



TAMPERE UNIVERSITY OF TECHNOLOGY

**THOMAS LIUKSIALA**  
**BLOOD CANCER LINEAGE IDENTIFICATION: A MACHINE**  
**LEARNING APPROACH**

Master's Thesis

Examiners: Professor Olli Yli-Harja and  
Professor Matti Nykter  
Examiner and topic approved in  
the Faculty of Engineering Sciences  
Council meeting on 6 May 2015

# TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO

Automaatiotekniikan koulutusohjelma

**THOMAS LIUKSIALA: Koneoppimismenetelmä verisyöpien kehityslinjatunnistukseen**

Diplomityö, 72 sivua, 2 liitesivua

Toukokuu 2015

Pääaine: Laskennallinen systeemibiologia

Tarkastajat: Professori Olli Yli-Harja ja professori Matti Nykter

Avainsanat: koneoppiminen, syöpä, kehityslinja

Syöpä, yksi nykyihmisen suurimmista tappajista, ilmenee virheellisen perintöaineiksen aiheuttamana hallitsemattomasti kasvavana ja henkeä uhkaavana kasvaimena. Muutos solun perintöaineksessa voi johtaa sen geeni-ilmentymisjärjestelmän uudelleenohjelmointiin. Jos tämä muutos johtaa rajattoman kasvun ja jakautumisen mahdollistavaan geeni-ilmentymistilaan, on syntynyt syöpäsolu. Kudoksen geeni-ilmentymistilan voi selvittää geenisirukokeella, joiden tuottamia mittauksia on suuri määrä julkisesti saatavilla. Eri biologisten ilmiöiden lukemattomia ilmentymiä käsittelevän korkealottuvuudellisen aineiston analysointiin liittyy kuitenkin laskennallisia haasteita. Erityisesti koneoppimismenetelmät ovat osoittautuneet arvokkaiksi suuriin biologisiin aineistoihin kätkeytyvien tiedonpalasten louhinnassa.

Tässä diplomityössä esitellään koneoppimismenetelmä syöpäkasvaimen lähimmän terveen geeni-ilmentymistilan määrittämiseksi sekä geenisäätelypoikkeaman mittaamiseksi. Tätä menetelmää sovellettiin hematologisiin syöpiin eli veri- ja imukudoskasvaimiin. Aluksi muodostettiin geeni-ilmentymisaineisto yhdistämällä 9,544 kasvain- ja normaalikudosnäytettä julkisesta tietovarastosta. Laadunvarmistuksella, normalisoinnilla ja systemaattisten virheiden korjaamisella pyrittiin mahdollistamaan satojen, ympäri maailmaa sijaitsevien laboratorioiden tuottaman aineiston yhteisanalyysi.

Pääkomponenttianalyysi, klusterointi ja ohjattu luokittelu vahvistivat, että aineisto todella mahdollistaa eri laboratorioiden tuottamien mittausten geeni-ilmentymistutkimukset. Hematologisten syöpien esittäminen normaalikudosten geenisäätelyllisinä poikkeamina paljastaa niiden kehityslinjan ja asettaa ne säätelyn näkökulmasta kantasolujen ja kypsien verisolujen väliselle alueelle. Tulokset avaavat uusia biologisia hypoteeseja, uuden lähestymiskulman syövän parantamiseen sekä rohkaisevat samankaltaiseen analyysiin myös eri syöpätyypeillä.

# ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY

Master's Degree Programme in Automation Technology

**THOMAS LIUKSIALA: Blood Cancer Lineage Identification: A Machine Learning Approach**

Master of Science Thesis, 72 pages, 2 Appendix pages

May 2015

Major: Computational systems biology

Examiners: Professor Olli Yli-Harja and Professor Matti Nykter

Keywords: machine learning, cancer, developmental lineage

Cancer, one of the most common killers of modern human, is caused by malfunctioning hereditary material manifesting as uncontrolled, life-threatening growth of a tumor. A change in the hereditary material within a cell may cause a re-programming of its gene-regulatory system. If the altered system leads to a gene expression state enabling limitless growth and replication, the cell has become cancerous. The state of a given tissue can be determined by gene expression array experiments. A massive amount of such measurements have been performed and placed under public access by the research community. Analyzing this type of high-dimensional data, spanning countless instances of separate phenotypes, however, poses computational and algorithmic challenges. Machine learning algorithms, in particular, have proven valuable in mining for pieces of knowledge hidden in biological big data.

This thesis presents a machine learning approach to estimate the nearest healthy gene expression state of a tumor and to quantify the regulatory divergence of the tumor from the normal state. The method was applied to hematological malignancies, or cancers of blood and lymph nodes. First, a hematological gene expression data set was integrated from 9,544 tumors and normal tissue samples available in a public data repository. Secondly, quality control, normalization and bias correction steps were performed to enable collective analysis of this data produced by hundreds of laboratories worldwide.

Principal component analysis at different scales, cluster analysis and supervised classification verified that the data set indeed allows for expression studies involving measurements from multiple laboratories. The characterization of hematological malignancies as gene-regulatory deviations from normal tissues uncovers the developmental lineage of the cancers and places them regulatory-wise between stem cells and mature blood cells. The results open up new biological hypotheses, a new approach to curing cancer and suggest that similar analyses in the context of other malignancies could be equally fruitful.

## PREFACE

This thesis project began at the Computational Systems Biology group at Tampere University of Technology and was finished at the Computational Biology group at the University of Tampere. The supervisor, all way long, has been Professor Matti Nykter, head of the latter research group.

Working with this project has introduced me to a stimulating, interdisciplinary field of research and a number of equally stimulating, brilliant minds. In particular, I express my gratitude to Professors Matti Nykter and Olli Yli-Harja for guidance and the possibility to work in such an inspiring environment. I am likewise grateful to Docent Olli Lohi, Docent Merja Heinäniemi and Kirsi Granberg, Ph.D., for providing ideas and their indispensable insight throughout the process. Lastly, I wish to thank my friends and family for their never-ending support during what appeared as a never-ending project.

