UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2083

Uncovering biomarkers and molecular heterogeneity of complex diseases

Utilizing the power of Data Science

SARA YOUNES

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2021

ISSN 1651-6214
ISBN 978-91-513-1310-8
URN urn:nbn:se:uu:diva-454997

Dissertation presented at Uppsala University to be publicly examined in A1:111a, Biomedical Centrum (BMC), Husargatan 3, Uppsala, Friday, 26 November 2021 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Aedin Culhane (University of Limerick, Ireland).

Online defence: https://uu-se.zoom.us/j/61211793326

Abstract

Younes, S. 2021. Uncovering biomarkers and molecular heterogeneity of complex diseases. Utilizing the power of Data Science. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2083. 71 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1310-8.

Uncovering causal drivers of complex diseases is yet a difficult challenge. Unlike single-gene disorders complex diseases are heterogeneous and are caused by a combination of genetic, environmental, and lifestyle factors which complicates the identification of patient subgroups and the disease causal drivers. In order to study the dimensions of complex diseases analyzing different omics data is a necessity.

The main goal of this thesis is to provide computational approaches for analyzing omics data of two complex diseases; mainly, Acute Myeloid Leukaemia (AML) and Systemic Lupus Erythematosus (SLE). Additionally, we aim at providing a method that would deal with integration issues that usually arise when combining complex diseases omics (specifically metabolomics) data from multiple data sources.

AML is a cancer of the myeloid blood cells that is known for its heterogeneity. Patients usually respond to treatment and achieve a complete remission state. However, a majority of patients relapse or develop treatment resistance. In paper I, we focus on investigating recurrent genomic alterations in adult and pediatric relapsed and primary resistant AML that may explain disease progression. In paper II, we characterize changes in the transcriptome of AML over the course of the disease, incorporating machine learning analysis.

SLE is a heterogeneous autoimmune disease characterized by unpredictable periods of flares. The flares are presented as different SLE disease activities (DA). Studies on the combinatorial effects of genes towards the manifestation of SLE DAs in patients' subgroups have been limited. In paper III, we analyze gene expression data of pediatric SLE using interpretable machine learning. The aim was to study the co-predictive transcriptomic factors driving disease progression, discover the disease subtypes, and explore the relationship between transcriptomics factors and the phenotypes associated with the discovered subtypes.

Recently, Metabolomics has been a crucial dimension in major multi-omics complex disease studies. Small-compound databases contain a large amount of information for metabolites. However, the existing redundancy of information in the databases leads to major standardization issues. In paper IV, we aim at resolving the inconsistencies that exist when linking and combining metabolomics data from several databases by introducing the new R package MetaFetcheR.

Keywords: Complex Disease, Cancer, Autoimmune diseases, Acute Myeloid Leukemia, Systemic Lupus Erythematosus, Bioinformatics, Machine Learning, Data Science, Statistical Analysis

Sara Younes, Department of Cell and Molecular Biology, Box 596, Uppsala University, SE-75124 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-1310-8
URN urn:nbn:se:uu:diva-454997 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-454997)

Cover Art

The cover image is inspired by rule-based networks that I used in my thesis. One can tell from the image that similar colors cluster together. This is one of the useful advantages of rule-based networks that I utilized to discover Systemic Lupus Erythematosus disease subtypes. The original image is a picture that I took of assorted chocolates that I personally organized then used a filter on the original image for artistic transformation. The Artistic image was generated by www.instapainting.com.

To myself
for hanging on

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Stratmann, S., Yones, S.A., Mayrhofer, M., Norgren, N., Skaftason, A., Sun, J., Smolinska, K., Komorowski J., Krogh Herlin M., Sundström C., Eriksson A., Höglund M., Palle J., Abrahamsson J., Jahnukainen K., Cheng Munthe-Kaas M., Zeller B., Pokrovskaja Tamm K., Cavelier L., Holmfeldt L. (2020) Genomic characterization of adult and pediatric relapsed acute myeloid leukemia reveals novel therapeutic targets. ***Blood Advances***, 5(3):900–912

II Stratmann S*****, Yones SA*****, Garbulowski M, Sun J, Skaftason A, Mayrhofer M, Norgren N, Herlin MK, Sundström C, Eriksson A, Höglund M, Palle J, Abrahamsson J, Jahnukainen K, Munthe-Kaas MC, Zeller B, Tamm KP, Cavelier L, Komorowski J, and Holmfeldt L. Transcriptomic analysis reveals pro-inflammatory signatures associated with acute myeloid leukemia progression. *Accepted manuscript in* ***Blood Advances,*** *In press, (2021),* https://doi.org/10.1182/bloodadvances.2021004962. *** Equal contribution**

III Yones SA, Annett A, Stoll P, Diamanti K, Holmfeldt L, Barrenäs CF, Meadows, JRS*****, and Komorowski J*****. Identification of combinatorial markers in pediatric Systemic Lupus Erythematosus and disease subtypes from gene expression data using interpretable machine learning. *Under Revision*. *** Equal contribution**

IV Yones SA, Csombordi R, Komorowski J, and Diamanti K. MetaFetcheR: An R package for complete mapping of small compound data. *Under Revision.* bioRxiv. Published online March 1, 2021. doi:10.1101/2021.02.28.433248

List of Additional Publications

I Herlin MK, Yones SA, Kjeldsen E, Holmfeldt L and Hasle H. What Is Abnormal in Normal Karyotype Acute Myeloid Leukemia in Children? Analysis of the Mutational Landscape and Prognosis of the TARGET-AML Cohort. ***Genes***. (2021); 12(6):792. https://doi.org/10.3390/genes12060792

II Garbulowski M, Smolińska K, Çabuk U, Yones SA, Celli L, Yaz E, Barrenäs F, Diamanti K,Wadelius C and Komorowski J. Machine learning-based analysis of glioma grades reveals co-enrichment. *Manuscript.*

Contents

Abbreviations

α	Alpha
∅	Empty set
ε	Epsilon
<	Less than
≤	Less than or equal to
μ	Mean
<<	Much less than
≠	Not equal
σ	Standard error
⊆	Subset
∑	Summation
AD	Autoimmune Disease
AML	Acute Myeloid Leukemia
ANN	Artificial Neural Network
Anti-dsDNA	Anti-double stranded DNA
AUC	Area Under the Curve
BILAG	British Isles Lupus Assessment Group
BM	Bone Marrow
cDNA	Complementary DNA
CGI	CpG Islands
ChEBI	Chemical entities of biological interest
CNV	Copy Number Variation
CpG	Cytosine–guanine dinucleotides
CR	Complete Remission
CTS	Chemical Translation Service
DAMP	Damage-associated molecular patterns
DGE	Differential Gene Expression
DNA	Deoxyribonucleic Acid
DNAm	DNA methylation
ES	Enrichment Score
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
FPKM	Fragments per kilobase of exon per million mapped fragments
FPR	False Positive Rate
GO	Gene Ontology

GSEA	Gene Set Enrichment Analysis
H0	Null Hypothesis
H1	Alternative Hypothesis
HC	Hydroxychloroquine
HMDB	Human Metabolome Database
HSCT	Hematopoietic stem cell transplant
KEGG	The Kyoto encyclopedia of genes and genomes
KM	Kaplan-Meier
LAI	Lupus Activity Index
LHS	Left Hand Side
LIPID MAPS	Lipid omics gateway
LOWESS	Locally Weighted Scatterplot Smoothing
MCFS	Monte Carlo Feature Selection
MDS	Myelodysplastic syndromes
ML	Machine Learning
MS	Mass Spectrometry
ncRNAs	Non-coding ribonucleic acid
NGS	Next Generation Sequencing
NLR	Neutrophil-lymphocyte ratio
NMR	Nuclear magnetic resonance
NP	Non-deterministic Polynomial
NUSE	Normalized Unscaled Standard Error
PC	Principal component analysis
PCA	Principal component analysis
PCR	Polymerase chain reaction
PGA	Physician's global assessment
PR	Primary Resistant
pSLE	Pediatric Systemic Lupus Erythematosus
PubChem	Database of U.S. national center for biotechnology information
QC	Quality Control
R/PR	Relapse/Primary Resistant
RHS	Right Hand Side
RI	Relative Importance
RLE	Relative Log Expression
RNA	Ribonucleic acid
RNA-Seq	RNA-Sequencing
RNADeg	RNA Degradation
ROC	Receiver Operating Curve
RPKM	Reads per kilobase of exon per million mapped fragments
SFI	SLE Flare Index
SLAM	SLE Activity Measure
SLE	Systemic Lupus Erythematosus
SLEDAI	SLE Disease Activity Index

SNV	Single Nucleotide Variant
SV	Structural Variant
SVA	Surrogate variable analysis
SVM	Support Vector Machine
T2D	Type 2 Diabetes
TARGET	Therapeutically Applicable Research to Generate Effective Treatments
TCGA	The Cancer Genome Atlas
TMM	Trimmed Mean of the M-values
TN	True Negative
TP	True Positive
TPR	True Positive Rate
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Introduction

Complex diseases

Complex diseases are largely caused by an interplay between genetic, environmental, and lifestyle factors most of which have not been identified yet. Some examples of complex diseases include Alzheimer's disease, multiple sclerosis, cancer and autoimmune disease. Although some genes associated with these diseases are inherited, genetic factors represent only part of the risk associated with complex disease phenotypes[1]. Therefore, studying all possible factors and their interactions can potentially improve understanding of the causes of complex disease and assist in developing targeted therapies. Features of complex diseases complicate the detection of their driving factors, as some drivers commonly can be confounded by others that might be shadowing their contribution to the disease manifestation. Furthermore, there are multiple interchangeable genetic and environmental factors.

Previously, studies have been conducted on either the genetic determinants or the environmental factors underlying complex disease. Today, with the vast availability of sequencing data and tools such as bioinformatics, statistical and machine learning tools the trend is shifting towards cross-disciplinary studies and collecting data on all aspects of various diseases such as genomics, transcriptomics, epigenomics and metabolomics. These various types of data collectively are called omics data and the cross-disciplinary studies are called multi-omics.

In this work we focus on analyzing different types of omics data for two types of complex diseases which belong under cancer and autoimmunity umbrellas, respectively. These diseases are Acute Myeloid Leukemia (AML) and Systemic Lupus Erythematosus (SLE). Moreover, we introduce a tool that we develop, which specifically deals with issues that arise with metabolomics data. The tool is aimed to aid the research community when analyzing metabolomics data for complex diseases.

Cancer as a Complex disease

Cells are the main building blocks of the human body. There are hundreds of cell types and trillions of cells that co-exist in a social setting where each cell has its own function. Cells communicate with each other sending, receiving,

and interpreting signals. If a cell starts to malfunction it is obligated to sacrifice itself for the survival of the whole system through a process called apoptosis (programmed cell death). Cancerous cells are cells that deviate from this social setting and, for instance, are able to evade apoptosis due to occurrence of special genetic alterations. The genetic alterations force the cell to attain special characteristics to be categorized as a cancer cell such as resisting cell death, sustaining proliferative signals, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality and inducing angiogenesis[2]. Cancer is a multistage progression disease where the cell has to go through several checkpoints[3]. The most common genetic alteration that would transform a normal cell into a cancer cell is point mutations. Point mutations are DNA sequence variations occurring within a single nucleotide. The point mutations that are acquired by cancer cells and enable them to gain cancer characteristics are called somatic mutations while inherited mutations are called germline mutations. Together, these mutations enable the cells to gain cancer characteristics. These characteristics are attained by either mutations that lead to loss of certain normal functions or gain of new functions. Loss or gain of function consequently enables the cell to bypass social duty. Genes that are associated to cancer by becoming activated by gain-of-function mutations are known as oncogenes. Examples of common oncogenes are *MYC* and *TERT* where somatic mutations lead to gain of function[4–6]. Tumor suppressor genes acquire loss-of-function mutations that lead to deregulation of protective mechanisms. One example of a common tumor suppressor is *TP53,* which acquires loss-of-function mutations that lead to its deactivation and consequently causes deregulation of apoptosis and cell cycle regulation [7]. There are common and special mutations for each cancer type that lead to cancer initiation. The special mutations are different for each cancer type since they involve different cell types which are different in function and use different signaling pathways. Gaining cancer characteristics is not only limited to point mutations but can also be a result of other genomic alterations such as sequence insertions or deletions (Copy Number Variation), translocations, inversions or fusions. Other factors that could cause deregulation of genes are epigenetic changes. These changes could lead to chromatin remodeling, which in turn can affect gene regulation and hence alter cell characteristics.

Cancer is a heterogeneous disease because different patients have different genetic profiles (genetic alterations, dysregulated pathways). The genetic heterogeneity has also limited developing efficient therapies for most cancer patients. Bioinformatics and data analysis tools come into play here to find significant cancer drivers in different cancer genomes (e.g., mutations affecting oncogenes or tumor suppressors) that would lead to identification of drug targets and development of effective therapies.

Acute Myeloid Leukemia (AML)

Acute myeloid leukemia (AML) is a form of hematopoietic (blood) cancer, which occurs due to transformation of myeloid progenitor cells into malignant cells. This leads to crowding of the dysfunctional blood cells in the bone marrow (BM) before entering the peripheral blood. The majority of AML patients receive chemotherapy treatment and/or Hematopoietic stem cell transplant (HSCT) and reach complete remission (CR); However, 40-60% of adults and 35% of pediatric patients relapse within two to three years [8–11] (Figure 1). Those patients usually have a poor disease outcome and acquire resistance to conventional AML therapies. In paper I and II we aim to uncover the genomic and transcriptomic landscape of relapsed AML in an attempt to understand how malignant cells in relapsed AML gain resistance to therapy and which genes and pathways can be potential new targets for treatment.

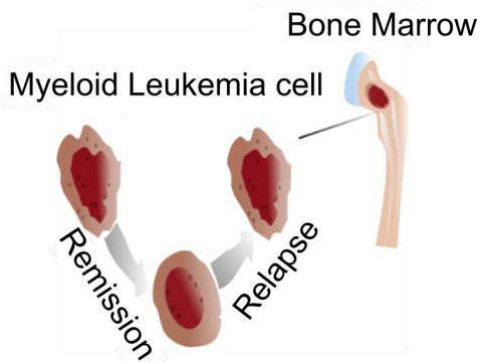

Figure 1. An illustration of the process when AML patients reach complete remission post-treatment then subsequently relapse and cells attain therapy resistance

AML classification

Previously, AML has been classified by information determined through simple clinical genetic tests and morphological examinations. However, over the decades with the technological advances that made it convenient for sequencing more genomes, a number of genetic and transcriptomic biomarkers have been discovered and currently used in clinical routine for the diagnosis and prognosis of AML. To date, more than 100 (Ref[12]) altered genes have been catalogued for AML. An average of 13 gene mutations and 1.5 gene fusion events were found per adult AML, which is a low mutational rate compared to other forms of cancer [13]. Examples of genetic biomarkers for AML include recurrent mutations in the genes *NPM1*, *CEBPA*, *RUNX1*, *FLT3*, and *TP53*, mutations in genes regulating DNA methylation (e.g., *DNMT3A*), in genes regulating histone modifications, such as *KMT2A* and *ASXL1/2*, as well as in genes associated with the *CTCF* and cohesion complex, which

regulate the three-dimensional conformation of chromatin. The cohesin complex also plays an important role during the separation of the sister chromatids during cell division. Finally, gene fusions commonly identified in AML include, for instance, *RUNX1-RUNX1T1*, *CBFB-MYH11*, *PML-RARA*, and *MLLT3-KMT2A*. It is worth mentioning that different subsets of AML patients harbor different genomic alterations, with AML being a highly heterogenous disease. There are a number of patients that do not harbor any known alterations in genes encoding epigenetic regulators, indicating independent epigenetic modifications [14]. In addition to genetic alteration biomarkers there are also transcriptomic biomarkers related to the expression of certain genes that aid in the prognosis of AML. For example, high expression of *HOX*-gene family members[15] have been associated with poor outcome and decreased therapy sensitivity in AML [16,17] and elevated expression of *BAALC*, *ERG*, *MN1*, *PRAME*, *CD34* and *WT1* have been frequently identified in AML patients with adverse outcome [18,19].

