datasets studied. Moreover, this miR-mRNA module was highly enriched for disease associated genetic variants (SNPs).

Even though we haven’t found evidence for the upstream role of miRs in disease pathogenesis, we have shown miR involvement in T-cell differentiation and CD4+ T-cell related diseases, which is in line with previous reports.

In summary, this thesis reveals the importance of shared-multi-disease genes and shared up-stream regulators. This work sheds light on the potential explanations why the same genes might contribute to the pathogenesis of different diseases i.e. a) differential gene expression in different diseases (for example deregulation of a key TF would affect expression of its target genes); b) alternative splicing of a key TF may affect its target group of genes (different splice variants of the same gene might have different target genes); and c) SNPs might cause alternative splicing.

\(^2\)Here we have concentrated on anticorrelation since miRs are thought to negatively regulate gene expression.
Materials and Methods

In the section below summary of materials and methods is presented, concentrating on bioinformatics methods, which were in my main focus during this PhD. For remaining information please refer to original papers and manuscript.

Study subjects

PROJECT I

1. CD4+ T cells from 48 patients with seasonal allergic rhinitis (SAR) stimulated with glucocorticoids (GCs). Eight of these were classified as high responders (HRs; see below for definitions) and eight as low responders (LR); SAR was defined based on a positive seasonal history, and a positive skin prick test or a positive ImmunoCap Rapid test to birch and/or grass pollen. Patients with perennial symptoms or asthma were excluded. The healthy controls did not have any history for SAR and had negative tests results;

2. CD4+ T cells from 50 patients with multiple sclerosis (MS) stimulated with humanized monoclonal antibody which binds to α4β1-integrin (natalizumab-NZB). Eight of the MS patients were classified as HRs and eight as LRs (see below for definitions).

PROJECT II

1. Naïve CD4+ T cells were freshly isolated from four healthy donors and polarized towards T helper cell type 1 (Th1), type 2 (Th2), type 17 (Th17) and regulatory T (Treg);

2. CD4+ T cells from 10 SAR patients and 10 healthy donors. SAR and healthy subjects were defined as in project I (see above);

3. CD4+ T cells from 14 SAR patients before and during two years of treatment with immunotherapy, and six healthy donors;

4. Eight patients with symptomatic influenza A or B, which was verified with PCR experiment of nasal secretions;
5. 10 patients with active pulmonary tuberculosis, and matched six controls from Lima, Peru;

6. Eight breast cancer patients sampled before radical surgery;

7. CD4+ T cells from 10 healthy donors and 10 MS patients diagnosed with definite relapsing-remitting MS. Patients did not undergo corticosteroids or immuno-modulatory treatment for at least two months before sample collection, and were relapse-free for at least three months.

PROJECT III

1. Naïve CD4+ T cells were freshly isolated from four healthy donors and polarized towards Th1, Th2, Th17 and regulatory T – same biological material as in project II;

2. CD4+ T cells from 10 SAR patients and 10 healthy controls stimulated with mock (PBS) or grass pollen extract for 17h; SAR was defined as in project I (see above).

Ethics statement
All studies participants provided written consent for participation.

PROJECT I
Studies were approved by ethics boards of University of Gothenburg and Linköping. Diary number: 2013/12-31, Mikael Benson, M180-07 and T130-09, Jan Ernerudh.

PROJECT II
Studies were approved by ethics boards of University of Gothenburg, Lima, and Linköping. Diary number: 2013/12-31, Mikael Benson, M180-07 and T130-09.

PROJECT III
Study was approved by the ethics board of Linköping University. Diary number: 2011-292/31, Mikael Benson.
Definition of low and high responders to treatment
Response to treatment was assessed by asking patients to mark their symptoms on a visual analogue scale. The symptoms were rhinorrhea, congestion, and itching. The patients marked their symptoms before and after two weeks of nasal treatment. LRs/HRs were defined based on decrease in symptom scores when treated for two weeks during the pollen season.

Fifty patients with MS were treated with NZB, and classified as LRs or HRs depending on if they had had at least one relapse, or no relapse at all during the three-year follow-up period. Expression profiling was performed before starting NZB medication.

GWAS data accession

PROJECT I
GWASs data were downloaded from GWAS Catalog National Human Genome Research Institute (NHGRI). Data included 256 diseases and traits. SNPs were mapped to the closest up- and down-stream genes at the significance threshold $10^{-5}$.

PROJECT II
GWASs data were downloaded from GWAS Catalog National Human Genome Research Institute (NHGRI) on 6th March 2014. All categories were first classified manually as disease traits (421) and non-disease traits (574). Non-disease traits were removed from further analyses. All remaining SNPs were mapped to the closest up- and down-stream gene. All genes with at least one disease SNP mapped to are referred as GWAS genes. All nominally disease associated SNPs i.e. SNPs with $P<10^{-3}$ were downloaded from [69] version 3 (2014/03/06) for SNP frequency enrichment analyses.

PROJECT III
GWAS SNPs were obtained from GWASDB2 (version 2, October 2015) [69]. SNPs in high linkage disequilibrium (LD; $r^2 > 0.8$, 1000 Genomes Phase 3 v5) [70] were identified with SNiPA-tool [71] (available at http://www.snipa.org, accessed October 2015). SNPs were further mapped to the genes within 5kb flanking genes region.
Enrichment analyses

PROJECT I

Pathway enrichment analyses of GWAS genes was tested using all pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG) [72], Ingenuity Pathway Analysis (IPA; Ingenuity® Systems) [73], and Gene Ontology (GO) [74]; all accessed on 15th November 2013. As a background all human genes annotated in NCBI (National Center for Biotechnology Information) were used. Mouse knockout phenotypes were downloaded from [75] on 31st January 2013. Therapeutic target drugs and disease or prognosis markers were downloaded from IPA database. Enrichment $P$-values were calculated using Fisher’s Exact Test (FET) using all human and mouse genes as background respectively. In order to control possible connectivity biases during enrichment analyses of shared disease genes we have performed permutation test of genes in STRING [76] preserving median and minimal degree to be the same as for the shared disease module. $P$-values were corrected for multiple testing with Benjamini Hochberg correction method. Size effect was calculated as fold enrichments (FE) i.e. frequency of genes in the tested set having a certain characteristic divided by the frequency of the characteristic genes among all annotated genes.

PROJECT II

Pathway enrichment analyses of $GATA3$, $MAF$ and $MYB$ predicted targets were tested with Ingenuity Pathway Analysis (IPA; Ingenuity® Systems) [73].

PROJECT III

Pathway enrichment was performed using Gene Ontology terms with DAVID using all microRNA target mRNAs in microRNA-gene regulatory network (mGRN) as background. For SNP enrichment analyses we performed permutation test ($10^6$ randomizations) having variant frequency as a test statistic.

Public T-cell disease data

PROJECT I and II

Public T cell related disease microarray data were downloaded from the Gene Expression Omnibus (GEO) repository on 31st December 2013 that consisted of i) unstimulated CD4+ T cells profiling from healthy controls
and patients with T cell-related non-virus diseases; ii) at least five patient samples and controls; and iii) the patients were not undergoing drug treatment and were not symptomatic, iv) samples were not activated in vitro (GSE4588, GSE50101, GSE13732, GSE14924, GSE14317, GSE8835, GSE12079).

PROJECT III
The same public gene expression data as in project I and II were meta-analysed, plus additional three (GSE6740, GSE9927, GSE17354) that followed the same criteria as before apart from that they had to contain at least three samples per patients and controls (in contrast to the previous requirement of 5 samples per condition). microRNA microarrays followed the same criteria on 7th June 2014 (GSE21697, GSE31629, GSE24022).

Gene expression microarray data processing
PROJECT I
The GCs and NZB materials were corrected for potential batch effects using COMBAT [77], using phenotype and stimulation as covariates. All probes were mapped to the corresponding genes. If there were multiple present probes - the median probe levels were used. The LIMMA package in R was used to compute differentially expressed genes.

PROJECT II
Microarray data were processed as described in project I with the exception that if a gene was represented by more than one probe mean instead of median values over the probes were considered. Differential gene expression analyses were done using maSigPro in order to find differentially expressed (DE) genes between the time-series (default parameters) [78]. Early differentially expressed genes were defined as a subset of maSigPro DE genes that were further DE at 6h or 24h, Th1 vs Th2 in either Tippets or Fishers method for combining LIMMA P-values. Individual genes were assumed to be DE if double-sided $P<0.05$ after correction for multiple testing with the Benjamini Hochberg procedure.