Tampere, May 2015

Thomas Liuksiala

# SISÄLTÖ

1. Introduction . . . . .	1
2. Theoretical background . . . . .	3
2.1 Cell . . . . .	3
2.2 DNA . . . . .	4
2.3 Gene expression . . . . .	4
2.4 Cancer . . . . .	6
2.5 Hematological malignancies . . . . .	7
2.5.1 Hematopoiesis . . . . .	7
2.5.2 Molecular pathology . . . . .	9
2.5.3 Classification . . . . .	9
2.6 Gene regulatory systems . . . . .	12
2.7 DNA microarrays . . . . .	15
2.7.1 Technology . . . . .	15
2.7.2 Data pre-processing . . . . .	16
2.7.3 Microarray databases . . . . .	18
2.8 Machine learning . . . . .	19
2.8.1 Dimensionality reduction . . . . .	20
2.8.2 Distance measures . . . . .	21
2.8.3 Cluster analysis . . . . .	24
2.8.4 Classification . . . . .	31
2.8.5 Classifier validation . . . . .	35
3. Material and methods . . . . .	39
3.1 Data . . . . .	39
3.2 Pre-processing . . . . .	41
3.3 Dimensionality reduction . . . . .	42
3.4 Cluster analysis . . . . .	42
3.5 Subtype prediction . . . . .	43
3.6 Lineage identification . . . . .	43
4. Results and discussion . . . . .	46
4.1 Gene expression landscapes of cancer . . . . .	46
4.2 Cluster analysis . . . . .	50
4.3 Subtype prediction . . . . .	52
4.4 Cancer lineage identification . . . . .	54
5. Conclusions . . . . .	58
References . . . . .	60
A. Appendices . . . . .	66

## ABBREVIATIONS

ALL	Acute lymphocytic leukemia
AML	Acute myelocytic leukemia
AUC	Area under curve
cDNA	Complementary DNA
CLL	Chronic lymphocytic leukemia
CML	Chronic myelocytic leukemia
CV	Cross-validation
DBSCAN	Density-based spatial clustering of applications with noise
DNA	Deoxyribonucleic acid
EM	Expectation-maximization
FAB	French-American-British classification of leukemia
GEO	Gene Expression Omnibus
GMM	Gaussian mixture model
GRN	Gene regulatory network
HSC	Hematopoietic stem cell
kb	Kilobase, unit of length for nucleic acids equal to 1000 nucleotide bases
KNN	$k$ -nearest neighbors
MAS 5.0	Microarray Suite 5.0
ML	Maximum-likelihood
MM	Mismatch
mRNA	Messenger RNA
NN	Nearest neighbor
PC	Principal component

PCA	Principal component analysis
PDF	Probability density function
PM	Perfect match
RBN	Random Boolean network
RNA	Ribonucleic acid
RCT	Reciprocal chromosomal translocation
RF	Random forest
RMA	Robust Multi-Array Averaging
ROC	Receiver operator characteristic
SVM	Support vector machine
TF	Transcription Factor
WHO	World Health Organization

# 1. INTRODUCTION

Cancer is a spectacularly broad family of deadly diseases. Common to all cancers is their biological cause: an error, or mutation, in the hereditary material of a cancer cell. This hereditary material, or DNA, contains a great number of genes. The genes hold the information a cell needs to produce proteins, which, via a variety of functions, form the basis of practically all cellular activity, and thus, life itself. As proteins may affect each other's production, or expression, they form a regulatory network whose complexity transcends human comprehension by far. When the DNA of a cell becomes mutated, there is a small chance of this gene-regulatory network going awry. A change in the expression of a single gene may cascade through the network, changing the entire expression profile of the cell. If this abnormal cell happens to have enhanced replicative potential, the cell might give rise to a tumor, or mass of cancerous cells. This horde of recklessly replicating monster cells threatens its well-behaving neighbor cells, and, ultimately, the life of the individual itself. [1; 2]

If a gene-regulatory network is viewed as a dynamical system, then the different cell types can be seen as attractors of that system, or expression states to which the cell tends to even if slightly disturbed. A cancer-causing mutation, in turn, re-models the system so that its attractor changes to a state enabling limitless replication. [3] Studying and modeling these expression states requires measuring the expression profiles of normal tissues and tumors. DNA microarrays have been widely used for nearly twenty years in measuring the expression levels of all known genes in a tissue sample. As most peer-reviewed journals require the authors of scientific articles to submit their measurement data to public, curated repositories, a great amount of gene expression data has accumulated over the years, accessible to anybody interested in down-loading it. [4]

The work presented in this thesis involves integrating a set of 9,544 gene expression measurements from blood and lymph node cancers and normal blood cells. The data set consists of measurements from hundreds of separate studies with a range of possible study-dependent sources of bias. Thus, a major challenge of the work is to integrate the data to enable multiple-study analyses and to validate that the variation in the data resource is explained primarily by the biological phenotypes it is supposed to represent and not by technical artifacts.

Normalization and the subsequent down-stream analysis of the expression data

set requires applying computational methods developed within the fields of bioinformatics and machine learning. Bioinformatics is a rapidly evolving interdisciplinary domain of research providing the field of biology with methods of applied mathematics and computer science. The rise of massively parallel — or high-throughput — measurement technologies has turned modern biology into a heavily data-intensive discipline, fueling the development of elaborate bioinformatic tools and resources. Machine learning, on the other hand, is a field studying algorithmic learning from data, or knowledge-mining. As learning is a quintessential feature of human intelligence, so is machine learning an important sub-discipline of artificial intelligence. Machine learning has proved a source of valuable approaches in making sense of the ever growing heaps of data biologists face today. Therefore, one of the roles of bioinformatics is finding the best machine learning algorithms to answer biological questions, and if necessary, developing novel algorithms for the task. [5]

In this thesis, we first walk through the very basics of molecular and cancer biology, gene regulatory systems, gene expression microarrays and machine learning. After introducing the theory behind the work, we present the data and methods used to analyze it. Finally, we divide the results section into four parts, the first three of which involve standard machine learning approaches to organize and interpret the data at various biological scales, or subsets of phenotypes. In the last part we present a novel pan-hematological view of the regulatory divergences between cancers and normal cells intended for generating new hypotheses of both normal and malignant hematopoiesis.