Currently, AML is classified according to the World Health Organization (WHO) classification system based on clinical features, morphological features, immunological phenotype and genetic data [20,21] into seven main categories:

- AML with recurrent genetic alterations.
- AML with myelodysplasia-related changes (poorly formed blood cells).
- Therapy-related AML and myelodysplastic syndromes (MDS).
- Myeloid sarcoma (the cancer begins in the bones and in the soft connective tissues).
- Myeloid proliferations related to Down syndrome.
- AML not otherwise specified.
- AML or MDS with germline predisposition.

De novo AML refers to AML in patients with no clinical history of prior MDS, myeloproliferative disorder, or exposure to potentially leukemogenic therapies or agents [22].

Autoimmune diseases as complex diseases

The immune system is a set of interconnected biological processes that protects the human body from external pathogens. The immune system is divided into innate and adaptive systems that are highly interdependent. The innate immune system is designed to act immediately when a pathogen is detected through a conserved pattern recognition receptor, and it results in a defensive response from the cells. The adaptive immune system consists mainly of T and B cells, which use manifold diverse receptors for detecting certain molecules called antigens. The receptors can recognize millions of foreign antigens and the adaptive immune cells form an immunologic memory

for newly recognized antigens. The immune system has mechanisms that prevents self-reactive T and B cells from uncontrolled immune self-reactivity or attacks (Figure 2). Alteration in specific genetic loci that prevents uncontrolled self-reactivity can result in autoimmune diseases (ADs).

ADs are a family of complex heterogeneous disorders with similar underlying mechanisms characterized by immune responses against self. There are more than 80 chronic illnesses characterized by immune system dysfunction that leads to loss of tolerance to self-antigens, and presence of increased level of autoantibodies, consequently resulting in chronic inflammation [23]. An example of mutations in genetic loci that result in ADs are mutations affecting the transcription factor Autoimmune regulator, leading to a relaxing selection of self-reactive T cells in the thymus (which is a gland made of immature T cells called thymocytes and that eventually releases mature T-cells) resulting in autoimmune polyendocrine syndrome [24]. Another example are the mutations affecting the *FOXP3* transcription factor in the AD IPEX, which causes aggressive autoimmunity attacks as a result of defects in the function of regulatory T cells [24]. A similar control mechanism is present in the B-cell portion of the immune system.

Regulation of T and B cells are controlled by cell-signaling events that varies among persons and among cell types, due to genetic and/or epigenetic diversity in the population. Causes of population diversity include mutations, migration, genetic drift and natural selection. Such population and personal diversity of immune systems throughout the decades have led to the complex and heterogenous nature of ADs.

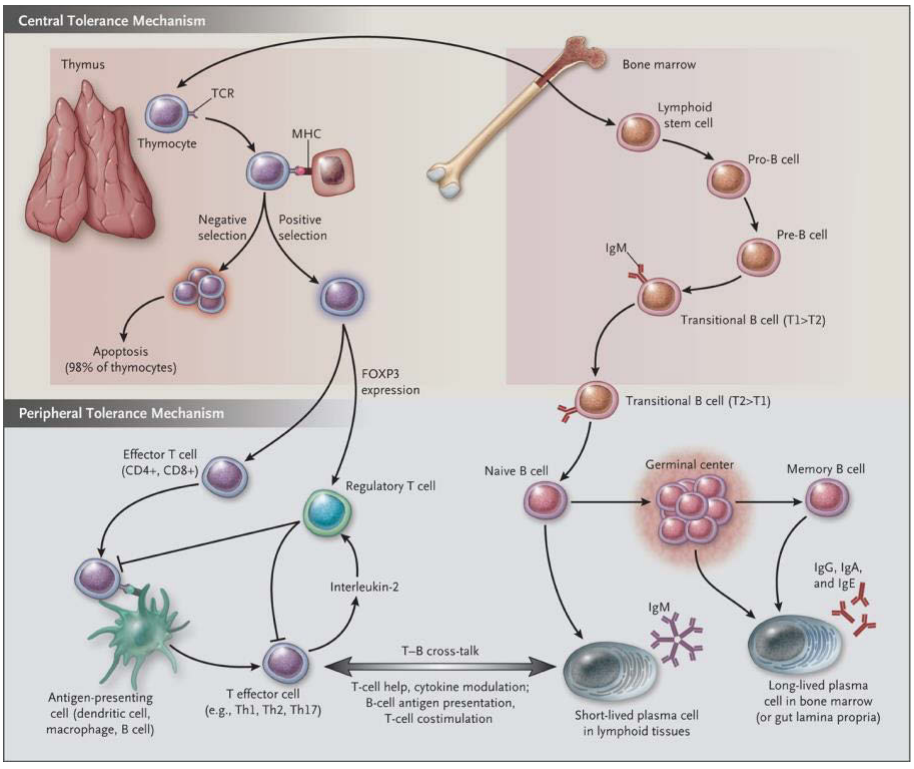

Figure 2. Central and Peripheral Tolerance mechanisms in the Adaptive Immune System. Figure adapted with permission from NEJM Publishers: Genomic Medicine., (Judy H. et al., 365:1612-1623) copyright 2011.

Systemic Lupus Erythematosus

Systemic Lupus Erythematosus (SLE) is a chronic autoimmune disease that spans a broad range of symptoms and manifestations in almost all organs and tissues (Figure 3). The production of a large number of autoantibodies is a prominent feature of SLE, which damages multiple tissues and organs. Typically, its natural history follows a relapsing-remitting course with highly variable outcome and significant morbidity [25].

With advances in medical therapy survival rates of SLE patients have improved considerably over the last decades. However, the majority of patients still experience repeated flares, which impact short- and long-term outcome [25].

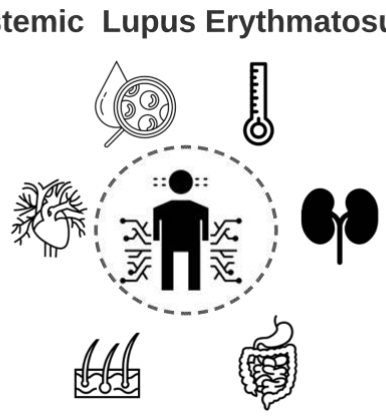

Figure 3. Illustration of the clinical and phenotypic heterogeneity in SLE

Diagnosis and classification of SLE flares

SLE is a worldwide disease but is more prevalent in some countries more than others. For example, China has a prevalence of SLE at approximately 70 cases per 100,000 people, which is high compared to other countries [26]. Moreover, SLE is more common in females than males [26]. Several SLE flare definitions have been developed based on clinical trials and are generally based on one of the following parameters [25]:

- Increase in disease activity score.
- Appearance of new or worsening of disease symptoms.
- Change in the physician's global assessment (PGA) scale towards more active/severe disease.
- Need for treatment intensification.

The commonly used SLE flare definitions are; SLE Disease Activity Index (SLEDAI), SLE Flare Index (SFI), physician's global assessment (PGA), British Isles Lupus Assessment Group (BILAG), SLE Activity Measure (SLAM) and Lupus Activity Index (LAI) [25]. The SLEDAI is an overall index that was developed in 1985 and it assess lupus activity index in the preceding 10 days [27]. It consists of 24 weighted clinical and laboratory variables of nine organ systems[27]. SLEDAI was modelled on the basis of clinicians' judgment [27]. The administrative burden with using SLEDAI is that the same physician has to keep following up on the patient history[27].

SFI is a composite measure that includes changes in the SLEDAI, individual organ manifestations not captured by the SLEDAI, changes in the treatment and the PGA, as well as the need for hospitalization due to lupus condition worsening [25].

The SFI is easy to use routinely in the clinic to classify mild to moderate disease flares and has been validated using numerous studies [27]. However, it has its limitations such as the lack of inclusion of certain drugs that have been more recently introduced in the treatment of SLE, it is not sensitive in capturing organ-specific flares and it does not differentiate between mild and moderate flares [27]. PGA reflects the physician's judgement of overall SLE disease activity [27]. BILAG is a more inclusive and sensitive-to-change scoring system [27]. It is computed based on the lupus activity in different organs separately, which is one of the main advantages of this method, since it includes an analytic description of activity from different organs and it includes varying levels of disease severity [27]. However, it is difficult to use in daily clinical practice. SLAM measures global disease activity within the previous month and includes 23 clinical manifestations in nine organs and seven laboratory features [27].

SLE flares manifestation

SLE flares often manifest in the form of non-specific symptoms (e.g., hair loss) or inherent symptoms (e.g., fever) that could be due to other causes such as fibromyalgia, drug reactions, infections and metabolic disorders including, for instance, iron deficiency [25]. A detailed history, complete physical examination, and checking laboratory results of certain features are required to exclude other causes [25].

Serological tests in the diagnosis of lupus flares

Specific serum autoantibodies and complement factors are commonly used for measuring SLE disease activity and flares. Antibodies that bind to double-stranded DNA (anti-dsDNA) are found in approximately 50% of SLE patients and their serum levels correlate with lupus activity, especially nephritis levels which affects the normal function of nephrons in kidneys[25]. Increases in anti-dsDNA (40–60%) usually precede disease aggravation by a few weeks or months, especially in the context of renal involvement[25].

Several complement proteins are associated to SLE pathogenesis and have been used to measure the disease activity for patients. Complement proteins are part of the complement system, which works with the immune system to enhance the process of identifying and fighting pathogens [28]. There are nine major complement proteins [29]. Specifically, c3 and c4 are the commonly tested ones and are associated with SLE manifestations. For example, serum c4 (but not c3) levels tend to decrease approximately two months before clinical appearance of a renal flare, reflecting the early activation of the classical complement pathway [25]. When the flare occurs, serum c3 concentrations are abnormally low suggesting that serum c3 levels may be more sensitive and specific than c4 to diagnose an SLE flare, specifically renal flares[25].

There is a subset (6–15%) of SLE patients who manifest prolonged persistent hypocomplementenemia (low levels of c3 and c4) and/or elevated anti-dsDNA antibody levels without any obvious clinical manifestations, which raised the concern of subclinical damage progression for this subset of patients[25]. In summary, from a clinical viewpoint, identifying SLE patients who are at greater risk to develop severe flares is important in designing and implementing preventive strategies[25].

SLE is currently managed with hydroxychloroquine (HC), corticosteroids, and immunosuppressive agents[30]. A human monoclonal antibody has been recently approved making it the only novel therapy that has been introduced for SLE treatment in the last 50 years[30]. The current treatments fail to target certain immune pathways for certain SLE patient subgroups confirming the molecular heterogeneity of SLE and the need for personalized treatments [30].

Omics technologies

The recent technological advances have enabled fast progression in high-throughput sequencing generating various types of omics data. These omics include quantification of expression levels of ribonucleic acids (RNA) (transcriptomics), identification of genomic alterations (genomics), quantification of non-coding RNAs (ncRNAs) and post translational modifications in histone proteins (epigenomics), as well as measurement of the abundancies of proteins (proteomics) and metabolites (metabolomics). The availability of the various types of omics imposes major challenges in identifying the interplay of multi-omics through computational approaches.

Omics technologies are usually divided into targeted and untargeted. Targeted omics include methods that measure a specific set of molecules. For example, quantification of expression levels of a pre-defined set of genes using RNA-microarrays is considered as targeted omics while capturing the whole transcriptome that can be annotated to all the set of known genes is considered untargeted. Targeted approaches may offer higher sensitivity, while the untargeted provide wider span of detectable molecules.

Genomics

Genomics appertains to analyzing DNA data to identify alterations and variations on the genomic level. Previously, methods allowed DNA sequencing of molecules one by one. These methods are currently being replaced by next-generation sequencing (NGS) technologies where billions of DNA molecules are sequenced simultaneously[31].

Determining the sequence of the entire human genome is referred to as whole-genome sequencing (WGS). Whole exome sequencing (WES) is sequencing an individual's exome (the coding sequence of the human

genome)[31]. WES is generally accomplished through capturing of the DNA containing the protein-coding regions of genes (exons) from the genome – either in solution or via an array[31]. This captured DNA is then sequenced. Since an exome represents only about 1% of the genome, it is considerably easier and less costly to sequence than a whole genome meanwhile containing greater than 85% of the disease-causing or pathogenic variants with strong effects on disease [31]. One of the main disadvantages of conventional sequencing results from the inability to prioritize individual genes among many candidates for diagnostic testing[31]. WGS approaches circumvent this difficulty because all candidate genes may be examined simultaneously [31].

Investigating the genome after being sequenced using NGS aids in the identification of many types of genetic variations that could be associated with disease phenotypes. These include bioinformatic tools identifying single nucleotide variants (SNVs), insertion/deletion mutations (indels), copy number variations (CNVs), and structural variations (SVs). The procedure in order to discover genetic variations in the sequenced genomes includes at least two elements; an aligner and a variant caller [32]. The aligner aligns the sequencing reads to a reference genome, and the variant caller assigns a genotype and identifies the positions of variants [32].

SNVs are DNA sequence variations occurring within a single nucleotide that are commonly called Single Nucleotide polymorphisms (SNP) if they are inherited and exist in at least 1% of the population[31]. The SNVs identified are then filtered to identify the ones which are pathogenic. In complex diseases such as cancer the SNVs acquired by the cells during disease progression are called somatic mutations. Variant callers usually detect indels during the variant calling process. Indels are either insertions or deletions in the genome measuring from 1 to 10 000 base pairs in length [33].

Another type of genetic variations that can be investigated are CNVs. A CNV is an event in which sections of the genome are amplified or deleted and these sections vary between individuals [34]. CNVs play an important role in human diversity and disease susceptibility, especially in complex diseases [32]. CNVs commonly also occur as somatic events in cancer cell genomes. A CNV is a type of SV of the genome that is generally defined as a region of DNA approximately 1 kb and larger in size [35]. There are various other types of SVs that could occur as a result of complex diseases such as genomic inversions and translocations.

Transcriptomics

In the central dogma RNA is a molecular intermediate between DNA and proteins, which are considered the primary functional read-out (transcription) of DNA [36]. RNA is divided into coding and non-coding. Coding RNAs are translated into proteins whereas non-coding RNAs are not. Transcriptomics is concerned with studying RNA levels genome-wide, both qualitatively (e.g.,

presence of specific transcripts and identifying different splice variants) and quantitatively (quantification of expressed transcripts)[36]. The advent of large transcriptomic studies in the past decade has shown that while only a small percent of the genome encodes proteins, up to 80% of the genome is transcribed [36]. The non-coding RNAs are divided into long, short and circular. Transcribed long non-coding RNAs play essential roles in many physiological processes, for example, endocrine regulation and neuron development [36]. Dysregulation of long non-coding RNAs has been implicated in various diseases, such as diabetes and cancer [36]. Similar to long non-coding RNA, short RNAs (microRNAs) and circular RNAs have growing evidence that points to the their dysregulation in various diseases [36].

Gene expression levels are commonly inferred by measuring the levels of abundance of mRNA in the cells or tissues. Measuring gene expression levels provide valuable information such as viral infection in a cell or signaling pathway disruption in cancer.

There are several approaches for gene expression profiling and they are either array or sequence-based methods. Quantitative polymerase chain reaction (PCR) and RNA microarrays measure the relative activity of previously identified target genes. An array or "chip" may contain probes to determine transcript levels for all known genes in the genome.