PROJECT III
Microarrays were processed and differentially expressed genes were defined as in project I. Paired test was used when possible.
Analyses of sets of genes

PROJECT I
In order to test if a set of genes was differentially expressed mean of the squares of the student \( t \)-test statistic was calculated and compared to random distribution obtained from permutation test (10\(^6\) randomizations). Additional analyses were performed using the -log \( P \)-values giving similar results. Small \( P \)-values were approximated using [79]. Fold enrichments (FEs) were calculated as described above.

PROJECT II
In order to define if a subset of genes was DE we have used similar method as described in project I. We have used mean -log10 \( P \)-values of the group as a permutation test-statistic (10\(^6\) resampling). Small \( P \)-values estimation was done as described before in project I. Effect size of enrichment analyses was calculated as odds ratios (ORs), i.e. ratio between mean value of the test statistic in the test group and average value in is complementary group.

PROJECT III
Left-sided –log transformed Pearson’s correlation \( P \)-value was used as test statistic for permutation test checking if a set of microRNAs and microRNA–target mRNA were anti-correlated (10\(^6\) permutations).

Disease-associated network modules

PROJECT I
Both the network and modules were constructed using a modified method described in [14]. All maximal cliques [80] in the human protein-protein interactions (PPI) database – STRING [81] (version using 9.03 interactions with a confidence score \( \geq 0.7 \)) were assigned a weight. This weight was defined based on the differential expression analysis as a sum of the -log(\( P \)-values) of all genes in the clique. Furthermore permutation test was performed randomizing \( P \)-values for all the genes in the differential expression analysis, and the randomized weight for each clique was calculated using same method as above. In order to obtain \( P \)-value of each clique’s weight the fraction of randomized weights was calculated that resulted in a higher value than the original, ‘real’ weight. Cliques that obtained \( P \)-value
smaller than 0.01 were than mapped back to PPI forming module. Whole procedure was repeated for each disease resulting in a module associated to each disease. For estimation how the disease modules intersection size differs from random, all the above analysis were repeated 100 times, again randomizing $P$-values. For each of the randomization step the intersection size between randomized disease modules was calculated. Obtained 100 module intersection sizes from permutations were next compared with the ‘real’ intersection size. Additionally those 100 sets of disease modules were also used to identify genes present in more disease modules than expected by chance.

PROJECT III
For disease association microRNA-mRNA modules were tested for differential expression in public disease expression data (see above). For each module $P$-values were combined with the Stouffer method [82]. Module has been assumed to be disease associated if combined $P$-value was lower than 0.05.

Shared disease module validation in several human interactome databases
PROJECT I
Latest versions of HPRD [83], Reactome [80], Intact [84], HI2 [84], and a high confidence database [85] were downloaded on 15th November 2013. Shortest distances between all pairs of nodes in the shared disease module that were also present in the largest connected component (LCC) were calculated. Next 1,000 random set of genes were selected from the largest connected component preserving the same number of genes as in the overlap of shared disease module genes and LCC. $P$-values were obtained using Wilcoxon test of the mean values of the shortest paths.

Classifications
PROJECT I
Classifiers were built using LASSO with MATLAB function lassoglm(). $\lambda$ parameter was chosen from leave-one-out cross-validation at the minimum deviance.
PROJECT II

In order to classify patients versus controls Support Vector Machine classifier was made with using MATLAB function `svmtrain()`. Leave-one-out validation followed by the calculation of area under ROC curve (AUC) were performed in order to assess classifier performance. AUC was estimated with MATLAB function `trapz()`. Expression of GATA3, MAF and MYB were tested as discriminant features, which allowed for good classification of 6 out of 8 diseases (AUC>0.85), whereas three transcription factors (TFs) downstream targets separated patients and controls with high AUC in all 8 diseases studied (AUC>0.95).

Th1/Th2 regulatory region analysis

Accessible DNA regions were defined by DNase peaks in Th1 and Th2 cells downloaded from Encode database. For those regions methylation expression of Naiïve T-cells (NT), and 7 days polarized Th1, and Th2 cells with custom methylation microarrays (Agilent Technologies) (79,789 enhancers in total) was measured. Quantile normalization has been used. Each regulatory region was mapped to the upstream gene. In order to determine if a region is active it has been compared to the distribution of all regulatory regions – i.e. enhancer was assumed non-active if its methylation probe level was higher than two standard deviations above the average probe level. All genes with at least one active regulatory region mapped to in NT, Th1 or Th2 cells were assumed to be possibly accessible for regulation in Th1/Th2 cells. Additional GATA3 ChIP-Seq was downloaded from supplementary material in [86] – those binding sites were used to validate TF-GRN in Th1 and Th2 cells.

Construction of the Th1/Th2 GRN

From the list of differentially expressed (DE) genes (see above) and Th1/Th2 accessible genes a subgroup of transcription factors (TFs) was selected. List of human TFs and co-factors was downloaded from Animal TF database [87] on 1st January 2013. Further sequence based TF-mRNA predictions (TFBS) were downloaded from [88].

Prior confidence of TF-target interaction was represented by confidence interaction matrix which for i-th TF and j-th target gene was defined as
follows:

\[ P_{ij} = \begin{cases} \frac{m_i}{n}, & \text{if TF } i \text{ was predicted to bind gene } j \\ 1 + \frac{m_i}{n}, & \text{if TF } i \text{ was not predicted to bind gene } j \end{cases} \]

Where \( m_i \) denotes the number of TF \( i \) predicted targets.

Expression difference (\( y \)) in Th1 (\( x^{Th1} \)) and Th2 (\( x^{Th2} \)) cells for \( i \)-th gene and each time point \( t_k \) of polarization was defined as:

\[ y_i(t_k) = x_i^{Th1}(t_k) - x_i^{Th2}(t_k) \]

The mean of forward and backward differences at the internal points was estimated as the rate of change:

\[ \dot{y}(t_k) = \frac{y_i(t_k + 1) - y_i(t_k - 1)}{2} \]

For time evolution, for all Th1/Th2 genes, linear model of following form was used:

\[ \dot{y}(t_k) = \sum_j w_{ij} y_i(t) \]

Values of \( w_{ij} \) were inferred using LASSO [89] with MATLAB function `glmnet()` selecting \( \lambda \) values from 10-fold cross-validation using one standard error criteria.

LASSO inference formula:

\[ \min_{[w_{ij}]_{i=1}^{K}} \sum_{k=1}^{K} \left( \bar{y}_i(t_k) - \sum_{j \in TF} \left( w_{ji} y_j(t_k) \right) \right)^2 + \lambda_i \sum_{j \in TF} P_{ij} |w_{ji}|, \forall i \]

**TF-target relationship correlation analyses**

**PROJECT II**

The correlation analysis for the TF-target relationships for the eight public diseases analysis was carried out from geometric mean of the \( P \)-values of its predicted targets for each of the eight diseases, and the corresponding \( P \)-value of the TFs (from LIMMA).
ChIP-Seq analyses

PROJECT II

All reads were aligned to genome version hg19 using Bowtie2 [90] (default settings). At this step technical replicates were merged. Peaks were determined using MACS14 [91], with the P-value cut-off of $10^{-4}$. During peak calling for GATA3, input was ignored since this sample has failed. For MAF and MYB appropriate inputs were used. All peaks were then further mapped to genes with transcription starting site in the range of 4kb upstream and downstream from the peaks (BEDTools [92]). Weighted permutation test was performed in order to validate if TF-GRN–predicted GATA3, MAF, and MYB targets are enriched for peaks with lower peak P-value (higher confidence) i.e. bootstrap P-value was calculated using weighted mean number of peaks mapped to TF-targets as test statistic checking if it is higher than for random gene set of the same size. $10^3$ permutations were done. Weights were calculated as 1- geometric mean of the peaks P-values.

Gene Regulatory Network and TFBS comparison using ChIP-Seq peak counts

PROJECT II

For comparison of TF-GRN–predicted TF targets with only sequence based predictions (TFBS) permutation test was performed using all network genes as a background. Test statistic was defined as the ratio between the average number of peaks mapped to the TF-GRN–predicted TF targets and mean number of peaks mapped to TFBS TF targets, $10^3$ permutations were performed.