This thesis is in part inspired by that of Heinäniemi *et al.* in which TF pairs with reversal patterns between hematopoietic lineages were uncovered from partly the same gene expression data used in this work [6]. Here, however, the scope is broadened to cover also malignancies of hematopoietic origin. A preliminary expression analysis with the data set used in this thesis has been carried out to identify known and novel expression patterns of small nucleolar RNAs across acute leukemias. [7]

## 2. THEORETICAL BACKGROUND

### 2.1 Cell

A cell is a fundamental building block of any living organism. The smallest of species constitute a single cell, while larger ones, such as humans, can be composed of a number of cells far beyond our comprehension, perhaps as much as a million billion distinct cells. Common to all cells, regardless of species, are certain components. These include the membrane which encloses the cell and holds it together, the cytoplasm which fills the space inside the membrane, as well as the hereditary material of the organism.

One of the most profound distinctions between organisms is based on the type of cells of which they are composed. Prokaryotes, comprising all bacteria and archaea, are mainly unicellular and lack organelles, membrane-enclosed subcomponents of the cell which are present in the other group, eukaryotes. The most notable distinguishing feature in eukaryotes — which include all animals and plants — is the nucleus, an organelle housing the genetic material. [8] In the context of this thesis, we will use the term "cell" to refer specifically to the eukaryotic cell, if not otherwise mentioned.

In a multicellular organism all cells originate from a single stem cell which has split into two daughter cells in a process called cell division. These daughter cells have further divided as have their daughter cells and so on — eventually forming a vast pedigree of cells. In this family tree of cells within an individual organism there are multiple lineages in which the cells have differentiated into distinct types. In a human there are hundreds of cell types, each carrying out its own tasks required to keep the individual thriving. Neurons, for instance, propagate electric signals around our body, whereas red blood cells, or erythrocytes, hold the responsibility of delivering oxygen from our lungs to muscles. [8; 9]

The relative amount of different types of cells in the human body is regulated by a complex system involving various signaling molecules to promote or reduce the replication and differentiation of the required cells. For example, low blood oxygenation increases the amount of a hormone called erythropoietin which, in turn, stimulates the production of erythrocytes. In addition to this regulation *between* cells, each cell has an *internal*, intricate regulatory network of protein production. The production of proteins forms the basis of nearly all cell functionality. [9] To

understand what it is about, let us first take a closer look at the aforementioned genetic material.

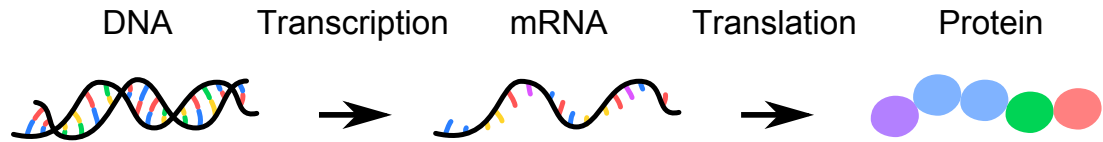
## 2.2 DNA

As mentioned, cells have a fascinating capability to differentiate and specialize in diverse, and often rather complicated, functions. The surprising complexity of these superficially humble entities rises from the vast amount of genetic information found in the cell together with the elaborate network regulating the utilization of that very information. The hereditary material is deoxyribonucleic acid, commonly known as DNA. A DNA molecule is composed of two coiled chains (or strands) of nucleotides, and each nucleotide, in turn, contains one of the following nucleobases: guanine (denoted by G), adenine (A), thymine (T) or cytosine (C). The two DNA strands have complementary sequences: at every position one strand has a adenine, the other one has a thymine and wherever there is a cytosine, the other strand has a guanine at the corresponding position. The hereditary information is encoded in the order of the four bases in the DNA. The entire genetic code of an organism, or the genome, can thus be represented as a sequence of these four letters. The genome consists of approximately 3.2 billion bases, or 3.2 million kilobases (kb). [8]

In eukaryotes, the DNA is organized as several densely packaged long molecules called chromosomes. The packaged chromosomes contain proteins and protein complexes around which the DNA is wrapped. Collectively the DNA and its attached molecules are known as chromatin. Humans normally have two pairs of twenty-three different chromosomes, i.e. forty-six chromosomes altogether. One pair of the chromosomes is sex-specific; females have a pair of X chromosomes, while males have one X and one Y chromosome. Each chromosome contains up to 2000 basic units of heredity called genes. There are an estimated number of 20,000 to 25,000 genes altogether in our genome, each constituting a stretch of DNA anywhere in the range of 0.2 kb to 2,500 kb. Each gene functions as a "recipe" for a functional gene product, as we will learn in the next chapter. [9]

## 2.3 Gene expression

The information encoded in our DNA has biological relevance only if it is interpreted. In other words, the genes must be expressed. In most cases, the result of gene expression is a protein specific to the gene in question. The genetic information of DNA is conveyed by ribonucleic acid, or RNA, to the protein level, in which it determines the shape, and ultimately, the function of the protein. The protein of each gene has its own function as a reaction catalyst, signaling unit or building block of cellular structures, to name just a few instances. The flow of information from



Kuva 2.1: The central dogma of molecular biology. Genetic information of the DNA is conveyed via mRNA to the protein.

DNA to protein via mRNA, known as the central dogma of molecular biology, is illustrated in figure 2.1. [8]

The first stage of gene expression is known as transcription, in which the sequence of the gene is read and a copy of it is produced in the form of RNA. The molecule responsible for the synthesis of RNA is known as RNA polymerase. Transcription begins as RNA polymerase attaches to the promoter region of the gene, located next to its actual coding part. For RNA polymerase to attach to the DNA, the promoter area must have an open chromatin structure and specific kind of proteins known as transcription factors (TF) bound to it. Once the RNA polymerase has bound to the promoter, it begins to slide down the DNA forming an RNA molecule complementary to one of the DNA strands. Thus the RNA is identical in sequence with the other strand, with one exception: in RNA, thymine is substituted by uracil (U). [8]

In the case of some genes, the RNA resulting from transcription is a functional unit with tasks varying from transferring other molecules to forming RNA-protein complexes with a range of complicated functions. Often, however, the RNA molecule is merely an intermediary product, called messenger RNA (mRNA), in the process of protein synthesis. In such cases the RNA is translated into a protein on a cellular machine known as a ribosome. By the ribosome, a protein is formed by connecting amino acids into a long polypeptide chain, the sequence of which is fully determined by the RNA sequence. There are 20 different kinds of amino acids, and each amino acid in the peptide chain corresponds to a codon, or set of three nucleotides in the RNA (and DNA, for that matter). Thus, the original coding sequence of the gene in the genome has three times the amount of nucleotides than there are amino acids in the final gene product, the protein. [8; 9]