Alternatively, RNA-Seq is a sequence-based method that uses NGS for measuring gene expression levels. It provides a relative concentration measure of different mRNAs in the cells. An advantage of RNA-Seq over array-based methods is that it allows for the measurement of the transcripts, with a known or unknown sequence. Additionally, since it is sequence based it can be used to identify SNVs, splice-variants and novel expressed genes. In RNA-Seq, all the transcripts are isolated and complementary DNA (cDNA) sequences are created by reverse transcription of the transcripts. Subsequently, barcode DNA adapters are added to the cDNA followed by the cDNA amplification and sequencing. Rigorous quality control is then applied and the reads with high quality are kept and then mapped to a reference genome. Reads mapped to genes are then quantified, annotated and normalized to account for biases such as gene length. RNA-Seq quantifies the average expression of genes in bulk samples resulting in a comprehensive illustration of the transcriptome.

Epigenomics

Epigenomics focuses on genome-wide modifications of DNA and DNA-associated proteins, such as DNA methylation (DNAm) and histone acetylation. Modifications of DNA and histones are major regulators of gene transcription and subsequently of cellular fate [36]. Those modifications can be influenced both by genetic and environmental factors and have shown evidence to be of importance in biological processes and disease progression such as metabolic syndromes and cancer [36].

For DNA methylation, the principal form is methylation of cytosines in cytosine–guanine dinucleotides (CpGs). The addition of a methyl group at the 5′ position of the cytosine residue is an essential mechanism in both gene expression and chromatin structure regulation [37]. Previously, differentially methylated regions were mostly studied in the promoter regions, CpG islands (CGIs) and enhancers [37].

However, recently it has become obvious that DNAm is highly active outside such regions (mainly non coding regions) [38]. On the other hand, for histone modifications the most common include mono-, di- or trimethylation, as well as acetylation of one or more amino acids in the amino-terminal tails of core histones [38]. The role of DNAm variation in complex disease has mainly been explored in the context of cancer. For non-malignant, common complex diseases, such as diabetes or autoimmunity, the epigenetic component is only just beginning to be investigated [38].

Pyrosequencing, methylation-specific PCR, and direct Sanger sequencing have been the most widely used methods for analysis of targeted regions, such as a promoter region of a single gene or a CpG island [37]. During the recent years several newer methods have been developed to overcome the limitations of the former ones such as low quantitative accuracy, short read length, low sample throughput and targeting only specific genes[37].

Microarray hybridization was one of the first technologies enabling the DNAm studies in a genome-wide level [37]. It enabled simultaneous analysis of samples which improved the former throughput[37]. However, the coverage is dependent on the array design[37].

NGS platforms have vastly improved DNAm analysis and allowed for construction of the genomic maps of genome-wide DNAm at a single base resolution [37]. To date, there are several NGS genome-wide techniques for DNAm profiling. Among the most common are affinity enrichment-based methods, restriction enzyme-based methods and bisulfite-based methods [37]. Each of these methods has its pros and cons and is suitable for different applications. For example, bisulfite-based methods are suitable for high resolution studies but not for site specific targeted studies and it requires high DNA input and higher costs, whereas restriction enzyme-based methods are more suitable for specific targeted studies and are less costly [37].

Metabolomics

Metabolomics is an omic science that deals with the global assessment of the metabolites present in a biological system (e.g., biofluids, cells and tissues) to evaluate the progress of a disease, select potential biomarkers, and provide insights into the underlying pathophysiology [39]. Metabolites are small molecule types, the substrates and products of metabolism that drive essential cellular functions, such as energy production, signal transduction and apoptosis [40]. Examples of metabolites include amino acids, fatty acids, and

carbohydrates [36,41]. Metabolite levels reflect metabolic function, and they are indicative of disease if they are outside normal range [36]. Measuring of metabolite levels has made it possible to discover novel genetic loci regulating small molecules [36]. Research for metabolic marker discovery spans a fast-growing array of prevalent disease areas, such as breast cancer, osteoarthritis and Alzheimer's disease [41].

Nuclear magnetic resonance (NMR) and Mass spectrometry (MS) are common analytical tools used to quantify hundreds of metabolites within a reasonable time frame [41]. These techniques must be coupled with appropriate statistical and computational algorithms to make sense of the data. Currently, there are several databases that reports the existence of over 100,000 metabolites in the human body such as Human Metabolome Database (HMDB)[42], chemical entities of biological interest (ChEBI) [43], the Kyoto encyclopedia of genes and genomes (KEGG) [44], database of U.S. national center for biotechnology information PubChem [45] and Lipidomics gateway (LIPID MAPS) [46].

Other omics

There are several other omics types that contribute to biomarker discovery in complex diseases. The list includes but is not bound to proteomics, Imiomics and HiCAP. Proteomics is typically a crucial dimension in any major multi-omics complex disease study. It involves the large scale studies of proteins and the exploration of proteomes from the overall level of protein composition, structure, and activity [47,48]. Imiomics deals with quantification and analysis of phenotypes (e.g., fat volume) from multiple tissues using integrated imaging platforms (e.g., positron emission tomography PET scans and magnetic resonance imaging MRI). HiCAP are studies that explore the interaction of gene regulatory elements such as promoters and distal elements (enhancers).

The here mentioned omics types are outside the scope of this thesis but it was worth mentioning their roles in large multi-omics studies.

Analyzing omics data

NGS technologies produce massive amounts of omics data. The availability of the huge high-throughput data comes with challenges in analyzing them. These challenges include, for instance, heterogeneity of the data sources, the noise of the experimental omics data due to the variety of experimental techniques and environmental conditions, high dimensionality of the data in terms of the number of features compared to the number of samples, existing outliers and missing values[49]. Big data analytics covers integration of heterogeneous data, data quality control, analysis, modelling, interpretation

and validation [49]. Application of big data analytics on omics data provides comprehensive knowledge discovery. Particularly, it enables identifying clusters and correlation between objects in a dataset (e.g., patient samples) based on their features' values (e.g., gene expression levels). Moreover, it can identify correlations and clusters between datasets from different cohorts, as well as develop predictive models using statistical and computational approaches such as machine learning[49].

Pre-processing of transcriptomics data

As previously mentioned, omics data is usually accompanied with inherent noise as a result of factors such as different environmental conditions and experimental techniques. Rigorous pre-processing steps have to be employed to assess the quality of the data and standardize it prior to any analysis step. Pre-processing steps usually enfold handling noisy data, outliers, missing values, data transformation and normalization.

Here, I further explain the various pre-processing steps that are generally employed on microarray and RNA-Seq data, which are within the scope of this thesis (Figure 4).

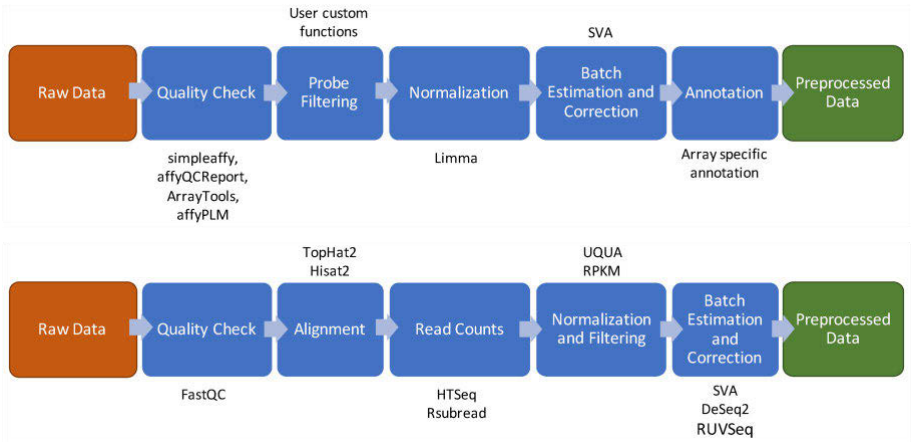

Figure 4. Data pre-processing schema for microarray (the top pipeline) and RNA-Seq (the bottom pipeline). The brown box indicates the input of the pipeline. The green box indicates the output of the pipeline. The blue boxes show the intermediate steps of the pipeline and above or below the boxes are listed the software/packages employed in the step. Figure adapted with permission from MDPI Publishers: Nanomaterials (Basel)., (Antonio. et al., 8;10(5)-903) copyright 2020

Quality Check

Gene expression measured using microarray experiments can be significantly affected by different sources of systematic and random errors that may occur at different levels of the experiment such as poor probe design or inappropriate

sample treatment[50]. Quality check (QC) methods aim at standardizing gene expression distributions across samples. A special handling in the QC of microarray data should be directed to the detection of RNA degradation signals. A commonly employed approach for detecting RNA degradation level, is to compute the Normalized Unscaled Standard Error (NUSE), the Relative Log Expression (RLE) and the slope of the RNA degradation curve (RNADeg) [50]. The sample outlierness can be detected by investigating the distributions of the values of the three metrics and visualizing the data using boxplots. The sample outlierness is evaluated based on the data distribution for each measure[50]. Similar to microarray experiments, RNA-Seq procedures sometimes suffer from certain biases. There are two common approaches to maintain the samples with good quality for downstream analysis. The first excludes the samples with substantial RNA degradation from further analyses based on establishing an arbitrary cut-off value [50]. The second one is to apply a standard normalization procedure in case the decay of RNA is comparable across samples which corrects the variation in gene expression estimates[50].

Normalization

Normalization of raw data is a crucial step for either microarray or RNA-Seq pre-processing. Normalization allows to adjust the individual hybridization intensities or gene expression levels across or within samples in order to achieve fair comparisons [50]. Prior to normalization it is essential to remove probes with low variance and low intensities in case of microarray data or low counts and low variance genes in case of RNA-Seq data[50]. This step will subsequently lead to an analysis with robust statistical significance due to the reduction of the number of hypotheses tested.

For microarray data, different methods have been proposed for normalization. The first approach is the scale normalization approach that allows the samples to have the same median and absolute deviation[50]. This strategy does not put into consideration that the shape of the distributions might vary between different arrays, which could be problematic. The second approach normalizes the samples by adjusting their variability, such as the Locally Weighted Scatterplot Smoothing (LOWESS) algorithm [50]. The third approach is the quantile normalization, which assumes that the statistical distribution of each sample is the same, hence applying a scaling approach that also accounts for the variability [50].

Specific normalization methods were developed for RNA-Seq data due to the nature of it (e.g., paired-end reads, single reads). The most classical and commonly used method for normalization of RNA-Seq data is the Reads per Kilobase of exon model per Million mapped reads (RPKM), which allows for "within sample" normalization, scaling the read counts based on the transcript length (expressed in kilobases) and "between samples" normalization (correcting the read counts based on the library size) [50]. For paired-end reads, the

normalization approach commonly uses Fragments per Kilobase of exon model per Million mapped reads (FPKM) since it considers the fragment (both pairs) rather than the single read [50]. The Trimmed Mean of the M-values (TMM) is another approach, which is commonly used for normalization. TMM is a between-sample normalization method in contrast to within-sample normalization methods (RPKM or FPKM) [51]. The TMM normalization method assumes that most of the genes are not differentially expressed [51]. Additionally, it normalizes the total RNA output among the samples and does not consider gene length or library size[51]. TMM is most effective in normalization of samples with different RNA sources (e.g., samples from different tissues) [51].

Batch effect estimation and correction

Gene expression data can be affected by biological and non-biological confounders. The variability in gene expression due to these types of variables is known as batch effect[50]. Batch effects can arise due to factors such as environmental conditions, amplification protocols, different sites/laboratories in which the experiments are performed and different chip or platform types[50]. These batches have a detrimental effect on the quality of the data and can ultimately lead to incorrect results[50]. A critical issue is to remove the effects associated with batch variables while maintaining the variation associated with biological variables[50]. Batch effects can arise from known or unknown sources. Examples of known sources are biological (e.g., treatment, age, tissue) and technical (e.g., array, temperature) variables that are provided by the user in the form of metadata. Unknown sources of variation are not linked to any of the phenotypic information provided as input but can be seen when the data is visualized and can be identified through surrogate variable analysis (SVA) R library[50].

An essential step in identifying batch effects prior to correction is by means of visualization of the data using principal component analysis (PCA), prince plot and confounding plots (Figure 5 A-B). Prince plot shows the correlation between the technical variables that can be potential causes of batch effects and the principal components of the expression matrix[50]. It is a powerful method to quantify the effect of batch variables, as well as to reveal the presence of unaddressed sources of batch behavior in the data[50]. Confounding plots is used to identify known technical or surrogate variables, which are associated with strong sources of variation and are not correlated with biological variables of interest[50]. Correcting for these technical variables in this case should be safe since it will not remove the information addressed by the biological variables of interest. Another approach that allows for studying batch effects is constructing a linear model (e.g., Limma R package) and adding known or unknown potential confounders as covariates in the constructed model. This allows to investigate the covariates effect on gene expression. Tools that are commonly used to correct for known unwanted

sources of variation are edgeR [52] and DESeq2 [53] R packages while SVA R package [54] can be used to detect and correct for unknown sources of variation.

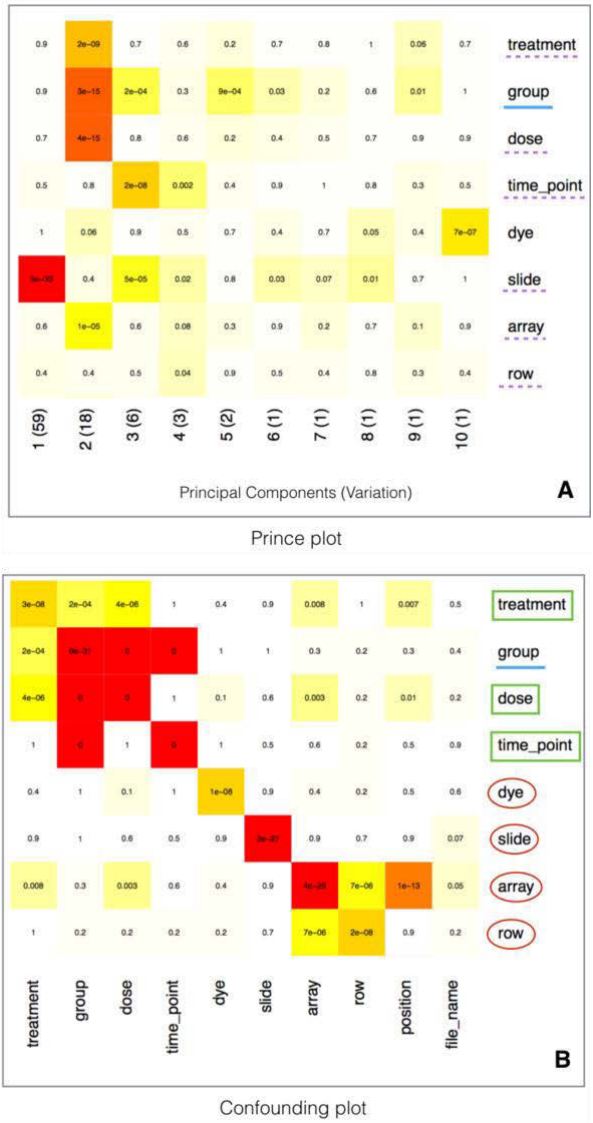

Figure 5. "Panel **A**—Prince plot showing the association between the technical variables and the principal components. The text and the background color in each cell represent the association *p*-value. The row label underlined with solid blue line represents the variable of interest. The row labels underlined with dotted purple line represent other sources of high variation. Panel **B**—Confounding plot, representing the correlation among the technical variables. The row label underlined with solid blue line represents the variable of interest. The green squares represent the variables confounded with the variable of interest or other batch variables. The row labels circled by red outline are batch variables suitable for correction." Figure adapted with permission from MDPI Publishers: Nanomaterials (Basel).,(Antonio. et al., 8;10(5)-903) copyright 2020

Interpreting omics data

Statistical analysis and machine learning play a major role in interpreting big omics data which is broadly true for any kind of data analyses and not specific to omics data analyses [55]. For statistical analysis, statistical tests or models that reject the null hypothesis and support the alternative hypothesis are used while machine learning estimates meaningful relation from the data to generate a predictive model [55].