Splice associated SNPs

PROJECT II

All GWAS SNPs were mapped to the predicted splice regulatory sites (SRS) of GATA3, MAF, and MYB with RegRNA2.0 online tool [93]. SRS were defined as exon splicing silencer, exon splicing enhancer, intron splicing silencer, intron splicing enhancer and splicing sites.
Expression quantitative trait loci
PROJECT II
*GATA3* expression quantitative trait loci (eQTL) analyses were performed as follows: both genotypes data and normalized exon counts were downloaded from the gEUVADIS repository [94]. By fitting linear equations association between genotypes and exon counts was computed. *P*-values were calculated with *F*-test. In this analyses only European samples were analysed. Here we have used Ensembl genome version GRCh37.

siRNA mediated gene knock down
PROJECT II
For identification of splice variant specific targets Pearson’s correlation analysis over all the 6 exon arrays were performed. All gene variants of all TF-GRN–predicted *MAF* targets (103 targets) were tested. Splice variant target association has been assumed if correlation *P*<0.05 and FDR<0.1.

Sublingual immunotherapy (SLIT)
PROJECT II
Subject samples were collected at the three time points: a) before treatment, immediately prior to initiation of SLIT; b) after one; and c) two years of treatment. The treatment for birch allergy was based on an 11-day progression of doses phase for Staloral® (Stallergenes SA), (from 10 to finally 300 IR/ml). Drops were kept under patients tongue for 2 min before swallowing. The dosage of 240-IR was kept for 5-6 months. Next 6-months were with no treatment. All patients responded positively to treatment.

Quantitative PCR (qPCR)
PROJECT II
For expression calculations comparative ΔCt method was used, having *ACTB* as an internal control. qPCR measured gene expressions were compared using one-sided Wilcoxon test.

PROJECT III
For expression calculations comparative ΔCt method was used having *GUSB* and hsa-mir-423-5p as internal controls for mRNA and microRNA, respectively.
MS gene expression in relapse and remission

PROJECT II

Meta-analyses were done of material originally published in [95] of 113 samples of MS patients and healthy controls with most of the samples being paired. In this study additional patient information not published previously was used such as the time-point for their remissions, the status of their recovery, and their treatments. This allowed for identification of patients with relapse at their first visit and remission at the second one. At both visits patients had a clinically isolated syndrome. Relapse was defined as within 90 days from first clinical symptoms, whereas remission as more than one year since last clinical symptoms occurred and where patients clinically recovered (judged by a physician). Those requirements let for identification of the subset of patients used in this study: 30 patients with remission (with 10 matched controls). Eleven of those patients were also sampled during relapse.

GWAS analysis of multiple sclerosis

PROJECT II

Original data from the MS consortium study [96] consisting of genotype array of all common SNPs of ∼25,000 individuals was reanalysed. SNPs mapped to GATA3, MAF and MYB with $P$-value<0.001 (total of 9 SNPs) were further queried at the regulome database [97] (accessed on 18th Jun 2015) in order to check if any of them might have regulatory effects. Analyses revealed, seven such SNPs (regulomedb score 1-6). Disease severity association of those seven SNPs was than tested in an independent study of 2,085 MS patients published in [98]. Additional, previously unpublished disease severity data was used: 252 (of 2,085) patients had severe disease variant. For significance assessment if patients with severe disease had more SNPs permutation test has been used.

Construction of miR-mRNA regulatory network and module identification

PROJECT III

For microRNA-mRNA regulatory network construction microRNA-mRNA sequence based predictions (TargetScan 6.0 [99] and miRanda [100]) and
microRNA-mRNA anti-correlation analyses (Pearson’s correlation, left-sided test) were combined. microRNA was assumed as a putative regulator of an mRNA if there was a predicted microRNA-mRNA interaction by both miRanda and TargetScan algorithms and if the microRNA and mRNA were significantly anti-correlated in the time series data of Th1/Th2 polarization. For module construction miRsynergy tool [101] was used with following parameters: alpha=2; merge.tol=2; density1.tol =1e-2, density2.tol=5e-3. Gene-gene interaction network was obtained from STRING 9.1 database [76].

Modified Shannon entropy score [102] was calculated using $-\log_{10}$ transformed Stouffer [82] combined $P$-values for each module and time-point in order to identify time-point specialization in the modules.
Results and Discussion

Since the advent of Genome Wide Association Studies (GWASs), many genomic loci have been associated with more than one trait and/or disease. One striking example is \textit{NOTCH4}, which has been associated with multiple, highly diverse diseases such as rheumatoid arthritis, ulcerative colitis, and lung cancer in ever smokers, age-related macular degeneration, shingles, plantar warts, and strep throat (GWAS Catalog; [103]). In contrast to the highly pleiotropic \textit{NOTCH4}, another example, \textit{CEP290}, illustrates that the same gene can have pleiotropic effects within the same organ system. Variants in this gene are associated with retinal degeneration and intellectual disability, both of which affect the nervous system [104]. Generally, disease gene pleiotropy\(^1\) has become a widely recognized phenomenon, and gave rise to the main question behind this thesis: \textit{How can the same genes may contribute to pathogenesis of different diseases?}

In order to address this question, the \textbf{first aim} of my thesis was to define a group of genes involved in the pathogenesis of more than one disease – shared disease genes. The \textbf{second aim} was to examine if it is possible to identify shared upstream regulator or regulators that control genes in multiple diseases. The \textbf{third aim} was to study potential mechanisms by which shared disease genes and regulators can be associated with different diseases.

\textbf{Project I: Identification of shared disease genes}

Even though each cell in our body has the same DNA sequence, tissues, organs and cell-types are highly diverse functionally and phenotypically. They express different genes at different levels. For example, \textit{LEF1} has been shown to be associated with tissue-specific functions in twelve different tissues [105]. Epigenetic factors such as DNA methylation or non-coding RNAs may also cause disease-associated changes that vary in different cell types or tissues. This complexity has two implications for the study of shared genes or their products: 1) mRNA may be an optimal starting point for such studies, because it may reflect both genetic and epigenetic variants; 2) to limit the confounding effects of cell- or tissue-specific variation in

\(^1\text{disease gene pleiotropy - gene association with more than one disease phenotype.}\)
mRNA expression, it may also be optimal to focus on one cell-type that is highly relevant to many diseases – i.e. a cell type whose impairing might lead, or highly contribute to, many disease phenotypes.

SNPs might affect genes in a variety of ways. In the case of complex diseases, many SNPs are located in non-coding regions (including introns) [10]. Those SNPs might affect gene expression, TF binding sites, alternative splicing, chromatin accessibility, DNA methylation, small RNAs, large intergenic non-coding RNAs (lincRNAs), RNA editing, and mRNA degradation [106]. On the other hand, intergenic SNPs might affect alternative splicing, open reading frames, amino acid inclusion leading to change of protein function. For example, around 2/3 of disease missense mutations result in the perturbation of the interactome, of which 50% lead to complete loss of interactions, mainly due to protein misfolding or impaired expression [107]. The other half leads to alteration of part of the interactions [107]. However only about 7% of GWAS SNPs can be found in protein-coding DNA regions [108].

Interpretation of GWASs results should be done with caution, because of biological complexity and technical limitations. SNPs are commonly mapped to the host gene in order to identify disease associated genes via genetic variants. Intergenic SNPs are often mapped to the closest upstream and/or downstream gene. The criteria for mapping of those SNPs may have important implications for pleiotropy, which can be difficult to dissect. For example, most human genes (nearly 80%) expressed in whole blood have a local eQTL (local variant affecting gene expression; [109]). Those eQTLs often have small effect sizes [109], and may also affect distal genes. Thus, different eQTL mapping to the same genes may have pleiotropic effects. Furthermore SNPs identified with GWASs represent all SNPs in the high linkage disequilibrium (LD) and therefore might not necessarily increase disease risk themselves but rather mark nearby DNA loci that does [108, 110].

As exemplified by NOTCH4, above, different genetic variants in the same gene may be one of the explanations why the same genes can be involved in multiple disease pathogenesis. However, it is also possible that the same gene and variant may have pleiotropic, but opposing, effects in different diseases. For example, SNP rs2476601 (mapped to gene PTPN22) increases risk for type I diabetes, but is protective against Crohn’s disease [11].

Furthermore GWASs generally tend to identify only common genetic variants [10], which in general only explain a small proportion of the ob-
served heritability [105]. Moreover, even though GWASs are generally genome wide – some regions are poorly covered, like for example copy number variable regions [10].