Gene expression is how the genetic makeup, the genotype, manifests as observable traits, the phenotype. However, the presence of an identical genotype does not necessarily lead to the same phenotype in each cell. This is due to gene regulation. Some genes are only expressed in a specific cell type or phase in the life of a cell. The massive orchestration of turning cells "on" and "off" allows for cells to differentiate to highly specified tasks. For example, only certain cells in the pancreas produce a protein called insulin, whereas beta-actin, a protein crucial in providing the cell

with structure, is expressed in nearly every type of cells. [8]

The machinery of gene regulation involves a diversity of transcription factors and proteins affecting the state of the chromatin. The abundance of such gene products affects the expression of other genes, some of which may have regulatory roles as well. Therefore, a change in the expression of a single gene may cascade through the network of genes, changing the whole expression profile of the cell. This notion of an expression profile, the abundance of each type of protein (or the corresponding mRNA) expressed in a cell is one of the most important ones in this thesis. We will approach it from a more technical viewpoint later on. But now, acquainted to the very basics of cell biology, we are prepared for the story of the most intriguing disease plaguing humankind ever since Adam made the definitive approach towards Eve.

## 2.4 Cancer

Fine-tuned by billions of years of evolution, the inter- and intra-cellular regulatory networks have become fairly robust to a range of disturbances. Yet sometimes — very rarely in fact, considering the amount of cells in a human being — a single cell may slip out of the control of its peers and begin to replicate recklessly, producing a tumor, a mass of monster cells of its kind. This bulk of malignant, or cancerous, cells threatens the life of its well-behaving neighbors, and, as a result, the life of the individual itself. And indeed cancer does pose a threat to all of us: during the lifespan of a modern human, there is a probability of over one-third to be diagnosed with some type of cancer and a majority of the patients eventually succumb to the disease.

Why does cancer still pester us, despite of being perhaps the most studied subject in recent biomedical research? To answer this question one must first learn to appreciate the complexity of the disease. The complexity stems from 1) the immeasurable distinct molecular causes of cancer 2) the intricacy of the regulatory networks it disturbs, and 3) the evolutive nature in which tumors develop to resist treatment and metastasize, or spread to distinct locations in the body. [2]

In the context of cell biology, the cause of cancer is most often a mutation. Mutations are events in which the nucleotide sequence of the DNA, i.e., the genetic code, in a cell is altered. This happens when a nucleotide is substituted with a different one, or a segment of DNA is deleted from or inserted to a chromosome. Mutations occur both due to replication errors when the chromosomes are copied before a cell division as well as by the influence of certain chemicals or ionizing radiation.

Normally, a mutation does little or no harm. Cells have a DNA repair machinery which is often able to fix minor mutations. If the DNA damage is beyond repair,

the cell may commit to apoptosis, a programmed cell death, or "suicide", in order not to cause damage to its neighboring cells by possible misbehavior. But even if the cell fails to repair its altered DNA or perform apoptosis, it is very unlikely a randomly located small mutation is of any significant consequence, as the majority of DNA does not code for any protein or have any other apparent crucial function. If a mutation, however, occurs in a region coding for a protein or with regulatory significance, such as a transcription factor binding site, it has a potential of affecting the structure (and function!) of the protein or the rate in which it is produced. [2]

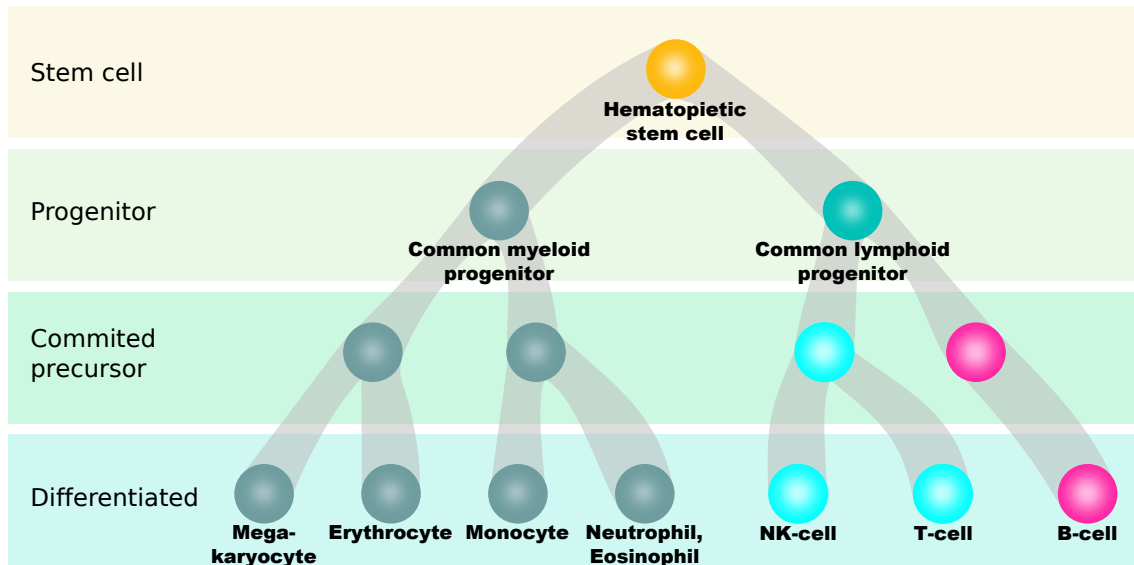
A decisive step toward cancer is taken if a mutation causes changes which lead to enhanced cell growth or resistance to apoptosis. An increase in the growth rate causes the DNA to be replicated at a higher rate, increasing the risk of further mutations. Also, in a population of frequently replicating cells, the ones with a highest growth rate are selected in an evolutive manner. Evading apoptosis, likewise, predisposes the cell to further DNA damage. It is easy to see that these traits form a loop of positive feedback: cancer-like properties cause even stronger cancer-like properties. This vicious circle may last for years as mutations with malignant potential accumulate in a cell in an accelerating rate. Eventually, when the tumor attains the capability to metastasize, forming new, separate daughter tumors, it is considered fully malignant and critically life-threatening. [2; 10; 11]

Cancer may arise in nearly any tissue, but some cells are more prone to uncontrolled growth than others. Certain tissues, including epithelia and blood, experience a high level of normal wear and tear, and must therefore renew their cells at a steady pace. Tissue renewal requires stem cells, which maintain the capacity to replicate and differentiate to a number of mature cell types needed in the tissue. This very replicative potential, while crucial in keeping us alive, perversely holds the door ajar for cancer. In the rest of the thesis, we shall restrict ourselves to malignancies of two related tissues undergoing constant self-renewal: blood and lymphoid tissue.