Statistical tests

Statistical tests provide means to compute the levels of significance or importance of variables. This is essential when dealing with biological data that constitutes thousands of background variables that shadow the effect of the important ones. Testing the significance of a hypothesis is a crucial step to draw solid conclusions. The aim of hypothesis testing when dealing with omics data is to select events that occur due to biological variance rather than random factors. In hypothesis testing, a null (H0) and an alternative (H1) hypothesis are compared against each other. H0 is the hypothesis that assumes there is no significant difference between populations included in the test and H1 assumes there is a significant difference. The main aim of hypothesis testing is to test whether H1, that models the observed data, is significantly different than H0, that models the random events. When the distributions from H0 and H1 are compared, a test-statistic is computed to represent the difference between H0 and H1. The test-statistic basically takes the observed data from an experiment (H1) and compares the results to expected results (H0). A significant level referred to as p-value is then computed using the test-statistic and it represents the probability of obtaining the observed data if the null hypothesis was true (by random chance). The p-value is then compared to a pre-determined threshold (α level) which is the probability of rejecting the null hypothesis if it was actually true. In other words, α level is the probability of making a wrong decision. If p-value is less than the α level then we can safely reject the null hypothesis and claim that the results obtained are significant. The α level is analogous to confidence intervals. The confidence interval is the range of likely values for a population parameter (e.g., population mean). The α level is usually chosen to be equal to 0.05. If an α level 0.05 is chosen then it means the null hypothesis H0 can be rejected with 95% certainty (one out of 20 times to falsely reject the null hypothesis).

For computing t-statistic it is usually essential to make a parametric assumption of the population from which the observed data is sampled, and this assumption is usually hard to meet. Resampling statistical methods are used to infer test-statistics from the sampled data empirically. Resampling statistical methods include bootstrapping and Monte Carlo permutation tests that generate empirical statistical distributions. In the bootstrapping method,

the original set is sampled with replacement to generate sets. The test-statistic for each sampled set is computed. Subsequently the test-statistics from all the sampled sets are used to build a bootstrapping distribution and the test-statistic from the original set is compared to it in order to estimate the statistical significance [56]. In a permutation test the distribution of the test-statistic is computed from all the possible rearranged sets of the original set, and then the original test statistic is compared to it to estimate the statistical significance. However, calculating test-statistics for all possible rearranged sets is computationally intensive and time consuming. To balance the accuracy of the results and computational time Monte Carlo methods are commonly used. Monte Carlo methods make random rearrangements that are large enough to approximate all permutations.

Statistical quality measures

When performing statistical tests there are chances of attaining false significant results which could lead to wrong conclusions also referred to as Type I errors or Type II errors (false positives (FP) or false negatives (FN)). In order to test the goodness of a statistical test some quality measures need to be computed based on the rate at which the test makes wrong or right decisions. The true positive rate (TPR) also referred to as sensitivity is the rate at which the test obtains correctly positive results and true negative rate TNR is the rate at which the test obtains correctly a negative result. False positive rate (FPR) is the rate at which the test incorrectly obtains positive results and false negative rate (FNR) is the rate at which the test incorrectly obtains negative results.

$$TPR = \frac{TP}{TP + FN} = 1 - FNR$$

$$TNR = \frac{TN}{TN + FP} = 1 - FPR$$

$$FPR = \frac{FP}{FP + TN} = 1 - TNR$$

$$FNR = \frac{FN}{FN + TP} = 1 - TPR$$

Correction for multiple testing

Thousands of statistical tests are usually performed simultaneously when dealing with omics data which increases the probability of false significant tests (Type I errors or Type II errors). Several methods have been introduced

to deal with errors that increase due to performing multiple statistical tests. One of the most widely used methods is False discovery rate (FDR). FDR is defined as the expected proportion of false positives among all significant tests. It is based on computing an expected value *Q*, which is the proportion of false discoveries among all discoveries and then the FDR method keeps FDR below a given threshold q [57]. Another commonly used method is the Bonferroni correction, which compensates for that increase of Type I errors by testing each individual hypothesis at a significance level of α/m, where α is the pre-determined alpha level and *m* is the number of hypotheses[58].

$$FDR = \frac{FP}{FP + TP}$$

Differential Gene Expression analysis

One of the most performed and essential analysis when dealing with omics data is the differential analysis. For transcriptomics data it is the differential gene expression analysis (DGE). DGE performs statistical analysis that uses normalized read count data to discover changes in expression levels between different experimental groups. DGE performs tens of thousands of statistical tests at once in one experiment, which theoretically increases the probability of false significant tests (which could lead to wrong conclusions) also referred to as Type I errors. Methods such as FDR and Bonferroni correction (introduced previously) are used within DGE analysis to handle this problem.

Some examples of the commonly used R packages for DGE are edgeR [52] and Limma [59]. The method of edgeR is based on modelling count data using an overdispersed Poisson model, and uses an empirical Bayes procedure to moderate the degree of overdispersion or variation across genes. Limma [59] on the other hand models the data in the form of a linear model (gene wise linear model) to perform the DGE.

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) is a method to identify whether a set of genes are over-represented compared to the expected background distribution of the genes in the gene ontology (GO) terms (biological process, cellular component or molecular function) [60]. It is essentially assigning a functional profile to the gene set under question. GSEA uses pre-defined gene sets that have been grouped together according to their involvement in the same GO term. It compares the input gene set to each of the terms in the GO and performs a statistical test on each term to see if it is enriched for the input genes [60]. The steps for GSEA can be summarized as follows:

1. An enrichment score (ES) is computed to determine the over-representation of the genes in the input gene set. To compute the ES the genes in the input gene set are compared to either the over expressed or the under expressed genes associated to a GO term using Kolmogorov-Smirnov statistical test [60].
2. Estimate the statistical significance of the ES using a permutation test in order to produce a null distribution for the ES [60].
3. Adjusting for multiple hypothesis testing of multiple gene sets [60].

Survival Analysis

Survival analysis is a statistical analysis that models the expected duration until an event occurs (time to an event) using a survival function $\hat{S}(t)$ [61]. The survival function represents the probability that a person survives the event longer than time t [61]. The survival function is usually estimated using the Kaplan-Meier (KM) curve, which is based on computing the Kaplan-Meier estimator. The Kaplan-Meier estimator is non-parametric and represented as [61]:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

t_i is the time when at least one event happened, d_i is the number of events that happened at time t_i, and n_i is the number of individuals known to have survived up to time t_i.

Machine Learning

Machine learning (ML) is a broad term that encloses a collection of computer algorithms aimed to identify shared patterns of variables in input dataset samples (training dataset). The training dataset is represented in the form of an input table where features (variables) constitute columns, and samples or objects constitute rows. Machine learning algorithms are further divided in to two large subgroups of supervised and unsupervised learning.

Unsupervised learning

Unsupervised learning algorithms identify groups or clusters of objects based on shared or similar patterns of features in the training data.

Clustering is a subcategory of unsupervised learning which encompass a group of algorithms. The broad aim of the algorithms is to identify groups of objects from the training dataset that have similar patterns of features' values. The algorithms differ mainly in the way the similarity between objects is measured. Hierarchical and k-means clustering are examples of clustering

methods that are widely used when dealing with omics data. Hierarchical clustering measures the similarity of all pairs of objects and then uses a dendrogram to visualize the distances by placing the most similar pairs close to each other. In k-means, the expected number of clusters k should be predefined and then iterative computational rounds are performed to assign objects to the nearest cluster. Each cluster is described by a computed mean of the samples in the cluster commonly called the cluster centroid. In each computational round the objects are reassigned to the cluster based on the minimum distance between each object and the clusters 'centroids. The algorithm eventually stabilizes when the object assignment is un-changed in each iteration.

Dimensionality reduction is another subcategory of unsupervised learning. These methods aim at representing large sets of variables in a lower-dimensional space using latent variables. Latent variables are variables estimated from the data which represents the variability of the original set. Principal component analysis (PCA) is a widely used algorithm under the dimensionality reduction subcategory. PCA transforms the input variables to a smaller set of orthogonal variables, which are called principal components (PCs). Each PC captures an aspect of variation in the original data. For example, PC1 represents the highest variation in the dataset whereas PC2 represents the second highest variation. The basic structure of data can be explored by plotting the first two principal components in a two-dimensional space. The plot helps in observing whether there are certain clusters of samples based on the PCs and by overlaying the information of a certain phenotypes on the plot it is possible to learn that there are possible existing batch effects or confounders in the dataset. However, since the clusters observed are based on latent variables which led to a change in the original co-ordinate system it is hard to interpret from PCA what are the real causal factors.

Supervised learning

Supervised learning contains methods that identify combinations of variables and their values (descriptors) from the training dataset that can optimally predict a predefined attribute also known as decision. Moreover, these methods learn a mapping between a set of input variables X and an output variable Y and apply this mapping to predict the outputs for unseen data.

Supervised learning can be further divided into regression and classification methods. Regression is a statistical model that is used to estimate relationship between variables by finding a relationship between the input and output variables and representing it as a continuous function. This function is later used to make predictions on unseen data. One of the most widely used regression methods is linear regression. In linear regression, the output or dependent variable is predicted from a linear combination of weighted set of

features and a constant called an intercept. The aim is to compute the intercept so that the output variable is optimally approximated and the distance between the predicted and experimental value (also known as residual) is minimized. Linear regression assumes a linear relationship between the input and output variables and that they are continuous, which is not always the case. Logistic regression can circumvent this limitation since the output variable is allowed to be binary or nominal and the independent variables are allowed to be a combination of continuous and/or categorical variables.

Classification methods do not assume a predefined relationship between input and output variables but learn the relationship in a data driven way. The input dataset will usually be divided into training and test datasets. The methods learn the model from the input training data set by going through several computational rounds. For each round they try to optimize the parameters of the model by minimizing the error between the observed data (training data) and the model's predicted data. In general classification methods cover three main phases including model building from training dataset, evaluation and tuning of the model, and then using the model for prediction-making and possibly interpretation of the results. Some of the commonly used algorithms in ML methods are artificial neural networks (ANNs) [62], which is inspired by how the human brain works, and support vector machine (SVM) [62], which is based on learning the coefficients of parameters for finding a hyperplane separating between two decision outcomes. Even though most of these methods are regarded as powerful tools common criticism when applying them to biological data is that they are 'black boxes', meaning their internal logic cannot be easily understood.

In this thesis we focused more on interpretation of the results rather than prediction accuracy. We thus favored to rely on interpretable classification models that are also know to perform fairly well (in terms of predictive performance) when the number of features outnumber the objects available in the dataset. Examples of interpretable models are decision trees and rule-based models such as rough set and fuzzy logic-based models. A decision tree is a set of true/false questions nested in a hierarchical structure and they are inherently interpretable because the content and order of each question can be directly observed. Rough set rule-based models are inherently interpretable because the predictive model is constructed using a set of rules and the original coordinate system is unchanged (unlike PCA). Further, interpretation of the model can be done by visualizing the set of rules in the form of networks where a relationship between co-predictive features can be easily comprehended and explored. Additionally, the rough sets-based models can provide multiple minimal solutions unlike other machine learning models. This is useful specifically when dealing with biological problems if the aim is to learn how biological systems can deal with a problem in multiple ways (e.g., different pathways).

Rough set rule-based machine learning

The Rough set rule-based method offers classification transparency since it can be used to build interpretable machine learning models [63,64]. This method uses feature selection to reduce the data dimensionality and provides minimal subsets of features and feature values that serve discernibility. Data is collected into decision tables where rows represent objects and columns represent features, the final column is the decision variable. The data in the decision table is then discretized, since in rough sets theory the features and the decision are expected to be discrete. However, the discussion on the different discretization algorithms that can be used is beyond the scope of the thesis. The decision table is then used as an input to the algorithm, which utilizes rough-set theory to group objects based on a certain criterion. The criterion is that objects that cannot be discerned based on all the features in the decision table are grouped together in an equivalence class. Subsequently, the approach constructs a discernibility matrix utilizing the equivalence classes. The discernibility matrix specifies the features that discern between each pair of equivalence classes. The algorithm applies Boolean reasoning to the discernibility matrix and constructs a discernibility function. The discernibility function is represented as all the possible conjunctions of the entries in the discernibility matrix (the features discerning between pairs of equivalence classes) which is consequently simplified using Boolean algebra. The result of the simplification are minimal subsets of features called reducts that can discern between all the equivalence classes. Reducts are overlaid on the objects in the decision table to create IF-THEN rules. The set of IF-THEN rules represent combinations of variables and their corresponding levels from all the reducts based on all the objects in the decision system. The set of rules constitutes the rule-based classification model. To aid interpretation, the rules generated by the model can be visualized as Rule Networks, where the features involved in rules are the nodes in the network and the nodes are connected together if they co-appear in multiple rules.

In this thesis we used R. ROSETTA [65] and VisuNet that is a visualization tool for rule-based classifiers [66]. An example of a hypothetical rule is shown in Figure 6 and an example of a rule set is given in Table 1.

Left Hand Side (LHS)	Right Hand Side (RHS)
IF X=value1 and Y=value2	THEN decision = value 3

Figure 6. A hypothetical rule obtained after computing reducts. Features X and Y are the result of computing a reduct, value1 and value 2 are the descriptors for features X and Y respectively (or discretized descriptors in case the features have a continuous scale). Decision is the class label and its descriptor is value3. Features and their value constitute the left-hand side (LHS) of the rule and decision is the right-hand side (RHS) of the rule.

Table 1. An example of a subset of a ruleset that is generated by the algorithm after computing reducts for a gene expression level dataset. Features computed from the reducts used to build the rules are presented in this example as genes. Discretized gene expression levels correspond to features descriptors in this example such that 1 is low expression value, 2 is normal expression value and 3 is high expression value. The decision is the decision label for the objects supporting the rule, in this example it represents the disease activity where 1 is mild and 3 is severe disease activity.

IF	*LOC649923* =1 and *CKAP4*=2	THEN disease activity is 1
IF	*LOC649923*=1 and *KLRB1*=2	THEN disease activity is 1
IF	*KLRB1*=1 and *MT1F*=3	THEN disease activity is 3

Performance measures of the rules

Quality of the rules are assessed using measurements of support, coverage and accuracy [65]. The rule support represents the number of objects that fulfil the rule conditions[65]. LHS support is the number of objects that satisfy the LHS of the rules [65]. RHS support is the number of objects from the decision system that supports the RHS of the rule. The rule coverage can be determined from the LHS or RHS support as a percentage of objects contributing to the LHS or RHS of the rule [65]:

$$\text{coverage}_{RHS}(\text{rule}) = \frac{\text{support}_{RHS}(\text{rule})}{n_d}$$

$$\text{coverage}_{LHS}(\text{rule}) = \frac{\text{support}_{LHS}(\text{rule})}{n_d}$$

where $\text{support}_{RHS}(\text{rule})$ is the number of objects satisfying the RHS rule conditions, $\text{support}_{LHS}(\text{rule})$ is the number of objects satisfying the LHS rule condition and n_d is the total number of objects from the decision table for a decision class d defined by the rule.

Accuracy of the rule represents its predictive strength [65].

$$\text{accuracy(rule)} = \frac{\text{support}_{RHS}(\text{rule})}{\text{support}_{LHS}(\text{rule})}$$

Rule significance or p-value is computed using the hyper-geometric distribution as described by Hvidsten et al[67].

$$P(x; N, n, k) = \sum_{i=x}^{min(k,n)} \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

Where x is the number of objects correctly classified by the rule, N is the total number of objects in the dataset, n is the support_{RHS} and k is support_{LHS}.