However, despite their limitations GWASs are the most comprehensive source of disease associated genetic variants. It has been shown that genetic variants identified in GWASs tend to perturb regulatory modules specific to cell types or tissues that are important for disease [40]. Therefore, we have hypothesised that a cell type expressing pathways enriched for genes with variants associated with multiple diseases, would in turn, be of relevance for multiple diseases. Therefore, we have proceeded with pathway enrichment analyses of genes to which at least one disease associated SNP have been mapped (henceforth referred to as ‘GWAS genes’). This revealed the T helper cell differentiation pathway to be most enriched for GWAS genes (Bonferroni-corrected $P < 10^{-15}$ Ingenuity Pathway Analysis-IPA [73]; $P < 2 \times 10^{-6}$ Gene Ontology [74]; $P < 4 \times 10^{-3}$ Kyoto Encyclopedia of Gene and Genomes [72]). Moreover, the T helper cell differentiation pathway was also enriched for genes annotated to diseases in Online Mendelian Inheritance in Man database [5] ($P < 2 \times 10^{-11}$). Since alleles are non-randomly associated at different loci we have also verified pathway enrichment based on all genes in linkage disequilibrium with the GWAS genes again finding T helper cell differentiation pathway enriched ($P < 3 \times 10^{-7}$).

However, this result might have been confounded by overrepresentation of immune-related diseases studied with GWASs. To assess this possible confounder, we manually classified diseases as immune and non-immune2, and repeated the analyses, once more finding highly significant enrichment of the T helper cell differentiation pathway ($P < 1 \times 10^{-7}$; IPA).

This is consistent with T cells having key regulatory roles in many different diseases, such as autoimmune and allergic diseases (5), atherosclerosis (6), cancer (7), and obesity (8). In malignant diseases, tumour-infiltrating T cells have been associated with prognosis in malignant melanoma [111] and ovarian cancer [112]. In HIV-positive patients with anal cancer, lower CD4+ T cell counts were associated with worse treatment outcome [113]. Similar to CD4+ T cells several reports associate CD8+ T cells with improved survival in different forms of cancer [114, 115, 116]. In contrast to CD8+ T cells, FoxP3+ regulatory T cells, have more complex links to prognosis. In some solid tumours these cells are associated with poor prog-

2Diseases were manually classified as immune and non-immune by a medical doctor.
nosis [115], while in colorectal, head and neck, and oesophageal cancers FoxP3+ Tregs are associated with good prognosis [115]. Taken together, these data support the hypothesis that profiling of T cells immune cells may help to understand common disease mechanisms.

For that reason and since T helper cell differentiation pathway is expressed in CD4+ T cells throughout this thesis we have focused on transcriptome profiling in CD4+ T cells.

Many studies have shown that disease relevant genes are highly connected, and co-localise [14,17,21,22] in the interactome forming highly interconnected modules, which, if dysfunctional may lead to disease [21,24]. The main advantage of a module-based approach is that modules are thought to prioritise limited numbers of highly disease-relevant genes. Even if not all genes in the disease-associated gene module are differentially expressed (corrected for multiple testing $P < 0.05$) the majority of genes would exhibit altered expression levels. In other words, we assume that a modest but consistent change of expression of many interacting genes might be as, or even more, important for disease pathogenesis as a pronounced, big change of an individual gene.

Therefore, we have proceed to identify disease modules based on expression profiling data from eight diseases (inflammatory and malignant), namely: four inflammatory diseases (allergy [(GEO) accession GSE50101], multiple sclerosis [GSE13732], rheumatoid arthritis [GSE4588], and systemic lupus erythematosus [GSE4588]), and four malignant or proliferative diseases (acute myelogenous leukemia [GSE14924], adult T cell leukemia [GSE14317], chronic lymphocytic leukemia [GSE8835], and hypereosinophilic syndrome [GSE12079]). Data sets choice criteria are described in the materials section. For all eight diseases we have identified genes that were significantly differentially expressed when comparing patients versus healthy controls. Next, we mapped these genes to their protein products in the Protein-Protein Interaction (PPI) network (STRING [102]). In agreement with previous studies of disease-associated genes we found that those genes instead of being dispersed in the human PPI were rather co-localised, forming modules (highly connected structures). This lead to identification of eight disease modules each consisting of 1,215 to 1,933 genes (interconnected differentially expressed genes).

Because of the heterogeneity of the studied diseases, we expected that the corresponding modules would be dispersed in the PPI network. Instead,
Protein-protein interaction map

Gene expression of patient/controls from T cells

Identification of disease susceptibility modules for each disease

Figure 3: The shared disease module (marked using a solid black circle) identified by analysis of expression profiling data from eight CD4+ T cell-associated diseases: allergy (A); acute myelogenous leukemia (AML); adult T cell leukemia (ATL); chronic lymphocytic leukemia (CLL); hypereosinophilic syndrome (HES). Black nodes represent proteins, lines possible gene interactions. From Gustafsson et al., Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment. Genome Med 6, 17 (2014). Reprinted with permission.

they tended to overlap (Fig.3). Consequently, we assumed that shared disease genes lie in the intersection of those modules. Those genes in the intersection of disease modules were generally highly interconnected forming a yet another module – here referred to as a ‘shared disease module’. The shared disease module consisted of 158 genes with 7,144 interactions between genes compared to 376 expected for 158 random genes ($P<10^{-300}$).

In order to confirm this finding we have analysed 5 additional versions of the human PPI networks (Materials and Methods) showing that the 158 shared disease genes had lower average shortest path between each other when compared to random sets of 158 genes ($P<10^{-3}$ in four versions and $P<0.011$ for the fifth version).

The shared disease module was enriched for key cell pathways like metabolism, proliferation etc., as well as enriched for known disease markers and therapeutic targets ($P<10^{-46}$ and $P<10^{-28}$ respectively), indicating its importance not only in healthy cells but also in active disease.

Modules in general are associated with robustness of biological systems [28, 117]. In general high connectivity is often linked to redundancy compensating loss-of-function mutations [117]. However same property might lead to cascading effect of local failures [117]. Indeed sparser net-
works are shown to be more robust [118]. Interestingly in [117] it is shown that modules exposed to greater environmental variation are generally more robust and contain less interactions [117]. On the other hand more essential modules (i.e. modules enriched for essential genes\(^3\)) are located central to the cell and are more densely connected [117]. Examples of internal processes include: transcription initiation, ribosomal subunit, mRNA processing modules. On the other hand cell wall organization, endosomal transport are examples of external cell processes less connected and not enriched for essential genes [117]. Even though this study is focused on *Saccharomyces cerevisiae* authors note similar results for other organisms including bacteria and roundworm and even identified similarities to the computational systems security [117]. Therefore, those findings might be generally applicable. In fact, we also observed that the shared disease module, that we believe is of importance for many immune related diseases, was highly interconnected with 19 times more interactions between genes than expected by chance (see above), which might explain why this specific part of the PPI is shared by many diseases.

However, although encouraging those results cannot be conclusive about the casual role of the genes in the shared module. Therefore, we proceeded to analyse a mouse knockout database, which showed that many of the shared module genes, when knocked out, lead to multiple phenotypes and diseases in mice \((P<10^{-50})\). The most significant mice phenotypes were abnormality of the hematopoietic and immune systems, and tumorigenesis. We also found a GWAS and cancer gene enrichment among shared disease genes \((P<10^{-15} \text{ and } P<10^{-41} \text{ respectively})\). Further, although the shared module genes were only based on analysis of eight diseases, we also found that those genes were generally enriched for associations to diseases and traits analysed with GWASs (and not only immune-related diseases and cancer; \(P<4x10^{-3}\)). All of those results support the general relevance of the shared disease module genes for involvement in disease pathogenesis.

The main limitation of this module-based approach is that it depends on the definition. Even though the core shared module genes should remain the same, the module boundaries are not strictly defined and can be altered upon change of module definition. For example, the shared module could be defined as genes shared by at least two diseases, as well as genes belonging to all eight disease modules (as used throughout this project). In this context

\(^3\)Here essential genes denotes genes whose knock-out leads to cell death.
we found it important to test if the shared disease module enrichment in GWAS and cancer genes depended on the definition. We found a strong correlation between degree of gene ‘sharedness’ – i.e. number of disease modules a gene belongs to, and fraction of GWAS (PCC = 0.91, $P < 4 \times 10^{-4}$) and cancer genes (PCC = 0.94, $P < 2 \times 10^{-4}$), (Fig.4). In other words, genes belonging to a larger number of disease modules, were more enriched for GWAS genes. This finding supports that shared disease module enrichment in both GWAS genes and cancer genes does not depend on the exact module definition.

It is also important to keep another potential limitation in mind – modules are defined using models of the human PPI network, in which many protein-protein interactions are missing [119].

Here, in this project we also examined the possibility of using shared disease genes to predict treatment outcome in order to further strengthen the clinical importance of those genes. We carried out a prospective study of 48 patients with Seasonal Allergic Rhinitis (SAR) (Materials and Methods). SAR patients were classified as high- or low-responders (HR and LR respectively) to a common drug, glucocorticoids (GCs), by clinical specialists (Materials and Methods). In order to mimic the molecular responses to treatment with this drug, CD4+ T cells were isolated from peripheral
blood from these patients during symptom-free periods (outside of pollen season). Next, the cells were stimulated in vitro with allergen extract and GCs. Based on gene expression microarray analyses we noticed an increased likelihood of a response to GCs among ‘more shared’ genes – i.e. genes belonging to more disease modules (PCC = 0.79, P = 0.0011; Fig.5A).