## **2.5 Hematological malignancies**

### **2.5.1 Hematopoiesis**

Blood, the "highway traffic" of our body, bears a close developmental relation to lymphoid tissue, the "defense bases" housing mainly immune cells. The cells comprising these two tissues descend from a common cellular ancestor residing in the bone marrow, the hematopoietic stem cell, or shortly, the HSC. The formation of new blood and immune cells from HSCs, known as hematopoiesis, is presented in the diagram of figure 2.2. To maintain a constant HSC level in the bone marrow, half of the daughter cells produced by HSC replication must remain stem cells and the other half is engulfed in a multi-step process of cell differentiation. [8]



Kuva 2.2: Hematopoiesis. New blood cells are formed as hematopoietic stem cells divide and differentiate into mature cells of the blood and the immune system.

Differentiation of a blood cell can be modeled as a journey beginning from a HSC and ending at a terminal branch in the family tree of blood cells, representing a mature hematopoietic cell. As the differentiation progresses, the cell's replicative potential decreases while the hallmarks of its cellular destiny gradually strengthen. An intermediary cell between a stem cell and a mature cell is called a progenitor or a blast.

As figure 2.2 illustrates, the hematopoietic differentiation contains several points in which the cell makes a "decision" between two lineages. Both stochastic events and interleukine-mediated control of hematopoiesis determine the lineage into which a cell ends up. The first major branch is between myeloid and lymphoid cells. Cells of the myeloid lineage differentiate further into erythrocytes, thrombocyte-forming megakaryocytes, monocytes and myelocytes. The lymphoid lineage produces mainly lymphocytes, central units of the immune system. [9]

Cancers arising from hematopoietic cells are called hematological malignancies, further divided into leukemia, myeloma and lymphoma. They all are cancers of blood cells, but only the first two manifest primarily in blood while the latter is a cancer of lymphoid tissue (especially lymph nodes). Leukemia and myeloma, therefore, are liquid tumors while lymphoma is solid. All hematological malignancies, however, may cause complications of both the circulatory and immune system. [2]

### 2.5.2 Molecular pathology

A common feature of hematological malignancies is a certain type of mutation, a reciprocal chromosomal translocation (RCT). In chromosomal translocations, two chromosomes swap a part. A notable example of this is the so called Philadelphia chromosome, which is caused by a translocation between chromosomes 9 and 22, denoted by  $t(9;22)$ . A translocation might connect the coding areas of two genes, creating a fusion gene. The product of this novel gene may behave in deleterious ways or be non-functional. The fusion gene is regulated by the regulatory machinery of only one of the fusion partners, causing domains of the other partner being expressed at an aberrant level. This minor abnormality may have major consequences at a cellular level via the regulatory network of genes. [2]

In blood cancers, chromosomal translocations are important driver mutations, meaning they increase the cancer-like properties of the tumor, and are even implicated as the pivotal event causing the cancer in many instances. In the aforementioned case of the Philadelphia chromosome, the translocation fuses genes *ABL1* and *BCR*, resulting in a fusion gene *BCR-ABL*. Normally, *ABL* produces a protein with important signaling roles in cell differentiation and division. Fused with *BCR*, it is an oncogene, cancer causing gene, by increasing replication and preventing maturation of hematopoietic blast cells. This results in a blast crisis, in which the blood is flooded with an increasing amount of immature blasts, impairing the circulatory and immune systems, and ultimately, if untreated, to death.

Nearly all chronic myelocytic leukemias and some acute lymphocytic leukemias are caused by the fusion of *ABL1* and *BCR*. Incidentally, a precision drug called imatinib has been found to inhibit BCR-ABL and, consequently, prevent blast crisis. Although precision medicines are lacking for most cancers, knowing the genetics behind the disease often has at least some therapeutical relevance. For this reason (and because the cost for genetic tests has radically decreased), the presence of a specific translocation is useful in defining subtypes of blood cancers, as we shall see in the next section. [2]

### 2.5.3 Classification

There is a vast range of malignancies arising from hematopoietic cells, with differences from both biological and clinical viewpoints. A detailed, systematic classification scheme is therefore necessary to ease diagnosis and treatment of these neoplasms, as well as their research. The main, coarse-level classification, as mentioned previously, divides hematological malignancies into the following three categories [12]:

1. leukemia, arising from precursor blood cells and manifesting in bone marrow

and blood

2. lymphoma, arising from lymphoid cells and manifesting primarily in lymphoid tissue and
3. myeloma, arising from mature plasma B cells and manifesting in bone marrow and blood. (As a lymphoid malignancy, myeloma is often classified as a type of lymphoma.)

To further divide these main classes into meaningful subtypes, several measurable attributes have been used, including [12; 13]

1. clinical features, such as where and how the disease manifests
2. morphology, i.e. what the cancer cells look like under microscope
3. immunophenotype, defined by the protein content of the cancer cell surface membrane (i.e. surface markers) and
4. cytogenetics, especially chromosomal aberrations present in the cancer cells.

Of these attributes, the clinical features have been traditionally the easiest to detect and, naturally, most relevant to therapy. Morphology is also relatively easy to observe and, thus, has been used extensively to classify blood cancers. However, its relevancy in any respect is questionable [14]. Immunophenotyping, while more costly and laborious than microscopic examination, is shown to reveal — to some extent — the lineage and maturation level of the cell from which the cancer is originated. Indeed, the presence of surface markers has been used in both diagnosis and study of blood cancers. [15]

Cytogenetical measurements of chromosomal aberrations, such as translocations or abnormal numbers of chromosomes, are useful in studying the cause of a disease. However, until the advent of precision drugs about fifteen years ago, they have been of limited clinical use. Now, as genetic measurements are commonplace in cancer research and precision medicines are being sought after, the relevance of genetic features as disease subtype classifiers has been acknowledged. Furthermore, not only chromosomal abnormalities have been shown to explain the disease and be useful in planning personalized treatment, but also more subtle differences in the genome, transcriptome, proteome and epigenome as well. These so called molecular markers of disease have been a central topic in cancer research for the recent years. [16; 17; 18]

## Leukemia

Within leukemia, there are two major distinctions between subtypes of the disease. The first one, based on the main hematopoietic lineage giving rise to the cancer, divides leukemias into myeloid and lymphocytic types. The second distinction, a clinical one, divides them into acute and chronic diseases. Combined, these two distinctions yield four main categories of leukemia: Acute myeloid leukemia (AML), Chronic myeloid leukemia (CML), Acute lymphocytic leukemia (ALL) and Chronic lymphocytic leukemia (CLL).