Reduct computations

Generally, reduct computation is a non-deterministic polynomial (NP) hard problem that requires extensive computational time. Computing approximate reducts is thus a good alternative. Approximate reducts can be computed using Monte Carlo based algorithms or evolutionary algorithms. R.ROSETTA [65] utilizes two algorithms for computing reducts in a reasonable time, which are the Johnson reducer and the Genetic reducers[65]. The Johnson reducer is a deterministic greedy algorithm, and the Genetic reducer is an evolutionary stochastic method based on the theory of biological evolution[65]. For the Johnson algorithm the main aim is to find a feature $a \in A$ that discerns the highest number of object pairs [65]. The Johnson algorithm for computing a single reduct is expressed as follows [65]:

1. Let $R=\emptyset$
2. Let $a_{max} \in A$, a_{max} ∈A is the feature that maximizes $\sum(S)$ where $w(S)$ denotes a weight for subsets $S \subseteq$S and S is obtained from computing the discernibility matrix based on all features. The sum is taken over all S from S that contain a_{max}.
3. Add a_{max} to R.
4. Remove all S from S that contain a_{max}.
5. If S=∅ return R. Otherwise, go to step 2.

The Genetic algorithm is based on Darwin's theory of natural selection [65]. This is a heuristic algorithm for function optimization that follows the "survival of the fittest" idea [65]. In each iteration subsets of S are created using concepts of crossover and mutations related to evolution [65]. In each iteration each of these

subsets (hitting sets) are assessed using a score given by a fitness function f that rewards hitting sets B [65]:

$$f(B) = (1-\alpha) \times \frac{\text{cost}\,(A) - \text{cost}\,(B)}{\text{cost}(A)} + \alpha \times min\left\{\varepsilon, \frac{|[S \subseteq S : S \cap B \neq \emptyset]|}{|S|}\right\}$$

where B are hitting sets such that $B \subseteq A$, S is a set obtained from discernibility matrix, α is a control parameter for weighting between subset cost and hitting fraction, and ε is the degree of approximation [65]. In other words, the hitting sets B that have a hitting fraction at least ε are kept in the list [65].

Performance measures of the rule-based model

Various methods are used to assess the overall prediction power of the model. The most commonly used measure is the prediction accuracy, which is the percentage of correct predictions the model achieves on a test data. If there are few samples it could be hard to split the data into train and test data. In this case, Cross validation can be used. Cross validation simulates the prediction accuracy of the model using the whole dataset. Cross validation divides the dataset into k number of folds. For each fold the data is divided into training data and testing data and a model is trained and then evaluated on the test data. The cross-validation accuracy is the mean prediction accuracy of all the models built for each fold.

Another essential performance measure is the AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve. AUC - ROC curve is a curve where true positive rate (TPR) is plotted against false positive rate (FPR) at various threshold values. The AUC is a measure of the ability of a model to discern between classes. The higher the AUC, the better the performance of the model at distinguishing between positive and negative classes. As AUC gets close to 1, the classifier tends to be able to almost perfectly distinguish correctly between the positive and negative classes. In other words, the classifier is then able to detect more True positives (TP) and True negatives (TN) rather than False positives (FP) and False negatives (FN). Figure 7 shows an example of a ROC curve.

TPR or sensitivity is the proportion of the positive class that got correctly classified by the model.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

TNR or Specificity is the proportion of the negative class that got correctly classified by the model.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

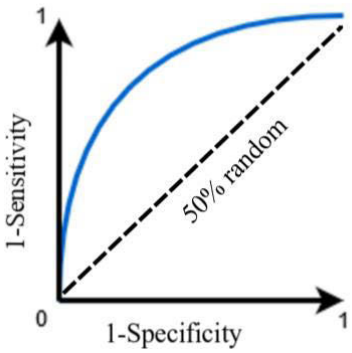

Figure 7. An example ROC curve. The linear line represents how the performance of the classifier would look like if it is close to random guess.

Significance of the model is computed by performing a permutation test. The test is performed by permuting the decision labels of the original dataset N number of times to create random datasets, and rule-based models are then created for these random sets. A normal distribution is built based on the computed model accuracies and an α level is chosen as a threshold of significance. The mean, standard deviation and the standard error for the normal distribution are then computed. The accuracy of the original model is then compared to the mean μ and standard error σ. If the accuracy of the original model was smaller than μ -σ or greater than μ +σ then the p-value (significance of the model in this case) is < α level.

Rule Networks

Rules offer also a new and unique possibility to investigate global properties of rule-based models.

Given rule (1):

IF $Gene_i$=down AND $Gene_k$ = up AND $Gene_z$ = no change THEN decision =DA1

A rule network can be built using rule (1) such that nodes in the network represent features in the LHS of the rule and edges are added for each pair of features occurring as shown in Figure 8.

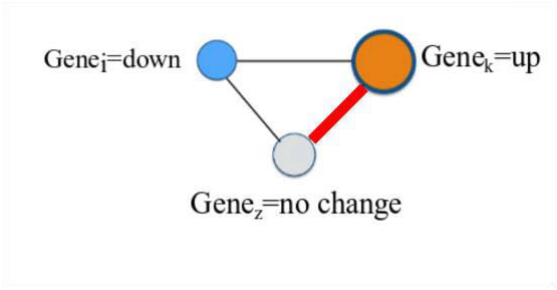

Figure 8. The rule network visualized for rule (1)

The thickness of the edge in the network represents the strength of the connection and is defined as:

$$\text{connection}(x,y) = \sum_{r \in R(x,y)} support(r) \times \text{accuracy}(r)$$

where r is a rule in rule set R and x, y are features. Node size is the mean support value of the rules in which the feature appears in, node border is the percentage of the rules that contain the feature and node color intensity is the mean accuracy value for the rules that the feature appears in.

Rule networks provide a global interpretation of the underlying rule model with the focus on co-prediction. For instance, the rule network can be split into subnetworks corresponding to each decision class. Figure 9 illustrates an example of this case. The figure represents a rule-network of gene expression data by Ye et al. [68] associated to different stimulus responses to CD4+ T cells and how it classifies between races (Afro American, Caucasian and Asian). The example shows how rule-networks succeed in revealing different gene expression levels and genes associated to CD4+ T cells stimulus that are co-predicting each decision class (represented in the network as three clusters). The result presented in Figure 9 is unpublished.

The features that occur most often in the rule set (represented as nodes) and are connected to several others forming hubs in the network suggest a strong contribution to prediction of objects belonging to the class represented by the subnetwork. Rule networks are displayed using a VisuNet R package [66].

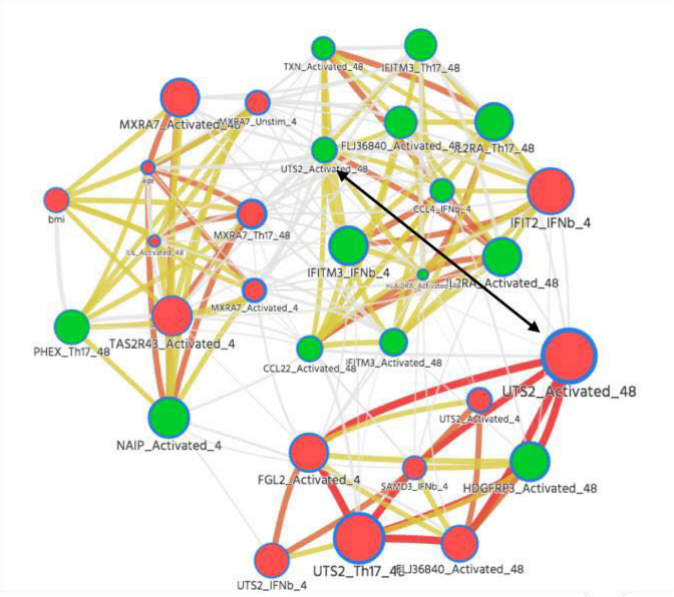

Figure 9. A rule-network of gene expression data associated to different stimulus responses to CD4+ T cells. The network classifies between decision classes represented as races (Afro American, Caucasian and Asian) using data by Ye et al.[68]. The three clusters formed by the rule network represent the states of the genes predicting each of the three decision classes. The result presented in this figure is unpublished.

Feature selection

Omics data is usually characterized by having a large number of features quantified for a limited number of samples. Building a prediction model on all the features in omics data could take an extensive amount of computational time and eventually could result in a dissatisfying prediction model with low accuracy. This is due to the inherent background noise that is usually associated with omics data represented by redundant or unimportant features that might be shadowing the effect of other important features. Applying a feature selection method prior to building a prediction model in this case has positive implications, such as reducing the computational time by removing redundant features. This, in turn, improves the prediction model quality by removing noise and potentially simplifies the interpretation of the model since the dimensionality of the data is reduced (fewer feature sets).

Feature selection techniques are classified into two categories: filter-based and wrapper-based approaches [69]. In the filter-based approach, feature selection is applied prior to the model construction [69]. Wrapper-based approaches depend on the classification model accuracy to enhance the feature set selected to eventually reach an optimal set [69]. For both approaches there are univariate and multivariate methods, which takes the interdependency of

features into consideration [69].The mentioned approaches and methods have advantages and disadvantages [69]. For example, filter-based approaches are faster than wrapper-based approaches [69]. However, wrapper-based approaches can be more accurate since they re-evaluate and re-enhance the selected feature set based on the prediction model accuracy until an optimal solution is reached [69]. Similarly, univariate methods are faster than multivariate methods, but they ignore feature dependencies[69].

Monte Carlo feature selection

In this thesis work we used the Monte Carlo Feature Selection (MCFS) [70] algorithm for feature selection. MCFS is a filter-based, multivariate feature selection to identify the most informative features from a classification point of view [70]. The Monte Carlo method is based on sample randomization techniques and uses decision trees intensively to measure the importance of a feature for classification. The idea is to create subsets with randomly selected features of the original dataset and then build numerous decision tree classifiers in order to compute a total score for each feature [70]. The total score (relative importance) represents the importance of each feature in the classification process. Figure 10 summarizes the MCFS procedure. Initially, *S* subsets of the input dataset are created with the same number of samples and *m* features from the overall feature set *d* such that $m << d$ [70]. Subsequently, each subset is split into *t* training and testing sets [70]. For each pair of training and testing set, a decision tree classifier is constructed, trained and tested [70]. The classification performance of each decision tree is assessed by calculating a weighted accuracy *wAcc* represented in equation (1) [70]. The relative importance score *RI* of each feature is computed using equation (2) and a ranked list of features is created based on their *RI* score [70].

$$wAcc = \frac{1}{c}\sum_{i=1}^{c}\frac{n_{ii}}{n_{i1} + n_{i2} + \cdots + n_{ic}} \tag{1}$$

c represents the number of classes, n_{ij} *is* the number of samples from class *i* classified as class *j*.

$$RI_{gk} = \sum_{\tau=1}^{st}(wAcc_{\tau})^{u}\sum_{n_{gk}(\tau)} IG\left(n_{g_k}(\tau)\right)\left(\frac{no.\,in\; n_{g_k}(\tau)}{no.\,in\;\tau}\right)^{v} \tag{2}$$

s^*t is the number of trees constructed; for each tree (τ). c is the number of decision classes, $IG(n_{gk}(\tau))$ is the information gain for node $n_{gk}(\tau)$, no.in $n_{gk}(\tau)$ is the number of samples associated to node $n_{gk}(\tau)$, no.in τ is the number of samples associated to the root node of the τ-th tree. *u* and *v* are weighting parameters.

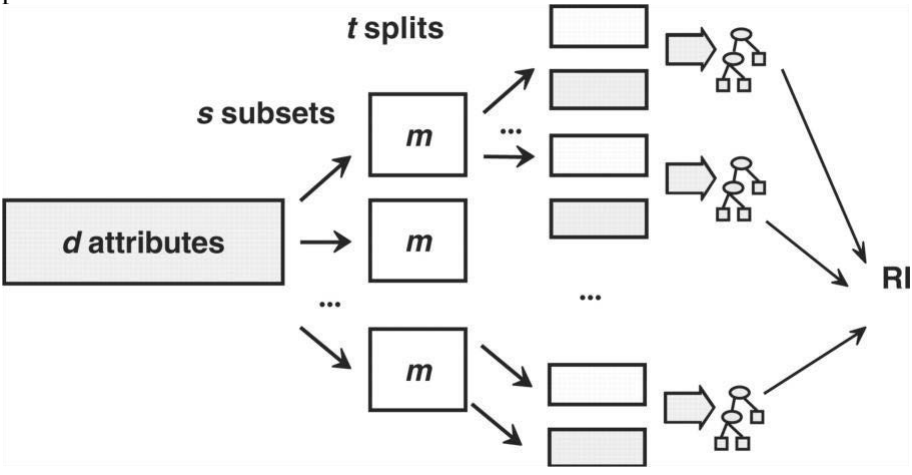

Figure 10. Summary of the MCFS procedure. Figure adapted with permission from Oxford Publishers: Bioinformatics, (Michal, et al., 24;1;110-117) copyright 2008

The rmcfs package [71] performs a permutation test to compute the statistical significance of the result [71]. The permutation test is done by randomly re-labelling the data and repeating the algorithm *N* number of times [71]. A mean and standard deviation of the permuted *RI* scores is computed for each feature and a distribution is constructed [70]. The *RI* scores are assumed to be normally distributed. To compute a p-value each original *RI* score is compared to the *RI* distribution created and subsequently, corrected for multiple testing using the Bonferroni method [70].

Aims

The main goal of this thesis was to provide a novel computational approach to analyze omics data of complex diseases. The approach furnishes new ways to view the interacting factors driving different phenotypic manifestations for disease subtypes. Additionally, we aim at providing a method that would help the scientific community to deal with unification and integration issues that usually arise when fetching and combining omics data from multiple data sources.

More specifically we aimed at:

- Analyzing genomics and transcriptomics data of various AML patient cohorts with statistical and interpretable machine learning methods. The main aim here is to identify novel genomic alterations and transcriptomics factors driving relapse of the disease for patients who have previously achieved complete remission and understand how these factors together co-predict the disease state (Papers I, II).
- Analyzing gene expression data of a pediatric SLE patients cohort using interpretable machine learning to learn about the co-predictive transcriptomic factors driving disease progression from low to high disease state, discover disease subtypes for each state and explore the relationship between the discovered transcriptomics factors and the clinical and phenotypic manifestations for each of the discovered disease subtypes (Paper III).
- Introducing a new method for the scientific community to resolve multiple inconsistencies that exist when fetching and linking metabolomics data from several small-compound databases (Paper IV).

Experimental design

In this thesis we used a combination of a number of publicly available and in-house-generated datasets to explore the alterations of genomics/transcriptomics in AML, transcriptomics in SLE and solving data fetching issues that arise with metabolomics data. Specifically, in Paper I we studied the mutational spectrum and the genomic alterations gained at relapse in AML, which represent potentially actionable therapeutic alternatives. Furthermore, we explored the differences in the mutational spectrum between adult and pediatric relapsed AML. For this purpose, we employed WGS and WES experiments for 48 adult and 25 pediatric AML patients (Cohort I). In Paper II we used Cohort I, the Therapeutically Applicable Research to Generate Effective Treatments (TARGET phs000465) AML cohort and the Cancer Genome Atlas (TCGA) AML[13] cohort to explore the transcriptomics data for complementing the findings in Paper I to further elucidate the biological differences between leukemia blasts at diagnosis and their counterparts at a later stage during tumor progression from the transcriptome perspective.