Furthermore, we created a classifier that correctly separated high- and low-responders (P < 3x10⁻⁴, Fig.5B). The classifier was built based on 311 genes affected by the drug (found as differentially expressed) and shared by at least 2 diseases (belonging to at least 2 disease modules).

We performed a similar study in patients with another relapsing inflammatory disease, multiple sclerosis (MS). These patients were treated with
a drug specifically targeting MS, natalizumab, and classified as HR or LR (Materials and Methods). In order to study their molecular responses to the drug, we first extracted CD4+ T cells from the patients during remission (symptom-free periods). Next, we treated the CD4 + T cells with nataluzimab and analysed them with gene expression microarrays.

In contrast to SAR, we found that MS disease-specific genes were targeted by the drug rather than shared disease genes (PCC = -0.61, P = 0.079; Fig.5A). Therefore, we hypothesized that MS disease specific genes (belonging uniquely to MS disease module) would serve as better discriminators when constructing classifier to separate HR and LR. Indeed, a classifier made based on 28 MS specific genes that were affected by natalizumab correctly stratified HR and LR (P<3x10^{-4}, Fig.5B).

**Project II: Identification of the upstream regulators**

In the section above, SNP enrichment in shared genes identified by expression profiling supported causality. Also, since those SNPs differed between patients with different diseases, this provided a potential explanation for why the same genes can contribute to pathogenesis of different diseases. However, the differentially expressed genes were identified during active disease, which may be secondary to early causal genes. Therefore, in the second project we aimed to identify early regulatory genes.

One approach may be to link disease-associated traits and diseases that have a shared genetic architecture [120]. For example, variants in the gene IKZF1 are associated with both mean corpuscular volume (trait) and acute lymphoblastic leukemia (ALL; disease) [120]. Increase of average corpuscular volume is being noted before ALL is diagnosed [120] and therefore the trait is assumed to increase disease risk. This would support the causal role of the shared genes. However, this is an inference with many possible confounders. Another approach may be based on using known protein interactions to infer up-stream regulators of genes that are differentially expressed during active disease [121]. However, this approach is limited to only direct upstream gene interactors and therefore might not be sufficient in search for causal regulators.

Identification of causal genetic drivers of human diseases has also been presented in [122]. Here authors tested candidate copy number variation genes whose copy number is informative of gene expression and have predictive value of transcription factors (TFs) activity. However, this approach
is limited by the requirement of a large number of sample-matched gene expression and genetic variant profiles.

In contrast, we hypothesized that upstream master regulators can be identified in a healthy T helper cell differentiation process. The background to this hypothesis is our previous finding that the T helper cell differentiation pathway is highly enriched for GWAS genes, as discussed above. Also, the T helper cell differentiation pathway is possible to mimic in vitro, using well-defined methods. This allows time-series analysis of the differentiation processes, starting with early stages. Based on this background, we proceeded to profile differentiation of four major CD4+ T cell subsets – namely T helper cell type 1 (Th1), T helper cell type 2 (Th2), T helper cell type 17 (Th17), and T regulatory cell (Treg), from Naïve to mature cells in an in vitro model. We measured gene expression profiles with microarrays at six time points (6h, 24h, 3 days, 6 days and 8 days) in four biological replicate biological experiments.

Subsequently, we identified that early Th1 and Th2 transcription factors (TFs, that exhibit altered gene expression early in the differentiation process, i.e. at 6 and/or 24h; Fig.6) were the most enriched for disease associated genes (GWAS genes; OR = 2.7, $P = 1.0 \times 10^{-7}$).

TFs are one of the key gene expression regulators in cells. Since one TF might regulate expression of many downstream genes errors due to, for example, disease-associated SNPs, would propagate through the signalling cascade altering TF-target gene expression. In order to identify such upstream key regulators we constructed a healthy TF-Gene Regulatory Network (TF-GRN) of Th1/Th2 differentiation.

A TF-GRN is a collection of TFs and their predicted gene-targets. It can be graphically represented as a set of nodes (genes) and edges (interactions) (Fig.1).

Network robustness and wiring might be altered by disease gene activity changes due to genetic or environmental factors, which in turn might lead to the change of the network output [104]. Commonly networks were representative of a static time point, and thereby missing insights into process dynamics [104]. In this project we have partially addressed this issue by introducing time-series transcriptome profiling of differentiating T cells. Time in general is an important factor. Therefore, we used time-series gene expression profiling of Th1/Th2 differentiation and Transcription Factor Binding Sites predictions (TFBS) to construct a TF-Gene Regulatory Net-
Figure 6: Differentially expressed genes in time series microarrays enrichment in GWAS genes (genes with at least one disease-associated SNPs identified by GWASs mapped to). Each T cell subset ("Th1", "Th2", "Th17", and "Treg") represent genes differentially expressed between the 6 and 24 hours of polarization in each cell type respectively. Enrichment $P$-values were obtained with Fisher’s exact test ($\ast\ast\ast\ast P < 1.0 \times 10^{-12}, \ast\ast\ast P < 1.0 \times 10^{-6}$). The number of disease associations for a gene are represented by the colour. From Gustafsson and Gawel et al., A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. Sci Transl Med 7, 313ra178 (2015). Reprinted with permission from AAAS.

work (TF-GRN). We hypothesized that genes with methylated upstream regulatory regions are unlikely to be regulated by a TF, and therefore we have also incorporated DNA methylation profiling.

Hub-proteins are assumed to be key players in dynamic networks (networks were network topology changes over time), and are found to often perform many functions [104]. Such multi-functionality might be seen as a gene product containing many domains with separate functions and/or perform tissue-specific functions [104]. Therefore, impairing such a gene may affect an abundance of other genes activity in one and/or in multiple different tissues and cell types. Interestingly, we found that all TFs associated with a disease via at least one SNP (GWAS TFs) had on average 3.7 times more targets than non-GWAS TFs (TFs with no mapped disease SNPs; Fig.7A).

Consequently, we hypothesized that ‘master’ TF regulators of T helper cell differentiation would be those TFs that regulate many downstream genes (so called hub-TFs). Indeed, we found that among the top 11 TFs
with the highest number of TF-GRN–predicted downstream targets, 10 were GWAS TFs. Furthermore, 91% of all the nominally disease-associated SNPs (genome-wide association of $P < 10^{-5}$) in all of the 10 GWAS hub-TFs were mapped to GATA3, MAF and MYB ($OR = 9.22, P = 6.6 \times 10^{-3}$; Fig. 7B). Since these hub-TFs that were most enriched for GWAS SNPs, and therefore more likely to have causal disease roles, we focused our further studies on these three TFs.

We have validated the TF-GRN–predicted TF-target interactions with ChIP-Seq experiment of GATA3, MAF and MYB. Using this technology, we analysed to which sites the three TFs bound in differentiated human Th1 and Th2 cells.

For all the three TFs, the ChIP-Seq identified binding sites were significantly enriched for GRN predicted targets ($OR_{GATA3}=7.30, P_{GATA3} < 10^{-81}$, $OR_{MAF} =1.92, P_{MAF} < 10^{-22}$, $OR_{MYB} =3.17, P_{MYB} < 10^{-27}$; Fig. 8).

Moreover, siRNA-mediated knockdowns of GATA3 and MAF in Th2 cells further supported the TF-GRN TF binding sites prediction accuracy as the knockdown affected genes showed significant enrichment of
TF-GRN–predicted early targets compared to all genes ($OR_{GATA3}=2.22$, $P_{GATA3}=2.9\times10^{-3}$; $OR_{MAF}=3.45$, $P_{MAF}=4.5\times10^{-3}$), as well as to publicly available predicted TF binding sites only.

In order to test the clinical relevance of *GATA3*, *MAF* and *MYB* in T cell associated diseases we first performed pathway analyses of TF-GRN–predicted targets of all three TFs showing enrichment for several disease-relevant pathways including cell activation and differentiation pathways, as well as diabetes, viral response, and cancer related pathways.

Here in this project we have not only identified possible disease drivers, but also three possible explanations why the shared disease genes may cause different diseases: A) shared master regulators might be deregulated at diverse levels and in different directions, or combinations, in multiple diseases; B) different SNPs at different positions might affect same gene in diverse ways; C) differential gene splicing.