The acute forms of myeloid and lymphocytic leukemia are typically caused by a low number of mutations (often one single translocation) in hematopoietic progenitor cells and they progress fast. Chronic leukemias, on the contrary, often take years to develop, cumulating a higher number of diverse mutations in typically slightly more mature blast cells than their acute counterparts. This is reflected in the fact that acute leukemias are far more common in children than chronic leukemias. Also, in the case of acute leukemias, translocation-based classification is more relevant as the chromosomal aberrations explain the disease to a longer extent. [2; 17]

Acute leukemias were previously classified using a morphology-based French-American-British (FAB) classification dating originally from 1976 [19]. As the clinical importance of chromosomal aberrations became apparent, the World Health Organization (WHO) released its own translocation-based classification in 2001, which was updated in 2008 [13; 20]. The WHO classification of AML is shown in table 2.1 and that of ALL in table 2.2. In the tables, the translocations are listed with their corresponding fusion protein. In the WHO classification of AML, the previously used FAB classes are included as provisional subtypes within AML cases which are not otherwise specified.

Taulukko 2.1: 2008 WHO classification of acute myeloid leukemia (AML). [20]

<b>AML with recurrent genetic abnormalities</b>
AML with t(8;21)(q22;q22); RUNX1-RUNX1T1
AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); CBFB-MYH11
AML with t(15;17)(q22;q12); PML-RARA
AML with t(9;11)(p22;q23); MLLT3-MLL
AML with t(6;9)(p23;q34); DEK-NUP214
AML with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); RPN1-EVI1
AML (megakaryoblastic) with t(1;22)(p13;q13); RBM15-MKL1
Provisional entity: AML with mutated NPM1
Provisional entity: AML with mutated CEBPA
<b>AML with myelodysplasia-related changes</b>
<b>Therapy-related myeloid neoplasms</b>
<b>AML, not otherwise specified</b>

Taulukko 2.2: 2008 WHO classification of acute lymphocytic leukemia (ALL). [20]

---

---

### **Precursor B-ALL**

Precursor B-ALL, NOS

Precursor B-ALL with recurrent genetic abnormalities

Precursor B-ALL with t(9;22)(q34;q11.2);BCR-ABL 1

Precursor B-ALL with t(v;11q23);MLL rearranged

Precursor B-ALL t(12;21)(p13;q22) TEL-AML1 (ETV6-RUNX1)

Precursor B-ALL with hyperdiploidy

Precursor B-ALL with hypodiploidy

Precursor B-ALL with t(5;14)(q31;q32) IL3-IGH

Precursor B-ALL with t(1;19)(q23;p13.3);TCF3-PBX1

### **T-ALL**

---

---

## **Lymphoma**

Lymphomas, as solid tumors, differ from leukemias in many ways in respect to classification. Unlike in leukemias, the anatomical location and tissue harboring the tumor is an important attribute used to define subtypes of lymphoma. Also, because solid tumors typically require a longer time to accumulate cancer-promoting mutations, the genetic makeup is more complex, decreasing the relevancy of a translocation-based classification as in acute leukemias. Table 2.3 lists the main categories of the 2008 WHO lymphoma classification, mainly based on the sub-lineage within the main lymphoid lineage [21]. Most categories harbor a large number of subtypes differing in various clinical, pathologic, or biologic features. In the WHO classification scheme, multiple myeloma is listed as a subtype of mature B-cell neoplasms. [22]

Taulukko 2.3: 2008 WHO classification of the main types of lymphoma. [21]

---

---

Mature B-cell neoplasms

Mature T-cell and NK-cell neoplasms

Hodgkin lymphoma

Histiocytic and dendritic cell neoplasms

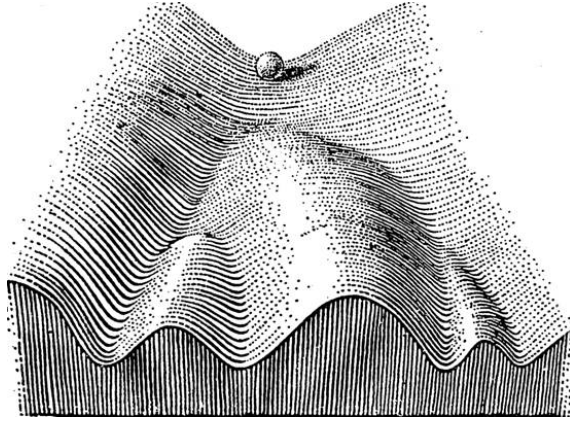
Posttransplantation lymphoproliferative disorders (PTLDs)

---

---

## **2.6 Gene regulatory systems**

In 1957, Conrad Waddington presented his well-known epigenetic landscape of cellular differentiation, pictured in figure 2.3 [23]. The landscape is a metaphor of gene regulation-mediated differentiation of the cells of an organism, in which a ball rolling down a grooved hill represents the cell as it matures. At the top of the hill, the ball represents a pluripotent stem cell, but as it rolls down the landscape, it ends up in a groove (developmental lineage) and further sub-grooves (sub-lineages),