In Paper III we used microarray gene expression data from Cohort II [30] (158 pediatric SLE patients) to investigate alterations to learn about the co-predictive transcriptomic factors driving disease progression, discover disease subtypes and filling the gap between disease subtypes and clinical and phenotypic manifestations. Finally, in Paper IV we used metabolomics data from Cohort III [72]and Cohort IV [73] to apply our new method for solving data fetching, integration and unification issues that arise when accessing small compounds databases.

Cohort I

Bone marrow or peripheral blood cells from 48 adult and 25 pediatric patients with AML from the Nordic countries, all of whom had relapsed or had primary resistant disease were collected[74]. Primary Resistant (PR) was defined as resistance to treatment without reaching first complete remission, while persistent relapse (R-P) samples were collected after relapse treatment for patients who did not reach complete remission after the respective relapse [74]. Cases that were included were those with relapse or PR specimens of sufficient quality and yield available via the Uppsala Biobank or Karolinska Institute Biobank, collected from 1995 through 2016 (Ref [74]). Cases of the

acute promyelocytic leukemia subtype, which is a clinically distinct type of AML subjected to a different treatment strategy, were excluded. Sixty-six patients had de novo AML, the remaining 7 had a prior diagnosis of MDS or other malignancy [74]. Informed consent was obtained according to the Declaration of Helsinki, and study approval was acquired from the Uppsala Ethical Review Board (Sweden) and the Regional Ethics Committee South-East (Norway) [74]. This cohort was used in Paper I and II.

TARGET cohort

The TARGET project is aimed at studying the molecular characterization in hard-risk or hard to treat childhood cancers in order to determine the genetic changes that drive the initiation and progression of the diseases [75]. One of the subprojects of TARGET is their AML project. DNA and RNA samples included in that study were extracted from peripheral blood or bone marrow tissues.

Several platforms were used for molecular characterization of the TARGET AML project. For Paper II, we used the publicly available mRNA-seq data. For differential gene expression analysis, we included 254 diagnosis samples, and for machine learning-based analysis we only included cases for which relapse samples were available (total of 38 patients, including 29 diagnosis and 38 relapse samples). The data used for the analysis is available in https://portal.gdc.cancer.gov/projects with dbGaP study id phs000463.

TCGA cohort

The Cancer Genome Atlas (TCGA) molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types [76]. In Paper II we used TCGA AML mRNA-seq data for validation of DGE analysis. The data were generated from bone marrow cells of 162 AML patients [13]. AML patients enrolled gave explicit consent for participating in the study, which was approved by the Washington University Human Studies Committee (WU HSC #01-1014) [13].

Cohort II

In Paper III we used data published in a study by Banchereau et al., [30] to explore SLE subtypes, understand the dynamics behind disease progression and the clinical and phenotypical manifestations for the discovered subtypes. The study provided longitudinal data of 158 SLE patients and contained clinical and blood transcriptional profiles followed longitudinally for up to 1,412 days, representing 924 unique clinical visits [30]. Children and adolescents with SLE were enrolled from the rheumatology clinics at Texas Scottish Rite Hospital for Children and Children's Medical Centre Dallas [30].

Informed consent was obtained from adults and the parents of patients younger than 18 years of age [30]. Assent was obtained from patients between 10 and 17 years of age. Blood was collected in Tempus blood RNA tubes (Life Technologies) for RNA-microarray analysis and laboratory measurements were recorded [30].

Cohort III

We utilized data from the study by Diamanti et al., [72] to compare the performance of the tool we developed in Paper IV to other existing tools for solving unification and integration issues arising from fetching metabolites data from small compound data bases. The original study aim was to explore the landscape of metabolic alterations in Type 2 Diabetes (T2D) across five tissues strongly connected to the pathogenesis of T2D [72]. For Paper IV we used 435 metabolites identifiers from the T2D study [72].

Cohort IV

For the same purpose as above for Cohort III, with regards to metabolic data handling, we also utilized data from Priolo et al., [73]. The original study aim was to explore the metabolic reprogramming of prostate tumors that reflects their molecular phenotypes and that could aid the development of metabolic diagnostics and targeted therapeutics [73]. For Paper IV we utilized 229 metabolites identifiers from the prostate tumor study [73].

Methods and Results

In this section I give a brief overview of the four papers included in this thesis. In Paper I and Paper II we mainly focused on analyzing genomic alterations and transcriptomics of AML. In Paper I we investigated the mutational landscape of relapsed and PR AML. Additionally, we used statistical methods to compare the recurrence of transitions and transversions events in multiple groups. Moreover, we used statistical tests for computing p-values for mutually exclusive and co-occurring events between genes with recurrent somatic variants. In Paper II we used differential gene expression analysis to find differentially expressed genes between AML disease states (diagnosis and relapse), and between AML cases characterized by long versus short event free survival for the adult and pediatric patients in Cohort I. Moreover, to further explore the difference between AML disease states we analyzed Cohort I and the TARGET cohort using interpretable machine learning and a rule-network approach to complement the findings of differential gene expression analysis and find potentially highly co-predictive factors affecting disease progression. In Paper III we adapted the same network and machine learning methodology as in Paper II for a similar objective but for a different complex disease, which is SLE and on RNA-microarray data. The aim was to explore the networks of genes that drive SLE progression, explore the subtypes of the disease and link the gene networks to the phenotypical and clinical traits for each subtype. In Paper IV we attempted to solve several data fetching, unification and integration scenarios. More specifically, the issues that exist when fetching and linking metabolomics data from the five small-compound databases HMDB, ChEBI, LIPID MAPS, PubChem and KEGG.

Paper I

"Genomic characterization of adult and pediatric relapsed acute myeloid leukemia reveals novel putative therapeutic targets". Stratmann, S., Yones, S.A., et al., 2021 ***Blood Advances***, 5(3):900–912.

Aims

To investigate recurrent genomic alterations in adult and pediatric relapsed and PR (R/PR) AML patients that may explain disease progression.

Methods

WGS using the HiSeq X Illumina platform was performed for a total of 111 AML samples from 60 different patients [74]. Patient-matched normal DNA was analysed for 60 patients. The genomic coverage was >90X for 99 AML samples, >30X for 12 AML specimens due to limited DNA material and >30X for patient-matched normal DNA. WES using the Ion Proton, Thermo Fisher Scientific platform was performed for 27 AML specimens from 20 different patients and for one patient-matched normal DNA sample. Samples analyzed using WES lacked patient-matched normal DNA or had an insufficient amount of DNA to carry out WGS. The mean coverage for WES reached 131X. Genomic variant calling and annotation were performed using the Sarek [77] pipeline. Calling of somatic SNVs and small insertions and deletions (InDels; ≤50 nucleotides) was done using the Strelka [78] variant calling pipeline [74]. Analysis of CNVs was done using ASCAT [79] and the results were validated utilizing CNVkit [80]. Calling of somatic structural variants including large InDels (≥50 nucleotides) was done using Manta [81] for samples analyzed by WGS.

Two types of survival analyses were carried out, which are the overall survival and the event free survival. The overall survival was defined as the time from initial diagnosis until death or last follow up, while event free survival was defined as the time from diagnosis until first relapse or death [74]. Kaplan-Meier plots were used to visualize the results of the survival analysis and Log-rank (Mantel-Cox) test was used for comparing between adult and pediatric groups using GraphPad Prism 7.02. For comparison and computation of p-values for transitions and transversions between the four groups (adult diagnosis; adult relapse; pediatric diagnosis; pediatric relapse) the Kruskal-Wallis test was used. For the comparisons between group levels a non-parametric pairwise Wilcoxon test was used and correction for multiple testing was done using the Holm-Bonferroni method. To compute p-values for mutually exclusive and co-occurring events between genes with recurrent somatic variants MAF tools R package 2.2.10 [82] was used. MAF tools uses pairwise Fisher's exact test on a 2 × 2 contingency table containing

frequencies of mutated and non-mutated samples. Odds ratio for all pairs of mutation events were calculated to indicate whether an event is more likely to be mutually exclusive or co-occurring.

Results

Our study revealed great plasticity during leukemic progression. Fifty-four percent (54%) of SNVs and small InDels persisted after disease relapse, 34% were gained after treatment, while the rest of SNVs were lost during disease progression [74]. Variants gained at relapse were mainly represented as transversions [74]. Structural variants (including InDels >50bp) and chromosomal gains and losses were stable or gained during disease progression [74]. Of note is that many of the structural variants and mutations that were found in this study for R/PR AML likely would not have been detected with confidence if WES or targeted gene panels had been used instead of WGS [74].

Amid the alterations that were found to be recurrent in adults at R/PR AML were mutations in *ARID1A* (6.3%) and *MGA* (10.4%) [74]. Mutations in the *ARID1A* and *MGA* genes were previously identified in de novo AML studies performed only on pre-treatment specimens, but then identified at very low mutational frequencies. This suggests an overrepresentation of these mutations at relapse and hence an important role for these mutated genes predominantly during leukemia progression.

Further, we reported novel specific differences in the mutational spectrum between pediatric versus adult R/PR AML, with recurrent internal tandem duplications in *UBTF*, encoding Upstream binding transcription factor, found solely in pediatric AML (n=3; 12.0%), while mutations affecting *H3F3A* (6.3%), *ARID1A* and *MGA* were specific for adult R/PR AML. Also, pediatric R/PR cases harbored a substantially higher frequency of mutations in cohesin-associated genes (adults: 10.4% of cases; children: 20.0%). Despite the relatively low frequency of several of the reported mutations described in this study, their identification indicates important roles during disease progression and/or therapy resistance, and they are thus of great interest in the setting of personalized medicine.

Paper II

"Transcriptomic analysis reveals pro-inflammatory signatures associated with acute myeloid leukemia progression". Stratmann S*****, Yones SA***** et al., 2021 ***Blood Advances****, Accepted Manuscript, In press.*

Aims

To investigate changes in the transcriptome in adult and pediatric R/PR AML patients that may explain disease progression.

Methods

We performed RNA-seq on 122 tumor specimens from 47 adult AML patients and 23 pediatric patients from Cohort I, as well as on CD34-expressing BM-cells from five different healthy individuals, used as normal controls. Library preparation was done using IlluminaTruSeq Stranded total RNA [ribosomal depletion] library kit and RNA-Seq using Illumina HiSeq2500 and/or Illumina NovaSeq6000 platforms. Calling of somatic SNVs and small InDels (≤50 nucleotides) was done using HaplotypeCaller GATK [83] with the default settings. Identification of fusion transcripts was done using STARFusion [84]. Quantification of genes and transcripts was done using FeatureCounts [85] and only the expressed protein-coding genes were utilized in the downstream analysis. The expressed protein-coding genes were pre-processed by first normalizing gene counts using the trimmed mean of M-values TMM and then log2-transformation was applied. Thereafter, differential gene expression (DGE) analyses were applied on the pre-processed gene expression data using Qlucore omics explorer 3.6 (Qlucore AB, Lund, Sweden). The resulting DGE results from Cohort I gene expression data were compared to DGE results of TCGA and TARGET validation cohorts.

In order to identify co-predictive biomarkers for disease progression (i.e., to distinguish between diagnosis and relapse) we applied interpretable rule-based machine learning. Three predictive rule-based models were built for adult and pediatric patients in Cohort I (64 samples for the former and 39 samples for the latter) and for the TARGET cohort (64 samples) using R.ROSETTA [65]. Prior to building the rule-based models each cohort's raw data were pre-processed using the same approach as applied above before DGE analysis. The data were then discretized into three bins, corresponding to low, medium and high expression levels for each gene. Subsequently, we applied feature selection on the pre-processed data (for each cohort) using MCFS [70] to identify the most important relapse predictive genes (features). Thereafter, we applied multiple iterative computational rounds on the features selected (Feature boosting) to choose the most optimal ones for building the final models. The final optimal set of features from each cohort was used to

construct three rule-based models using R. ROSETTA. The rules of the rule-based models were visualized to easily identify the highly co-predictive genes of the models using VisuNet R package [66].

Due to the relatively small number of pediatric samples in Cohort I (39 samples) it was hard to detect an obvious highly co-predictive feature. To overcome this obstacle and add power to the predictive model we merged the features from the feature selection step of the pediatric samples in Cohort I and the TARGET cohort into one set and used these features to build more powerful rule-based models for both cohorts. Subsequently, we compared the highly co-predictive features that appeared by comparing the rule networks representing both models. To objectively compare the networks a clustering approach on the predictive strength of the genes (as represented by nodes in the rule network) with respect to the decision classes of the models (diagnosis and relapse) was carried out. The clustering approach for each dataset was proposed by Garbulowski et al. [86]. The clustering was performed on the most informative nodes, using Kendall rank correlation coefficient as a distance metric. Additionally, based on the clustering, the topmost co-predictive genes were selected from the network and visualized as arc diagrams using arcdiagram R package [87].

Results

Utilizing fusion transcript detection on RNA-Seq data all structural variants that had previously been detected in Paper I and were predicted to generate in-frame gene fusions, could be verified. SNVs and small InDels were harder to validate at the transcriptomic level. Ninety one percent (91 %) of variants identified at the genome level and that were located in regions with sufficient read coverage in the RNA-seq data, could be validated at the RNA-level.

DGE analysis was performed to determine prognostic factors in AML. Using an event-free survival (EFS) analysis on the Cohort I diagnosis samples, we detected an association between short EFS and upregulation of *GLI2* and *IL1R1* expression, as well as downregulation of *ST18*. These results that were seen in Cohort I could be validated in two independent AML cohorts; TCGA as an adult AML validation cohort and TARGET as a pediatrics validation cohort. The EFS and DGE analysis indicated an association between pro-inflammatory expression signatures and poor outcome. This association provides a rationale for targeting key regulators of pro-inflammatory pathways in AML such as NF-κB and MAPK. Moreover, relapse samples showed significantly different expression levels of genes such as *CR1* and *DPEP1* compared to the samples that were collected at time of initial diagnosis. The expression profiles of these two genes are expected to facilitate a tumor-promoting environment.

The interpretable rule-based machine learning analysis identified a relapse associated rule-network represented by *CD6*-overexpression and *INSR*

downregulation in adult AML relapse samples. This suggest that AML cells benefit from aberrant production of the lymphoid-associated surface glycoprotein CD6, which potentially might result in AML cells tending to adhere to a protected niche and thus may aid in therapy evasion. Additionally, the AML cells seem to benefit from lower *INSR* levels, which hypothetically may lead to decreased cell proliferation and thus a more quiescent cell state associated with greater chemotherapy resistance. Moreover, restored high expression of *NFATC4* and *KATNAL2* were associated with relapse as predicted through network comparison between the local pediatric cohort and the TARGET cohort. This suggests that low *KATNAL2* and *NFATC4* protein levels promote leukemia onset, while higher expression of the *KATNAL2* and *NFATC4* genes are selected for at relapse. Overexpression of *NFATC4* has been associated to cell quiescence and chemotherapy resistance in ovarian cancer which might be as well the case for AML. The *KATNAL2* gene encodes a microtubule-severing enzyme. This raises the potential of introducing microtubule targeting drugs as novel treatment alternatives at diagnosis.

Paper III

"Identification of combinatorial markers in pediatric Systemic Lupus Erythematosus and disease subtypes from gene expression data using interpretable machine learning". Yones SA et al., *Submitted*

Aims

- To analyze gene expression data of a pediatrics SLE patients cohort using interpretable machine learning to discover disease subtypes, learn co-predictive transcriptomic factors driving disease progression (potential biomarkers) and explore their relationship with the phenotypic manifestations for each of the discovered disease subtypes.

Methods

In this study we used data from Cohort II. Initial pre-processing was done by combining and averaging gene loci that are represented by more than one probe, before each gene locus was log transformed. Batch effects were identified using Variance Partition R package [88] and corrected using SVA R package [54].