The above results supported that we had identified three putative major, shared disease drivers. However, further analyses also supported two possible explanations for why the same genes can cause different diseases: 1) differential expression of the same genes; 2) SNPs might affect alternative splicing.

In support of the first explanation, we found that the same three TFs were differentially expressed in six out of eight diseases (publicly available
### Figure 9: 

*GATA3, MAF* and *MYB* transcript and predicted targets differential expression in multiple diseases. 

**Upper** Bars mark the difference in the average $P$-values TF-GRN–predicted targets of *GATA3, MAF* and *MYB* compared to all genes (*$P < 0.05$, **$P < 0.01$, ***$P < 0.0001$ from permutation test*) showing that TF-GRN–predicted targets of the three TFs are generally more affected by disease than non-target genes (have lower $P$-values; patients compared with healthy donors) in nine T cell–related diseases: hyper eosinophilic syndrome (HES), adult T cell leukemia/lymphoma (ATL), acute myeloid leukemia (AML), systemic lupus erythematosus (SLE), multiple sclerosis (MS), seasonal allergic rhinitis (SAR), influenza (IZ), breast cancer (BC), and tuberculosis (TB); 

**Lower** log2 fold changes $[\log2(FC)]$ of the expression of each TF in patients compared to controls are represented by the arrows. Up-facing and down-facing arrows depict $\log2(FC) > 0$ and $\log2(FC) < 0$, respectively. *From Gustafsson and Gawel et al., A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. Sci Transl Med 7, 313ra178 (2015). Reprinted with permission from AAAS.*

---

We also found that most of the TF-GRN–predicted targets of all the three TFs (*GATA3, MAF* and *MYB*) were differentially expressed in all eight diseases (Fig.9). We have made similar observation in three additional original expression profiling studies of CD4+ T cells in breast cancer, tuberculosis and influenza (Fig.9; ad. A).

This might suggest that even though a TF is not differentially expressed itself its function might be altered so that its target genes exceed differential expression. That might be a result of protein misfolding, altered amino acid content, or by differential splicing. All of those errors might be due

---

4Standard microarrays are not designed to distinguish between splice variants and therefore a gene might seem as not differentially expressed when in fact expression profiles of its splice variants are shifted.
to the presence of SNPs. As we have shown GATA3, MAF and MYB are highly enriched for multiple disease associated SNPs. Thus, different SNPs associated with the same gene may lead to different disease phenotypes, through different mutations [107]; (ad. B).

In fact, there is a possibility that some of the SNPs in the nearest surrounding of the three TFs may have an impact on the transcriptional regulation and/or post-transcriptional regulation of those TFs. Indeed, we have found several regulatory SNPs (based on rSNPBase – a database with SNP regulatory annotations; [123]) in linkage disequilibrium with GWAS SNPs. Regulatory SNPs (rSNPs) are located in regulatory regions like proximal and distal transcriptional regulation sites inside elements associated with DNA accessibility, in the RNA binding proteins-associated regions, or SNPs affecting microRNA-mRNA interference. Indeed, we have found 14 disease associated SNPs with a predicted regulatory effect and 252 disease SNPs in LD block with rSNPs (106 for GATA3, 61 for MAF and 85 for MYB). We have therefore analysed all known SNPs (not only disease associated) downloaded from the UCSC database [124] and found many in the splice regulatory motif regions to be in linkage disequilibrium (LD) with GWAS SNPs mapped to GATA3, MAF or MYB. Furthermore, in order to test possible regulatory role of those SNPs, for all three TFs we have predicted splice regulatory motif regions, using a tool called RegRNA 2.0 [93]. Splice regulatory motifs are exonic and intronic splicing enhancer and silencers, splicing sites or open reading frames. This led us to identify 29 SNPs for GATA3; 6 SNPs for MAF; and 46 for MYB that might affect splicing events. Since MYB has many splice-variants, and for MAF there was no RNA-Seq data available, for exon eQTL analyses we have prioritized GATA3. GATA3 exists in 2 splicing variants that differ only in exon number 4, which makes it an interesting example to study SNP effects on exon inclusion (see Materials and Methods). We have identified one such example where SNP was affecting the splicing of GATA3 (Fig.10).

Therefore, we proceeded to test the three TFs splice variants potential upstream role during pre-symptomatic stages of diseases where we have used two relapsing diseases as models. In summary, our analyses revealed that splice variants of those key three TFs were miss-expressed in remission (symptom–free phase of disease) whereas their predicted targets were deregulated in relapse (active disease). Thus, differential splicing is a specific example of why the same gene might be involved in the pathogenesis
of multiple diseases (ad. C).

Ideally, we would validate importance of GATA3, MAF and MYB in early disease pathogenesis by studying initially healthy subjects that later on develop disease. Since patients generally present to health care when they are symptomatic, one would need to conduct a study following thousands of healthy individuals over decades. Such study would be demanding both cost- and time-wise. Instead, here we hypothesized that symptom free phases of relapsing diseases could serve as models of pre-symptomatic disease stages. Here, we have studied Seasonal Allergic Rhinitis (SAR) and Multiple Sclerosis (MS) as models of such stages.

Analyses of 10 un-stimulated CD4+ T cell from MS patients in symptom-free phase (during remission) and 10 healthy controls showed significant downregulation of GATA3 splice variant 2, in patients compared to healthy controls ($P = 0.019$). Those results encouraged us to collect samples in a prospective gene expression study of MS patients seen both during relapse (active disease) and remission (symptom-free). Our analyses not only confirmed significant decrease of GATA3 during remission compared to con-

Figure 10: Expression quantitative trait loci (eQTL) analysis of the exons of GATA3. In the plot there are presented normalized RNA sequencing (RNA-Seq) counts across all six GATA3 exons. Each exon counts are further divided by the genotypes of the variant rs501764. Asterisks denotes significant exon differential expression across these genotypes (** $P < 0.01$, *** $P < 0.001$). From Gustafsson and Gawel et al., A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. Sci Transl Med 7, 313ra178 (2015). Reprinted with permission from AAAS.
controls, but also revealed a decrease in remission compared to relapse. Moreover, we have noted significant differential expression of GATA3, MAF and MYB, TF-GRN–predicted targets during active disease (relapse) compared to controls ($P_{GATA3} < 10^{-12}; P_{MAF} = 5.3 \times 10^{-3}, P_{MYB} = 1.6 \times 10^{-4}$).

Further analyses of the MS consortium GWAS study, which included ~25,000 patients and controls showed significantly MS-associated SNPs that could be linked to all three TFs ($P_{MAF} = 1.9 \times 10^{-5}, P_{MYB} = 5.8 \times 10^{-6}, \text{and } P_{GATA3} = 6.6 \times 10^{-5}$). Intriguingly, all three TFs were among top 1% most enriched genes for MS-associated SNPs ($P = 3 \times 10^{-6}$).

Another disease with symptom-free stages is SAR. Here, CD4+ T cells isolated from patients outside of pollen season were stimulated with allergen in vitro. We have hypothesized that unstimulated cells would represent symptom-free disease stages whereas allergen-stimulated cells would serve as a proximity to active disease stage.

For that purpose we have collected samples from 10 SAR patients and 10 healthy controls. We have profiled splice variant expressions with exon arrays. Differential expression analyses showed that splice variants of all three TFs were miss-expressed in the symptom-free disease stage ($P < 0.05$; unstimulated cells) whereas their TF-GRN–predicted targets were DE in active disease ($P < 0.0001$; allergen challenged cells) when compared with healthy controls. Further strengthening their putative role as possible key disease drivers we have found that during asymptomatic stage all measured MYB splice variants, MAF splice variant2 (NM_001031804.2) and GATA3 splice variant 1 (NM_001002295.1) were differentially expressed (Fig.11A).

A combination of the average expression levels of the three TFs splice variants separated patients from controls with high accuracy (AUC= 0.83, $P = 2.9 \times 10^{-3}$, Fig.11B). Moreover same score correlated significantly with symptom severity in patients (PCC=0.67, $P = 0.019$; Fig.11B).

To further validate our findings we have replicated the results in an independent material consisting of 14 patients and 6 healthy controls, where the above mentioned score again separated patients and controls with high accuracy (AUC=0.77, $P = 0.039$). This suggests the potential of the three TFs as for disease prediction and disease severity markers.