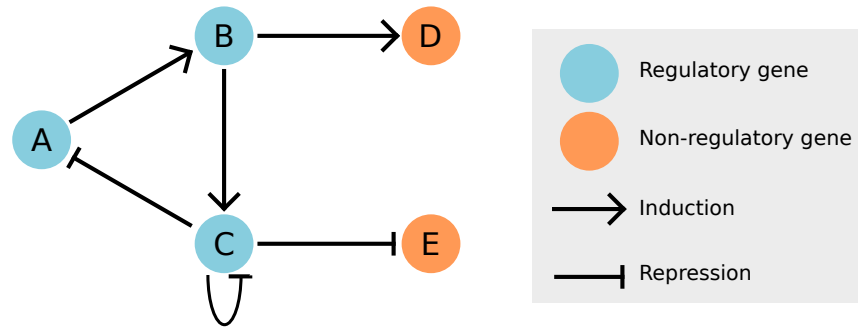


Kuva 2.3: Waddington's epigenetic landscape. The ball rolling down the grooved hill represents a differentiating cell. The valleys at the root of the hill correspond to distinct cell types. [23]

finally stopping at the root of the hill in a valley representing a specific, fully mature cell type. With the word epigenetic, Waddington refers to all gene-regulatory activity defining the phenotype as opposed to the actual genetic information, or DNA-sequence, which is assumed to stay unaltered during differentiation. Later, epigenetics has been defined more strictly as heritable modifications of a chromosome not affecting the DNA sequence. For convenience, however, we will stick here to Waddington's broader definition of the term. [24]

Besides being an illustrative metaphor of the complex system giving rise to multiple, distinct cell types in spite of identical hereditary material, Waddington's landscape model does not defy exact formulation. One of the first models of epigenesis, filling the causative gap between DNA and the diverse range of distinct cell types it enables, was proposed in 1967 by Stuart Kauffman. Introducing random Boolean networks (RBN) as discrete, dynamical models of gene regulatory networks (GRN), Kauffman suggested that different cell types are attractors, or stable states, of the network. The concept of an attractor is central in the theory of dynamical systems. In a dynamical gene network it corresponds to a gene expression state to which the system tends to even if slightly disturbed. A sufficiently large perturbation, however, may cause the system to drift to another attractor. In Waddington's landscape, the attractors are the valleys and an attractor-switching perturbation would be one which pushes the ball over the hill separating two valleys. [26]

Perturbations of the GRN may be due to impulses from the environment as well as the inherent stochasticity of inter- and intracellular regulation. During the process of differentiation, randomness plays an important role in defining cellular fate [25]. However, once the cell is terminally differentiated, robustness to disturbances is desirable to maintain the current state, or gene expression profile. Therefore, GRNs



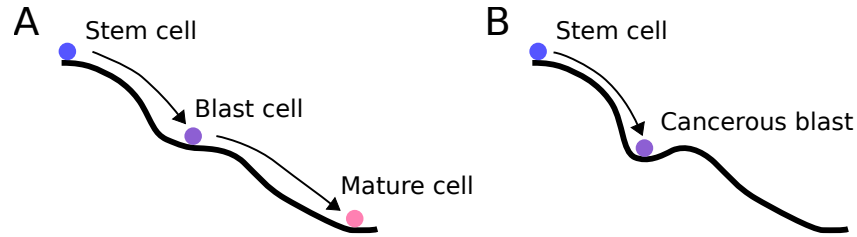
Kuva 2.4: A gene regulatory network. Genes A and B induce other genes, or increase their expression. C, on the other hand, is a repressor: it decreases the expression of genes, including itself. By self-repression, C regulates its own expression level. A regulates its own expression level, too, but indirectly by inducing a gene (B) which induces its repressor (C). Genes A, B and C form the regulatory core of this GRN while D and E merely express the effects of their regulators.

wired such that their attractors are sufficiently stable enjoy a selective advantage. Essential for any self-regulative system are negative feedback loops. Negative feedback ensures that changes in the state of the system are moderated by adversary effects. Gene regulatory-wise, this means that GRNs must contain self-repression. [28]

Figure 2.4 shows a small GRN. Three of its five genes have regulatory functions: they induce or repress other genes (or themselves). The regulatory genes form the core of the GRN as changes in their expression affects the expression state of the whole system. Non-regulatory genes have no effect on the expression profile outside their own expression. In a real-life GRN, the regulatory core composes of TFs and other genes with regulatory effects. Two types of control circuits are presented in figure 2.4: direct and indirect self-repression. Both contribute to the stability of the expression profile. [1]

Any change in the wiring of a GRN may change its attractor. Biologically, a mutation affecting any regulatory connection of the GRN is such a change. As the system changes, so does its epigenetic landscape. Thus, cancer can be seen as a re-wired GRN which leads to new attractors being formed, corresponding to gene expression states enabling limitless replication. Figure 2.5 illustrates the case. The normal maturation process is halted by a new attractor which retains the replicative potential of the cell. In leukemia, for instance, such an attractor-state is the cause of blast crisis. This model, thus, places tumorigenesis in the context of cellular differentiation, only with an altered epigenetic landscape. [28]

So far, modeling a genome-wide GRN has been a daunting task. Uncovering all the regulatory connections and quantifying their degree of inhibition or repression, not to mention their combinatorial effects, requires a great amount of biological research.



Kuva 2.5: Cellular maturation. In cancer, normal differentiation (A) is halted by rewiring the gene regulatory network, forming a new, cancerous attractor in the gene expression space (B).

Therefore, systemic properties of whole-genome regulatory models have been studied with GRNs making very strong simplifying assumptions, such as binary expression states and random regulatory effects in RBNs. Another approach is to restrict the GRN to a small sub-network and attempt to model its regulatory connections in detail using differential equations. This, of course, enables analyzing only a fraction of the functionality of the entire system. [27]

## 2.7 DNA microarrays

### 2.7.1 Technology

For the last fifteen years, the most widely used technology to measure genome wide features from tissue samples or cell cultures has been the DNA microarray. Most importantly, microarrays allow measuring the expression levels of all genes in the genome in parallel at a reasonable cost. Whole-genome expression measurement yields a snapshot of the expression state and can be used to characterize and study the mechanisms of intracellular processes, both normal and abnormal. [29; 30]

DNA microarrays are small chips with DNA oligonucleotides called probes attached to its surface. In a microarray experiment, complementary DNA (cDNA), obtained by reverse-transcribing RNA from a sample, is labeled by laser-excitable cyanine dyes. Then it is hybridized to the probes on the chip. Un-hybridized cDNA fragments which do not match any probe, are washed out. Finally, the amount of cDNA hybridized to the probes is measured by detecting the emission intensity of the labels when disposed to laser of the dye-specific wavelength. Some microarrays use a two-channel method in which two separate samples (disease versus healthy control, for example) are hybridized to the same chip with separate labels. In this thesis, however, we focus on analysis of data measured with single-channel chips in which one chip per sample is used. [31]