Expression values were discretized before any feature selection steps or rule-based model construction. For each gene, the control data expression mean (μ) and standard deviation (σ) were calculated, and then all expression

data for the gene was projected onto this threshold frame and discretized (Low $\leq \mu - 2\sigma <$ Medium $> \mu - 2\sigma \leq$ High; Numeric values 1, 2, 3).

An initial model was built by first collecting the data into a decision table where unique clinic visit identifiers (objects) were represented as rows and (n=629), and genes (n=33,006) were the features and constituted columns. Each object was assigned a decision label, which is the disease activity (DA1 or DA3). Next, the Monte Carlo Feature selection (MCFS) algorithm [70] was used as a feature selection step to obtain a ranked list of informative features with respect to classifying the objects. A significance cut-off for selecting features from the ranked list was obtained by a permutation test (p-value $\leq$ 0.05). Thereafter, multiple iterative computational rounds on the ranked feature list were used to build multiple classification models (Feature boosting) in order to choose the optimal features that lead to a model with the best overall accuracy. The optimal features were used for building the initial model using R.ROSETTA [65]. The rule-based model was then visualized with the VisuNet R package [66].

The initial rule-based model defined above was used as a base to further improve classification. Data (DA1 or DA3 visits) that did not match the left-hand side support of any significant rules in the previous model were removed (p-value < 0.05). The MCFS [70] process was then repeated after object removal. Prior to building the enhanced rule-based model, a Feature boosting step was again performed on the newly selected features. Following Feature boosting the enhanced rule-based model was built and visualized using R.ROSETTA and VisuNet R packages, respectively [65,66].

In order to identify patient subgroups, a matrix was constructed with maintained observations (visits) as rows and rules as columns. The cells for all observations that supported a rule were all assigned 1 or otherwise 0. Hierarchical clustering based on binary distance as the distance function was applied on this matrix.

For correlating the identified patient subgroups with continuous clinical and phenotypic variables a one-way ANOVA following a post-hoc Tukey HSD test was used to compute significance. A Fisher's exact test was used for the assessment of categorical variables to subgroups.

The association between a cluster's supported rules and clinical phenotypes was assessed by contrasting phenotype values for supported samples (patient visits) of each rule versus the non-supported samples (categorical variables, non-parametric Wilcoxon test; binary variables, Fisher's exact test).

For calculating the significance of the predicative enhanced model, we used a permutation test to compute a p-value where the decision label of the dataset (DA1 or DA3) was permuted 1,000 times and rule-based models were created for these random sets. A normal distribution was built for the model accuracies and an alpha of 0.05 and a 95% confidence interval was used to determine the significance of the p-value. The mean, standard deviation and the standard error for the normal distribution were computed. The accuracy of the original

model was compared to the mean μ and standard error σ. If the accuracy of the original model was smaller than μ -σ or greater than μ +σ then the p-value in this case was < 0.05.

Overrepresentation of gene sets belonging to each cluster and the gene sets belonging to rules in DA1 and DA3 were determined using the R package clusterProfiler [89]. The background list was set as the initial set of 33,006 available loci.

Results

In this study we identified the key regulatory networks that underlie the two disease states, DA1 and DA3 of pediatric Systemic Lupus Erythematosus (pSLE). This was done by reducing the high dimensionality of data drawn from 33,006 gene expression measures across 629 pediatric patient visits to co-predictive rule-networks linked via genes.

Five sub-networks were the result of clustering the patient visits based on the rules in the rule-based model; two subnetworks distinguishing DA1 as a result of treatment response, which are C1 and C2, and three subgroups not related to treatment, within the more severe DA3 disease state (C3, C4 and C5).

Studying the three DA3 subgroups identified using the co-predictive networks we were able to observe that SLE is a condition that spans the axes of both autoinflammatory and autoimmune disease. The C3 sub-group sits on the autoimmune side, and had the clinical hallmarks of hypocomplementemia (low Complement factor c3 and c4 clinical measures) in combination with high anti-dsDNA values, whilst the C4 sub-group likely represented the autoinflammatory side, with normal complement levels and low anti-dsDNA values. Cluster C5 likely represented the intermediate stage between C3 and C4, where a significant shift between neutrophil and lymphocyte involvement is observed. This could translate to an immune complex driven disease state in C5, where the type I interferon process was active. Thus, Network analysis and unsupervised clustering combined both the c3/c4 ratio and the NLR (neutrophil-lymphocyte ratio) biomarker sets and resulted in three separate groups spanning these factors. The clusters were linked to co-predictive rule networks to further understand and explain the progression of the disease from DA1 to DA3.

Important hub genes were identified from the co-predictive rule networks. For DA1 (e.g., *IFI35*, *KLRB1*) and DA3 (e.g., *CKAP4*, *OTOF*; Figure 2 in Paper III). *IFI35* expression is stimulated in response to IFN-α/γ [90] and it can act intracellularly as a negative switch of the innate immune pathway and extracellularly. The IFI35 molecule can act as a Damage-associated molecular pattern (DAMP) in macrophages. In DA1, *IFI35* expression was observed within the medium range, but a change in this value could be key in driving DA1 patients back to a remissive or inactive SLE state. Likewise, the

maintained medium expression of *KLRB1* (encoding the surface receptor CD161) suggests a role for other cell sets, including natural killer (NK) cells as the CD161 surface receptor mark this cell type. NK cells respond to innate cytokines and so promote innate inflammation.

CKAP4 was shown as a highly expressed hub gene in DA3, and the protein product is known to induce autophagy. Dysregulated autophagy can affect the regulation of T and B cell populations [91], and increased autophagy can promote the NF-κB pathway response [92] which can play an important role in the pathogenesis of SLE in a number of ways. *OTOF* was an important hub gene in DA3. This is an interferon inducible gene and it was recently suggested that through interaction with melatonin, OTOF may have a role in proteasome inhibition, which could affect the downstream signal transduction pathway of NF-κB [93]. An anti-inflammatory role of melatonin in SLE pathogenesis has been reported previously [94,95]. Gene networks focusing on *OTOF* may help to explain this anti-inflammatory action, and suggests that further investigation of melatonin treatment in SLE flare could be warranted.

Paper IV

"MetaFetcheR: An R package for complete mapping of small compound data". Yones SA et al., *Submitted.*

Aims

To introduce a new method for the scientific community to resolve multiple inconsistencies and incompleteness that exist when fetching and linking metabolomics data from several small-compound databases.

Methods

In this study we developed MetaFetcheR, which is an R package that unifies data from five open access and widely used small compound databases including HMDB [42], ChEBI [43], PubChem [45], KEGG [44] and LIPID MAPS [46]. The algorithm takes as input a sparse table of known identifiers of a collection of small compounds and works on mapping them to identifiers of other databases by filling in the empty fields.

This is led by a queue-based algorithm. The algorithm handles exceptional cases of empty returns from a query by reiterating exhaustively until all identifiers have been retrieved or cannot be further resolved (Paper IV, Figure 1).

MetaFetcheR package was developed using R version 3.5. All the database queries for installation and mapping tasks are performed using PostgreSQL version 12.

Performance of MetaFetcheR was benchmarked based on three case studies using two datasets (Cohort III and Cohort IV) and three existing tools. The three tools are MS_targeted, MetaboAnalystR along with MetaboAnalyst 5.0 web tool and Chemical Translation Service (CTS) [72,96–98].

Case 1

We compared the performance of the algorithm for mapping metabolite identifiers to the identifiers mapped by MS_targeted on the Cohort II dataset. The comparison was based on the rate of mapped and unmapped metabolite identifiers from both tools. Subsequently, the results from MS_targeted were manually curated and the concordance between MetaFetcheR and the manual curation of MS_targeted results was assessed.

Case 2

We compared MetaFetcheR mapping performance to that of the compound ID conversion function of MetaboAnalystR and MetaboAnalyst 5.0 webtool using data from Cohort III and Cohort IV. The comparison of the mapping performance was based on the rate of mapped and unmapped metabolite identifiers when using metabolite names as input for both MetaboAnalyst tools. Unlike MetaboAnalystR, MetaboAnalyst 5.0 accepts metabolite identifiers as input. In addition to the previous comparison, we also compared the number of metabolite identifiers that MetaboAnalyst 5.0 webtool mapped when the input was the available HMDB, KEGG and LIPID MAPS identifiers in the Cohort III dataset, and the available KEGG identifiers in the Cohort IV dataset.

Case 3

We compared MetaFetcheR mapping performance to that of CTS using data from Cohort III and Cohort IV. CTS accepts lists of metabolite names or metabolite identifiers of the same kind as input and does not support PubChem identifiers. To achieve a fair comparison, we ran CTS and MetaFetcheR three times with the available HMDB, KEGG and LIPID MAPS identifiers in the Cohort III dataset and the available KEGG identifiers in the Cohort IV dataset. ChEBI identifiers were discarded from the comparison using Cohort III since there were only two available entries.

Utilizing MetaFetcheR we were able to assess the quality of the data in small compound databases. To achieve this, we ran a test by selecting 1000 random identifiers from one of the five databases as input to MetaFetcheR and then we investigated the quality of the collection of retrieved identifiers. The test was performed 100 times for each database. The quality of the databases was assessed using three different metrics: *i)* percentage of consistency, *ii)*

percentage of ambiguity, and *iii)* percentage of unresolved cases. Consistency represents the percentage of one-to-one associated cases across all identifiers. Ambiguity is the percentage of original metabolite identifiers linked to multiple identifiers from other databases. Unresolved cases represent the percentage of cases that the original metabolite identifiers failed to link or were absent in all other databases

Results

For the comparison of the mapping performance of MetaFetcheR and MetaboAnalystR using the two datasets (Cohort III and Cohort IV), the mapping rate of MetaFetcheR was ~81% (non-empty fields), while MetaboAnalystR achieved ~48% mapping rate on Cohort III (Paper IV; Figure 2A). For the dataset from Cohort IV MetaFetcheR achieved ~95% non-empty fields rate, while MetaboAnalystR resulted in ~73% mapping rate (Paper IV- Figure 2B). Furthermore, we compared the mapping performance of MetaFetcheR to the one of MetaboAnalyst 5.0 web tool and, for both datasets MetaFetacheR performed better (Paper IV; Figure 2B-2C). A similar performance was observed in the test utilizing CTS and MS_targeted (Paper IV; Figure 3). The mapping rate of MetaFetcheR was on average ~70%, and 68% (non-empty fields), while CTS achieved ~38% and 61% mapping rate on Cohort III and Cohort IV, respectively. For MS_targeted the mapping performance was on average ~34% compared to MetaFetcheR ~71% on Cohort III.

For the test utilizing MetaFetcheR for investigating the quality of the data in small compound databases, KEGG showed the highest consistency percentage (~65%) and the lowest fraction of unresolved cases (~23%) compared to HMDB, which had highest fraction of unresolved cases (~71%) (Paper IV; Figure 4).

Conclusions

Paper I

- Analyzing WGS and WES data from AML patients, we identified novel R/PR specific recurrent genomic alterations (*CSF1R* and *H3F3A*), as well as genes with higher mutational frequencies at relapse than what have previously been reported in diagnosis only studies (e.g., *ARID1A* and *MGA*).
- Further, we reported novel differences in the mutational spectrum between pediatric and adult R/PR AML, with alterations in *UBTF* for children and mutations in *ARID1A*, *MGA* and *H3F3* for adults.

Paper II

- Analyzing RNA-Seq data, we identified novel gene fusions comprising known cancer related genes (e.g., *FOS-PSAP, SRSF3-PLAG1*, *CEBPE-CEBPA* and *REXO1-NF1*), as well as fusions gained during leukemic progression, including recurrent *BCR-ABL1* fusions.
- Using DGE analysis we identified an association between a pro-inflammatory signature associated with AML relapse.
- Using machine learning analysis, we identified relapse specific gene signatures that could function as potential novel biomarkers (e.g., *CD6, INSR, KATNAL2* and *NFATC4)*.

Paper III

- Utilizing machine learning analysis, we identified five patient subgroups for pSLE spanning the axes of autoinflammation and autoimmune signatures, which is an aspect that could further explain the disease progression.
- By correlating the results of the machine learning with the clinical and phenotypic variables we identified that c3/c4 and NLR could be used as biomarker sets to distinguish between pSLE patient subgroups and not just as individual biomarkers.
- The network analysis identified important hub genes that have not previously been associated to pSLE with their observed expression

state in the network (e.g., *CKAP4*, *OTOF* and *KLRB1*). Those genes could potentially be used to stratify pSLE patient subgroups for clinical trials or personalized medicine based on their disease state at a particular time.

Paper IV

- We created the R package MetafetcheR, which allows complete and unbiased mapping of metabolites identifiers across the five most widely used publicly available databases for small compounds.
- MetafetcheR was shown to outperform other existing tools (MS_targeted and MetaboAnalystR, MetaboAnalyst 5.0 webtool and CTS) on two datasets (Cohort III and Cohort IV)
- MetafetcheR was able to provide insights on the data quality of small compound databases.

Svensk sammanfattning

Att avslöja kausala drivkrafter till komplexa sjukdomar är ännu en svår utmaning. Till skillnad från enstaka genstörningar orsakas komplexa sjukdomar av en kombination av genetiska, miljömässiga och livsstilsfaktorer, som vanligtvis ännu inte har identifierats. Komplexa sjukdomar är också kända för att vara heterogena över patientgrupper. Detta lägger till ytterligare ett lager av komplexitet för att identifiera patientundergrupper och deras orsakssamband. För att studera de olika dimensionerna för komplexa sjukdomar är analys av olika omics data en nödvändighet.

Huvudmålet med denna avhandling är att tillhandahålla beräkningsmetoder för analys av omics data rörande två komplexa sjukdomar; akut myeloisk leukemi (AML) och systemisk lupus erytematosus (SLE). Ett av tillvägagångssätten som presenteras i avhandlingen ger ett nytt sätt att se de samverkande faktorerna som driver olika fenotypiska manifestationer för sjukdomsundergrupper. Dessutom siktar vi på att tillhandahålla en metod som skulle hjälpa det vetenskapliga samfundet att hantera enhets- och integrationsfrågor som vanligtvis uppstår när man hämtar och kombinerar omics- (specifikt metabolomics) data från flera datakällor.

AML är en cancer i myeloiska blodceller som är välkänd för sin heterogenitet. Patienter svarar vanligtvis på den första kemoterapibehandlingen och uppnår ett fullständigt remissionstillstånd. En majoritet av dem återfaller dock, och återfallsklonerna utvecklar oftast resistens mot behandlingen. Olika studier har funnit genetiska förändringar och kromosomavvikelser kopplade till AML, till exempel ”The Cancer Genome Atlas”-studien av vuxen AML (TCGA) och ”Therapeutically Applicable Research to Generate Effective Treatments”- (TARGET-) studien av AML hos barn, men lite är ännu känt om de specifika drivkrafterna bakom återfall och resistens mot behandlingen. I Papper I fokuserar vi på att undersöka återkommande genomiska förändringar vid AML-återfall hos vuxna och barn, samt vid primärt resistent sjukdom, vilka kan förklara sjukdomsprogression och behandlingsresistens. I Papper II identifierar vi förändringar i det så kallade ”transkriptomet” hos AML-patienter under sjukdomsförloppet, inklusive maskininlärningsanalys samt information om patientmatchad genomisk bakgrund med hjälp av data från tre olika kohorter (Vår lokalt genererade kohort, TARGET och TCGA).