Furthermore we have asked whether we could functionally analyse the effects of splice-variants. For that reason we have performed siRNA-mediated gene silencing experiment followed by Th2 polarization and exon microarray splice variant expression profiling. Since GATA3 splice variants differ in
only one nucleotide of one exon and since MAF has multiple splice variants here we have concentrated on MAF for which we were able to bioinformatically predict and experimentally verify siRNA targeting one specific MAF splice variant – namely splice variant 2 (differentially expressed in the asymptomatic stage of SAR). In this experiment knockdown efficiency of MAF splice variant 2 was 38.5%, MAF splice variant 1 was 1.7%. We have found that 63% (65 out of 103) of TF-GRN–predicted MAF targets were affected by the knockdown (Fig.12). Moreover those genes also had lower $P$-values in the symptomatic SAR stage ($P<10^{-5}$).

MAF splice variant 2 predicted targets (TF-GRN–predicted targets affected by siRNA knockdown) were enriched for cell programed death, intracellular signal transduction, regulation of protein metabolic process, and response to growth factor (Gene Ontology pathways enrichment analyses). This further strengthens the possible importance of MAF in disease patho-
Correlation -log(P-value) between $MAF$ splice variant 1 and other genes

Figure 12: $MAF$ splice variant specific targets. Stars represent TF-GRN–predicted $MAF$ targets. $MAF$ splice variant specific targets were identified using Pearson’s correlation analyses between $MAF$ splice variant 1 and 2 and TF-GRN–predicted targets in $MAF$ siRNA mediated knockdown experiment. From Gustafsson and Gawel et al., A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. Sci Transl Med 7, 313ra178 (2015). Reprinted with permission from AAAS.

genesis.

Another advantage of SAR as a model disease is that allergic patients can undergo treatment (immunotherapy) in order to achieve tolerance to allergen. We have followed 14 patients over 2 years of their treatment. Splice variants expression patterns were measured at 3 time points: before immunotherapy, after 1 year and after 2 years.

We found that the therapy gradually normalised the expression of the three TFs. After 2 years of treatment TF expression in patients was comparable with healthy controls (Fig.13). This finding further supports the importance of $GATA3$, $MAF$ and $MYB$ in SAR.

Generally, this project gives an insight into possible disease regulations, and reasons why the same genes may be involved in pathogenesis of multiple diseases – here shown in example of differential splicing of master TFs in two diverse diseases: seasonal allergic rhinitis and multiple sclerosis. One limitation of this study is that master regulators were identified based on the healthy TF-GRN. During disease new disease hub TFs might emerge, which we wouldn’t find in this model.

Moreover, we have tested deregulation of the three TFs in approximations of early disease stages, namely symptom-free stages of two relapsing diseases (MS and SAR). We don’t know if such approximations are relevant. One validation option would be to analyse samples from biobanks
Figure 13: Change of normalized mean splice variant expression of GATA3 splice variant 1, MAF splice variant 2, MYB splice variant 4, and MYB splice variant 5, over immunotherapy (T) before, after 1 and 2 years of treatment compared to healthy controls. From Gustafsson and Gawel et al., A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. Sci Transl Med 7, 313ra178 (2015). Reprinted with permission from AAAS.

-- biorepositories that store biological samples collected from large number of humans for use in research. However samples volume is limited, and therefore isolation and analyses of T cells in those samples are complicated. Another option is to conduct a prospective study where one should ideally analyse samples coming from individuals that do not have a disease at the time of sample collection but develop it later in life. However, this approach is very complex as we don’t know beforehand who is going to develop disease, and therefore large cohorts of initially healthy individuals would need to be followed over time. Moreover, it’s currently unknown when to expect gene expression changes before symptoms occur. That time would most probably differ depending not only on the specific disease but also interpersonally. Therefore, one would need to perform measurements relatively often which would increase study costs and complexity. In the absence of such samples, using symptom-free phases of diseases suggests a direction for further studies – here we have reduced number of possible hypothesis from hundreds to three – namely GATA3, MAF and MYB.

In this project, we have concentrated on transcription factors as early TFs were most enriched for variants identified by GWASs. Due to the nature of TFs (regulating gene expression) we have prioritized hub-TFs – meaning TFs with the highest number of targets. However, as discussed below, there have been reports that apart from hub nodes other type of
nodes in the interactome are of importance in disease.

Most biological networks have skewed node degree distribution (often referred to as scale-free networks) [2]—meaning that they are dominated by a few key hub nodes. In this type of networks hub nodes in the networks have been proven to be crucial for maintaining network integrity as discussed above. Due to the fact that hub nodes have a very high number of neighbours [2]. That in fact might be an advantage in case of drug design where targeting one gene with drug many have a therapeutic effect on multiple downstream genes. However, an attack on the hub nodes can be extremely destructive. Interestingly, viruses evolved to target hub nodes [125]. As much as hub nodes are of considerable importance in scale-free networks there are other network features that may imply node importance.

For example, betweenness centrality is a measure that mirrors how many shortest paths between other nodes lead through a given node. Nodes with high betweenness score can be seen as the main connectors within network topology [104]. The higher betweenness centrality the bigger destructions within network would occur upon node removal, and therefore more likely to result in a disease phenotype [24,126,127,128].

On the other hand, disconnectivity reflects how important a node/edge or a set of nodes/edges are for maintaining the communication between pairs of nodes in the network [128]. This measure allows for assessment of topological redundancy.

Having all of those, and other, options in mind, one could search for key, putatively disease driving nodes (genes) in the network in multiple ways.

Apart from the variety of node types from the network topology perspective there are other gene regulators in cells that could be taken into account—like microRNAs that may have a key role in disease pathogenesis.

Project III: microRNAs in T helper cell differentiation and in T cell related diseases

MicroRNAs were previously shown to be associated with multiple diseases [63, 129,130,131,132], even though individual microRNAs weakly repress their targets [67]. Thus, their impact on most of their targets often results in no phenotypic effect [67]. One hypothesis is that the role of microRNAs lies in system stabilization, and that the repressive effects of microRNAs on GRN stability is cumulative [67]. This hypothesis explains the wide spread of microRNA targets over multiple mRNAs without substantial target expres-
Figure 14: Average expression profiles of microRNA processing machinery over time in CD4+ T cells Naïve T cells differentiation towards mature Th1 and Th2 cells.

...sion change (negative feedback loops and wide spread of targets are more efficient in maintaining system stability [67]). Moreover, testing one single microRNA hypothesis assumes that there is no redundancy in the system for phenotype control [67]. Therefore, we have proceed by analysing groups of microRNAs predicted to regulate shared group of target genes.

In this project, we have again focused on T helper cell differentiation process as the shared point of multiple diseases. Since Th1 versus Th2 differentially early expressed genes show the highest enrichment of GWAS genes among other cell types and comparisons (apart from early TFs discussed above; Fig.6), we again focused on differentiation of those two cell types.

Analyses of the previous time series mRNA profiling of differentiating Th1 and Th2 cells revealed early expression level change of microRNA processing machinery suggesting an important role of microRNAs in that process and therefore possible importance in disease progression. Both DROSHA and DGCR8 – core components of microRNA processing machinery together with parts of RISC complex: AGO2, DICER and TARBP2, and microRNA transport machinery (XPO5) showed expression changes (Fig.14).

Encouraged by those findings we proceeded to profile expression of micro-
RNAs in the same setting, and in the same biological material used for analysing mRNA expression. Namely, we have polarized Na"ıve CD4 + T cells toward Th1 and Th2 cells in vitro. We have measured 800 microRNAs expression profiles at six time points (0hr, 6hr, 24hr, 3 days, 6days and 8 days). We have found 256 microRNAs to be expressed (detected in more than 50% of the samples) during differentiation.

Again, using the same principals as for TFs, we constructed a microRNA–Gene Regulatory Network (mGRN). mGRN was constructed based on the microRNA–gene target predictions and anti-correlation of microRNAs and mRNAs in time series of maturing Th1 and Th2 cells from the Na"ıve state over all 44 samples. The mGRN consisted of 174,592 edges (interactions) out of which 1,182 were microRNA-mRNA interactions.

Existing tools for microRNA–target predictions are not robust, mainly due to the high complexity of the microRNA–target regulation process [61] and by knowledge incompleteness.

In this project, we incorporated predictions made with two different algorithms, namely miRanda [100] and TargetScan [99]. miRanda has been shown to be very sensitive in comparison to other tools at the price of precision [61]. TargetScan on the other hand has higher precision [61]. Here sensitivity was defined in two ways: a) as percentage of experimentally supported microRNA predicted targets and b) as a proportion of correctly predicted microRNA targets to total correct microRNA-mRNA interactions (two different studies). Precision on the other hand was defined as proportion of correctly predicted microRNA targets to total predicted microRNA–mRNA interactions [61].