SLE är en autoimmun sjukdom som kännetecknas av oförutsägbara perioder av skov som drivs av slumpmässig aktivitet hos ett komplext genetiskt program. Skoven presenteras som olika SLE-sjukdomsaktiviteter. På grund av de slumpmässiga skoven, sjukdomskomplexiteten och heterogeniteten bland individer är det mycket svårt att skräddarsy en personlig behandling för SLE-patienter. Flera studier har genomförts för att utforska genetiska skillnader mellan friska kontroller och SLE-patienter. Ansträngningarna för att undersöka de kombinatoriska effekterna av gener i association med manifestationen av olika SLE-sjukdomsaktiviteter i olika patienters undergrupper har dock lett till relativt begränsade framgångar. Patientundergrupper indikerar eventuellt olika SLE-undergrupper. I Papper III analyserar vi transkriptom data från en pediatrisk SLE-patientkohort med hjälp av maskininlärning. Syftet var att undersöka de samprediktiva transkriptomiska faktorerna som driver sjukdomsprogression från ett milt till aggressivt sjukdomstillstånd, upptäcka sjukdomsundergrupper för varje tillstånd och utforska sambandet mellan de upptäckta transkriptomiska faktorerna och de kliniska och fenotypiska manifestationerna för var och en av de upptäckta sjukdomsundergrupperna.

Nyligen har metabolomics tillkommit som en viktig faktor i stora komplexa sjukdomsstudier inkluderande flera olika typer omics-analyser. Metabolomics är en omicsvetenskap som behandlar den globala bedömningen av de metaboliter som finns i biologiska system för att utvärdera utvecklingen av komplexa sjukdomar. Olika databaser innehåller en stor mängd information rörande metaboliter och metaboliska vägar. Den stora mängden av sådana databaser och redundansen av deras information leder dock till stora problem med analys och standardisering. Brist på förebyggande medel kan leda till felidentifierade föreningar och minskad statistisk upplösning. I Papper IV syftar vi till att lösa flertalet inkonsekvenser och ofullständigheter som uppstår när man hämtar och länkar metabolomics data från olika databaser genom att introducera det nya R-paketet MetaFetcheR till det vetenskapliga samfundet.

Acknowledgments

I would like to thank my supervisors Linda Holmfeldt and Jan Komorowski for giving me this great opportunity to pursue my PhD in Sweden. I remember when I first learned that I got accepted to the position I was very sceptical since I never thought of moving to Sweden. Instead, I was considering other English-speaking countries like the US or the UK. I thought it would always be cold, dark and depressing. On the contrary, I feel like I was reborn in Sweden. I gained a wealth of life experiences during these 5 years that I would have never thought of gaining. It was a roller coaster of emotions and knowledge. I can call it a crash course of maturing and independence. I re-discovered myself during the 5 years. It was extremely difficult with all the new variables (coming from a different culture and educational background since I attained my Master's degree in Computer Science). I felt like I was going to be sucked down by a black hole. I felt that learning all the biology was like learning rocket science at the beginning. Turns out that it is doable with some persistence and patience. I would also like to thank myself for hanging on.

To my dear and precious family in Egypt. Thank you so much for believing in me and your patience during those five long years. Thank you for teaching me to always be persistent and aim for success.

To all my mentors and professors who taught me during my undergraduate and graduate years in Uppsala and Egypt. Thank you for sharing your precious knowledge that made me who I am today.

To all my amazing collaborators and students whom I shared authorship of my papers with, specifically (Svea Stratmann, Jennifer Meadows, Klev Diamanti, Alva Annett, Patricia Stoll, Rajmund Csombordi, Mateusz Garbulowski and Fredrik Barrenäs). I honestly couldn't have made it to this day without your support. The projects would have been definitely dangling by now if it wasn't for you. Your input has been so precious to my PhD. Thanks a million. I also would like to thank Morten Herlin for inviting me to contribute with an analysis and including me in his paper as a co-author.

To my former and current colleagues in Komorowski, Holmfeldt and Wadelius labs, specifically (Svea, Nicholas, Karolina, Mateusz, Klev, Husen,

Behrooz, Zeeshan,Marco and Gang). Thank you all so much for your support and kindness. You were all very generous with sharing whatever knowledge you had with me and cheering me up whenever I felt down. Thanks a lot.

To SciLife lab. Thanks so much for making me part of the bioinformatics support program. You gave me an opportunity to share my work and provided me with constructive ideas. Specifically, I want to thank Bjorn Nystedt, Manfred Grabherr and Markus Mayrhofer.

To my friends whom I have known in Uppsala and became my second family, who stood by me during hardships, sickness and depression. I wonder how would I have survived without your existence?! I think the answer would be that I wouldn't. Thank you, my second family, for taking care of me and always being kind and understanding.

To all my friends and relatives in Egypt who are still in contact with me, are very keen to meet up whenever I am there and are always excited to see me. Those individuals are actually few but to me they are very precious. I call them the genuine group. Those are the ones whom I would never want to lose. Thank you all for your unconditional true feelings.

Lastly, I would like to thank everyone and anything who/that has contributed in keeping me somehow mentally stable and healthy during this period. This includes of course my dear friends in Uppsala, the gym, the yoga, psychologists and doctors. If it wasn't for you, I would have been drowning by now.

References

1. Complex Diseases: Research and Applications | Learn Science at Scitable. http://www.nature.com/scitable/topicpage/complex-diseases-research-and-applications-748.
2. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
3. Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).
4. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
5. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
6. Dang, C. V. c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol. Cell. Biol.* **19**, 1–11 (1999).
7. Evan, G. I. & Vousden, K. H. Proliferation, cell cycle and apoptosis in cancer. *Nature* **411**, 342–348 (2001).
8. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).
9. Karlsson, L. *et al.* Outcome after intensive reinduction therapy and allogeneic stem cell transplant in paediatric relapsed acute myeloid leukaemia. *Br. J. Haematol.* **178**, 592–602 (2017).
10. Verma, D. *et al.* Late relapses in acute myeloid leukemia: analysis of characteristics and outcome. *Leuk. Lymphoma* **51**, 778–782 (2010).
11. Bejanyan, N. *et al.* Survival of patients with acute myeloid leukemia relapsing after allogeneic hematopoietic cell transplantation: a center for international blood and marrow transplant research study. *Biol. Blood Marrow Transplant. J. Am. Soc. Blood Marrow Transplant.* **21**, 454–459 (2015).
12. My Cancer Genome. https://www.mycancergenome.org/.
13. Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
14. Li, S. *et al.* Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.* **22**, 792–799 (2016).
15. Hirsch, P. *et al.* Genetic hierarchy and temporal variegation in the clonal history of acute myeloid leukaemia. *Nat. Commun.* **7**, 12475 (2016).
16. Bullinger, L. *et al.* Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N. Engl. J. Med.* **350**, 1605–1616 (2004).
17. Andreeff, M. *et al.* HOX expression patterns identify a common signature for favorable AML. *Leukemia* **22**, 2041–2047 (2008).

18. Metzeler, K. H. *et al.* ERG expression is an independent prognostic factor and allows refined risk stratification in cytogenetically normal acute myeloid leukemia: a comprehensive analysis of ERG, MN1, and BAALC transcript levels using oligonucleotide microarrays. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **27**, 5031–5038 (2009).
19. Zhu, Y.-M. *et al.* Gene mutational pattern and expression level in 560 acute myeloid leukemia patients and their clinical relevance. *J. Transl. Med.* **15**, 178 (2017).
20. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
21. Bennett, J. M. *et al.* Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br. J. Haematol.* **33**, 451–458 (1976).
22. Abuhelwa, Z., Al Shaer, Q., Taha, S., Ayoub, K. & Amer, R. Characteristics of De Novo Acute Myeloid Leukemia Patients in Palestine: Experience of An-Najah National University Hospital. *Asian Pac. J. Cancer Prev. APJCP* **18**, 2459–2464 (2017).
23. Ramos, P. S., Shedlock, A. M. & Langefeld, C. D. Genetics of autoimmune diseases: insights from population genetics. *J. Hum. Genet.* **60**, 657–664 (2015).
24. Cho, J. H. & Gregersen, P. K. Genomics and the Multifactorial Nature of Human Autoimmune Disease. *N. Engl. J. Med.* **365**, 1612–1623 (2011).
25. Adamichou, C. & Bertsias, G. Flares in systemic lupus erythematosus: diagnosis, risk factors and preventive strategies. *Mediterr. J. Rheumatol.* **28**, 4–12 (2017).
26. Song, W. *et al.* Advances in applying of multi-omics approaches in the research of systemic lupus erythematosus. *Int. Rev. Immunol.* **39**, 163–173 (2020).
27. Mikdashi, J. & Nived, O. Measuring disease activity in adults with systemic lupus erythematosus: the challenges of administrative burden and responsiveness to patient concerns in clinical research. *Arthritis Res. Ther.* **17**, 183–183 (2015).
28. Charles A Janeway, J., Travers, P., Walport, M. & Shlomchik, M. J. The complement system and innate immunity. *Immunobiol. Immune Syst. Health Dis. 5th Ed.* (2001).
29. Complement: MedlinePlus Medical Encyclopedia. https://medlineplus.gov/ency/article/003456.htm.
30. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 551–565 (2016).
31. Conley, Y. P. *et al.* Current and emerging technology approaches in genomics. *J. Nurs. Scholarsh. Off. Publ. Sigma Theta Tau Int. Honor Soc. Nurs.* **45**, 5–14 (2013).
32. Wu, J., Wu, M., Chen, T. & Jiang, R. Whole genome sequencing and its applications in medical genetics. *Quant. Biol.* **4**, 115–128 (2016).
33. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19**, R131–R136 (2010).
34. McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42 (2007).
35. Freeman, J. L. *et al.* Copy number variation: new insights in genome diversity. *Genome Res.* **16**, 949–961 (2006).
36. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).

37. Barros-Silva, D., Marques, C. J., Henrique, R. & Jerónimo, C. Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. *Genes* **9**, 429 (2018).
38. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12**, 529–541 (2011).
39. Zhang, A., Sun, H., Yan, G., Wang, P. & Wang, X. Metabolomics for Biomarker Discovery: Moving to the Clinic. *BioMed Res. Int.* **2015**, 354671 (2015).
40. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
41. Lee, M. Y. & Hu, T. Computational Methods for the Discovery of Metabolic Markers of Complex Traits. *Metabolites* **9**, 66 (2019).
42. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
43. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344–D350 (2008).
44. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
45. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
46. Sud, M., Fahy, E., Cotter, D., Dennis, E. A. & Subramaniam, S. LIPID MAPS-Nature Lipidomics Gateway: An Online Resource for Students and Educators Interested in Lipids. *J. Chem. Educ.* **89**, 291–292 (2012).
47. Anderson, N. L. & Anderson, N. G. Proteome and proteomics: New technologies, new concepts, and new words. *ELECTROPHORESIS* **19**, 1853–1861 (1998).
48. Blackstock, W. P. & Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17**, 121–127 (1999).
49. Ristevski, B. & Chen, M. Big Data Analytics in Medicine and Healthcare. *J. Integr. Bioinforma.* **15**, (2018).
50. Federico, A. *et al.* Transcriptomics in Toxicogenomics, Part II: Preprocessing and Differential Expression Analysis for High Quality Data. *Nanomaterials* **10**, 903 (2020).
51. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
52. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2010).
53. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
54. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma. Oxf. Engl.* **28**, 882–883 (2012).
55. Yamada, R., Okada, D., Wang, J., Basak, T. & Koyama, S. Interpretation of omics data analyses. *J. Hum. Genet.* **66**, 93–102 (2021).
56. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (Springer US, 1993). doi:10.1007/978-1-4899-4541-9.
57. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
58. Dunn, O. J. Multiple Comparisons among Means. *J. Am. Stat. Assoc.* **56**, 52–64 (1961).

59. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
60. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
61. G. Miller, R. *Survival Analysis*. (John Wiley & Sons Inc, 1998).
62. Lin, E. & Lane, H.-Y. Machine learning and systems genomics approaches for multi-omics data. *Biomark. Res.* **5**, 2 (2017).
63. Komorowski, J. 6.02 - Learning Rule-Based Models - The Rough Set Approach. in *Comprehensive Biomedical Physics* (ed. Brahme, A.) 19–39 (Elsevier, 2014). doi:10.1016/B978-0-444-53632-7.01102-3.
64. Skowron, A. & Dutta, S. Rough sets: past, present, and future. *Nat. Comput.* **17**, 855–876 (2018).
65. Garbulowski, M. *et al.* R.ROSETTA: an interpretable machine learning framework. *bioRxiv* 625905 (2020) doi:10.1101/625905.
66. Smolinska K, Garbulowski M, Diamanti K, et al. *VisuNet: an interactive tool for network visualization of rule-based models in R.* (2021).
67. Hvidsten, T. R. *et al.* Discovering regulatory binding-site modules using rule-based learning. *Genome Res.* **15**, 856–866 (2005).
68. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
69. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517 (2007).
70. Dramiński, M. *et al.* Monte Carlo feature selection for supervised classification. *Bioinformatics* **24**, 110–117 (2008).
71. Dramiński, M. & Koronacki, J. rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. *J. Stat. Softw.* **85**, 1–28 (2018).
72. Diamanti, K. *et al.* Intra- and inter-individual metabolic profiling highlights carnitine and lysophosphatidylcholine pathways as key molecular defects in type 2 diabetes. *Sci. Rep.* **9**, 9653 (2019).
73. Priolo, C. *et al.* AKT1 and MYC Induce Distinctive Metabolic Fingerprints in Human Prostate Cancer. *Cancer Res.* **74**, 7198–7204 (2014).
74. Stratmann, S. *et al.* Genomic characterization of relapsed acute myeloid leukemia reveals novel putative therapeutic targets. *Blood Adv.* **5**, 900–912 (2021).
75. GenomeOC. TARGET project overview. *Office of Cancer Genomics* https://ocg.cancer.gov/programs/target/overview (2013).
76. The Cancer Genome Atlas Program - National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (2018).
77. Garcia, M. *et al.* Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research* **9**, 63 (2020).
78. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma. Oxf. Engl.* **28**, 1811–1817 (2012).
79. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).
80. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
81. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinforma. Oxf. Engl.* **32**, 1220–1222 (2016).

82. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
83. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33 (2013).
84. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213 (2019).
85. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
86. Garbulowski, M. *et al.* Interpretable Machine Learning Reveals Dissimilarities Between Subtypes of Autism Spectrum Disorder. *Front. Genet.* **12**, 73 (2021).
87. Sanchez, G. *arcdiagram*. (2021).
88. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
89. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).
90. Bange, F. C. *et al.* IFP 35 is an interferon-induced leucine zipper protein that undergoes interferon-regulated cellular redistribution. *J. Biol. Chem.* **269**, 1091–1098 (1994).
91. Clarke, A. J. *et al.* Autophagy is activated in systemic lupus erythematosus and required for plasmablast development. *Ann. Rheum. Dis.* **74**, 912–920 (2015).
92. Zhong, Z. *et al.* NF-κB Restricts Inflammasome Activation via Elimination of Damaged Mitochondria. *Cell* **164**, 896–910 (2016).
93. Yalcin, E. *et al.* Evidence that melatonin downregulates Nedd4-1 E3 ligase and its role in cellular survival. *Toxicol. Appl. Pharmacol.* **379**, 114686 (2019).
94. Bruck, R. *et al.* Melatonin inhibits nuclear factor kappa B activation and oxidative stress and protects against thioacetamide induced liver damage in rats. *J. Hepatol.* **40**, 86–93 (2004).
95. Bonomini, F., Dos Santos, M., Veronese, F. V. & Rezzani, R. NLRP3 Inflammasome Modulation by Melatonin Supplementation in Chronic Pristane-Induced Lupus Nephritis. *Int. J. Mol. Sci.* **20**, 3466 (2019).
96. Pang, Z., Chong, J., Li, S. & Xia, J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. *Metabolites* **10**, 186 (2020).
97. Pang, Z. *et al.* MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res.* **49**, W388–W396 (2021).
98. Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T. & Fiehn, O. The Chemical Translation Service--a web-based tool to improve standardization of metabolomic reports. *Bioinforma. Oxf. Engl.* **26**, 2647–2648 (2010).

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2083

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-454997

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2021