There are several differences between those tools. The most significant difference is that TargetScan assumes perfect complementarity in the seed region, whereas miRanda allows for mismatches. That may to some extent explain the low sensitivity of TargetScan since, in animals, strict complementarity between target site within mRNA sequence and microRNA is rare [61]. Furthermore, both miRanda and TargetScan are generally based on microRNA sequence complementarity in the 3’UTR (untranslated region) of mRNA. However microRNA target sites can be also found in 5’UTR and ORFs [60,61].

Since a group of microRNAs may regulate a shared group of genes and pathways we have then created microRNA–mRNA modules (groups of microRNAs that are predicted to regulate connected group of mRNAs)
using mirsynergy tool [101]. We have identified 25 such modules. In contrast to our primary hypothesis that microRNAs might have upstream role in Th differentiation the majority of the modules had late time-point specificity, especially at the microRNA level (Fig. 15).

To get an idea of the role of the mRNAs in the created modules we have performed pathway-enrichment analyses. We have found that modules 2, 8, 9 and 22 were most enriched ($P < 0.05$) for GO terms. Those terms were related to glycosylation (2), metabolism (8) and Wnt signalling (9 & 22), respectively.

In order to investigate modules associations with diseases we analysed 11 CD4+Tcell associated diseases from public repositories namely: acute myeloid leukaemia (AML), human immunodeficiency virus infection (HIV), multiple sclerosis (MS), hypereosinophillic syndrome (HES), Sézary syndrome (SeS), adult T cell leukaemia/lymphoma (ATL), type 1 diabetes (T1D), chronic lymphoid leukaemia (CLL), seasonal allergic rhinitis (SAR), rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE). All modules were differentially expressed in at least one disease ($P < 10^{-6}$). Strikingly one module (no 8) was DE in all 15 studied datasets (Fig. 16). Surprisingly some diseases, including AML, HIV, MS, HES and ATL, show a degree of similarity at the mRNA level as the same modules are differentially expressed in all those diseases (modules 1-8,10,14,17,21-23,25). Whereas other diseases like T1D, CLL, SAR and HIV at the microRNA level were highly associated ($P < 10^{-6}$) with only module 8.

Figure 15: Normalized specialization of microRNAs and mRNAs in modules presented in heatmaps was calculated using modified Shannon’s entropy score.
Figure 16: Heatmap of microRNA-mRNA modules associations with 11 diseases: Sézary syndrome (SeS), adult T cell leukemia (ATL), human immunodeficiency virus (HIV), rheumatoid arthritis (RA), chronic lymphocytic leukemia (CLL), type-1 diabetes (T1D), hypercosinophilic syndrome (HES), multiple sclerosis (MS), acute myeloid leukemia (AML), systemic lupus erythematosus (SLE), seasonal allergic rhinitis (SAR).

In further support of the importance of module 8 we found that it contains 10 microRNAs with known disease association to T cell related diseases like Sézary syndrome [129], multiple sclerosis [133], myasthenia gravis [134] and many cancers [135,136].

Moreover, we have performed GWAS gene enrichment analyses of the modules finding that number of enriched disease terms ranged from 1 for modules 15 and 20, to 160 for module 8. Notably module 8 was enriched for T cell related diseases like SLE, Diabetes, HIV, and terms related to allergy: ‘Allergic Rhinitis’; ‘Atopy’; ‘Asthma and Hay fever’; ‘Eosinophilic Esophagitis’ and ‘Self-reported Allergy’ ($P < 1x10^{-5}$). Generally for 12 out of 25 modules (including module 8) we have noted significant overlap between disease association based on both expression and genetic variants (GWAS; $P<0.05$).

Since two independent analyses suggested involvement of module 8 in allergy we have proceed to validate mGRN predicted edges in a prospective study of CD4+ T cells isolated from 10 healthy controls, and 10 SAR patients (samples collected out-side of the pollen season) stimulated with allergen or diluent (control) in vitro for 17h. This time point was chosen in order to capture early microRNA–mRNA dynamics. Both microRNAs and mRNAs were profiled with microarrays. MicroRNA–mRNA mGRN–
predicted interactions within the module 8 were present in the new material as they were higher anti-correlated than expected by chance over all samples (OR=1.19; \( P=2.4\times10^{-4} \)). Moreover predicted interactions were present in both healthy controls and patients (OR\textsubscript{control}=1.11; \( P=1.0\times10^{-11} \); OR\textsubscript{patient}=1.17; \( P=1.3\times10^{-23} \)).

Moreover, we have found that module 8 was significantly differentially expressed at both microRNA and mRNA level comparing patients and controls during an active disease (CD4+ T cells stimulated with allergen).

In summary, we have found strong microRNA–disease associations, however in contrast to primary hypothesis we haven’t found an evidence for the upstream disease role of microRNAs. However current knowledge about microRNAs contains many unknowns and uncertainties. This means that there is possibility that other than widely accepted negative miR–gene regulation mechanisms might occur in early T cell differentiation. Which means that further studies are warranted before one can reject upstream disease miR hypothesis.

For example, microRNAs were shown to enhance translation and transcription [60], which has not been taken into account in this study. In [137] authors presented an example of miR-363-3 that enhances gene expression by binding to AU–rich regions. However, this regulation is context–dependent as it occurs in the cell-cycle arrested cells but not in proliferative cells. On the other hand, in [138] it has been shown by the example of miR-373 that by binding to complementary sites within gene’s promoter regions, gene expression is promoted rather than repressed. However, not all genes with complementary sites in promoter regions are in fact positively regulated suggesting that this regulation is also context- dependent [138].

Another aspect of putative microRNA role in disease pathogenesis is their possible role in cell to cell communication. Extracellular microRNAs can be extracted from most biological fluids [60]. Although their function is currently largely unknown, there is growing evidence that ex-miRs are involved in disease pathogenesis. For example ex-miRs profiles has been shown to be predictive of infection [60]. Furthermore, there are reports showing that ex-miRs can be passed between leukocytes [139,140,141]; suggesting their role in cell to cell communication. Further strengthened by reports in [142] where authors show that the ex-miRs produced in Treg cells are causing repression of Th1 cell proliferation in mouse [142]. This gives another hint of the putative upstream role of microRNAs in disease,
since immunological studies have shown that many immune-related diseases are not only a result of gene mutation or differential expression but also are characterized by elevated number of T cells and disproportion of T cell subsets [10].

Moreover, microRNAs might also regulate gene alternative splicing. It has been shown that knockdown and overexpression of microRNA processing machinery (AGO1, AGO2, and DICER1) leads to impairment in exon inclusion [143,144]. Targeting an intronic region of the FN1 pre-mRNA with a siRNA resulted in local chromatin remodelling and altered splicing of the adjacent exon [143]. Moreover, siRNAs targeting the pre-mRNAs of SMN2 and DMD were capable of inducing either exon inclusion or exon skipping at therapeutically relevant exons [145]. In [143] authors have found evidence that siRNA targeting intragenic regions may affect alternative splicing though elongation inhibition. However, more studies are required in order to confirm or refute this hypothesis.

Here we have focused on the role of single molecular measure (mRNA and microRNA) to identify disease ‘sharedness’. However to fully understand disease pathogenesis it would require comprehensive analyses of multiple sources of evidence, like not only gene-gene or microRNA–gene interactions but also epigenetic modifications, regulations of posttranscriptional modifications, shared environmental factors, affected tissues, cell type proportions and interactions, organs etc. [146,147,148].
Concluding Remarks

This thesis aimed to identify shared disease genes and shared disease regulators as well as identification of potential mechanisms that may explain how the same genes may contribute to pathogenesis of multiple diseases.

I hold that the direction of studying disease similarities has many advantages, including understanding of disease mechanisms and inter-individual variability in both disease development and response to treatment. Moreover, studies of shared disease mechanisms may reveal opportunities for drug repositioning.

As an example, using principles from [149] in an ongoing project, we have identified a drug targeting genes proximal to the shared disease gene module. Even though this drug is primary used in treatment of hyperlipidaemia, we have successfully repurposed it to treat rheumatoid arthritis in mouse (data not shown).

However, one shouldn’t underestimate the power of disease specific genes. For example, in some diseases the shared disease genes may be the core of the disease but the disease-specific genes are the disease drivers directing disease pathogenesis towards specific disease.

In summary, this work can be seen as another step towards understanding shared disease mechanisms. However, many more extensive studies are required before these mechanisms can be exploited for diagnostic and therapeutic purposes.
Bibliography


[146] B. A. Maron, Y. Y. Zhang, K. White, S. Y. Chan, D. E. Handy, C. E. Mahoney, J. Loscalzo, and J. A. Leopold. Aldosterone inactivates the


Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-147047