Whole-genome expression microarrays are highly parallel, or high-throughput,

measurement platforms. For this reason, the analysis of the results requires algorithms and computation, beginning from the very normalization of the raw intensities detected from the chip. A multitude of elaborate down-stream analyses of gene expression data have been developed during the years, including enrichment analyses, time-series analyses, and several approaches harnessing the full power of supervised and unsupervised machine learning. [30; 31]

### 2.7.2 Data pre-processing

In any microarray experiment, the raw intensity measurements should be subjected to quality control and normalization to address technical biases and artifacts of various types. Additional platform-dependent pre-processing steps may also be necessary in transforming the probe intensities to gene expression values. Several algorithms have been developed by technology providers and academia with various approaches to normalization and other pre-processing steps. For Affymetrix GeneChips arrays, the most popular pre-processing algorithms are Microarray Suite 5.0 (MAS 5.0) and Robust Multi-array Average (RMA) [32; 33; 34].

#### Preliminary quality control

Before normalization, it is appropriate to carry out preliminary quality control for the raw intensities of microarray experiments in order to rule out experiments failed due to a technical or human errors. Data should be deemed unusable if a high number probe intensities are saturated at the maximum value or totally undetected. In addition, if the distribution of the raw intensities appears irregular (for example, multimodal distribution), it is reasonable to consider the experiment failed. If the study comprises several microarray experiments, the distributions of intensities can be compared by a visual inspection of box plots, for example, to detect outlier distributions. [30]

#### Background correction

Most pre-processing algorithms perform background correction, i.e., noise removal, to the raw intensities before normalization. Background correction is based on the assumption that the intensity values can be represented as a sum of the true signal and noise. In MAS 5.0, the noise is estimated from the 2 % of probes with lowest intensities in a local manner [32]. In RMA, the intensity  $s$  is assumed to be composed of the true signal  $x$  and noise  $y$ . Signal  $s$  is modeled using a random variable  $S$ :

$$S = X + Y, \tag{2.1}$$

in which the true signal is modeled as an exponential distribution  $X \sim \text{Exp}(\alpha)$  and the noise as a normal distribution  $Y \sim N(\mu, \sigma)$ . This allows the expected value of  $X$ , given  $s$  to be found by

$$E(X|S = s) = a + \sigma \frac{\phi(a/\sigma)}{\Phi(a/\sigma)}, \quad (2.2)$$

where  $a = s - \mu - \alpha\sigma^2$ ,  $\phi$  is the probability density function of a normal distribution and  $\Phi$  is its cumulative density function. The parameter values are estimated using all of the intensities on the array. RMA, thus, does not assume location-dependent noise unlike MAS 5.0. [33]

### Normalization

The objective of normalization is to remove the so called chip-effect, or the batch effect between multiple microarray experiments. Without normalization, comparison studies involving several chips are likely to suffer from the chip-effect. All approaches to cross-chip normalization attempt to reduce the difference in intensity distributions between arrays. In MAS 5.0, the data is scaled so that the mean (excluding the lowest 2% of values) of each array is equal. On log-scale, the scaling corresponds to shifting without affecting the shape of the distribution. [32] RMA uses a stronger method called quantile normalization, which forces the distributions of each chip to be equal. The steps in quantile normalization are:

1. Represent the  $d$  intensities of  $n$  arrays as a matrix  $\mathbf{I} \in \mathbf{R}^{d \times n}$
2. Obtain  $\mathbf{I}^s$  by sorting  $\mathbf{I}$  column-wise
3. Calculate  $\mathbf{m} \in \mathbf{R}^d$ , the row-wise mean of  $\mathbf{I}^s$ :  $\mathbf{m}_i = \frac{1}{n} \sum_{j=1}^n \mathbf{I}_{i,j}^s, \forall i \in [1, d]$
4. Obtain the quantile-normalized matrix  $\mathbf{I}^{qn}$  by replacing each element in the unsorted  $\mathbf{I}$  by  $\mathbf{m}_i$ , where  $i$  is the row index of the corresponding element in the sorted matrix  $\mathbf{I}^s$

Assuming the ideal, unbiased data are from equal distributions, quantile normalization is a more powerful approach for cross-chip studies. It is, however, vulnerable to outliers in data, making preliminary quality control necessary. [33]

### Correction for nonspecific hybridization

To address the issue of nonspecific hybridization, i.e. cDNA fragments with non-complementary sequences hybridizing to a probe, Affymetrix GeneChips have a mismatch (MM) probe for each normal probe, denoted perfect match (PM). The MM probe is equal to the PM probe except for the base in the middle of the probe,

which is replaced by its complementary base. The signal intensity at MM probes are supposed to reflect the amount of nonspecific hybridization of the corresponding PM probe. MAS 5.0, therefore, robustly estimates the level of nonspecific hybridization from MM intensities and subtracts them from the PM intensities. However, using the MM intensities has proven problematic in several ways, which is why RMA, among other methods, ignores the MM probes altogether and, thus, does not attempt to correct for nonspecific hybridization. [32; 33]

### Summarization

In Affymetrix GeneChip arrays, multiple probes (a probe set) are used in measuring expression of a single gene. Therefore, a separate summarization step is required to transform the probe intensities into gene expression values. Both MAS 5.0 and RMA use robust estimations of probe set medians. MAS 5.0 uses a method called Tukey's bi-weight while RMA uses median polish. In median polish, the  $k$  probes belonging to the same probe set are used to obtain gene expression values as follows:

1. Represent the  $k$  probe intensities of  $n$  arrays as a matrix  $\mathbf{P} \in \mathbf{R}^{k \times n}$
2. Initialize  $\mathbf{P}^*$  to  $\mathbf{P}$
3. Update  $\mathbf{P}^*$  by subtracting the row-wise median from its columns
4. Update  $\mathbf{P}^*$  by subtracting the column-wise median from its rows
5. Repeat steps 3. and 4. until the medians converge to zero or maximum number of iterations (five by default) is reached
6. Obtain the median-polished matrix  $\mathbf{P}^{\text{mp}}$  by subtracting  $\mathbf{P}^*$  from  $\mathbf{P}$
7. Obtain the probe set expression value for the  $i$ th array by taking the mean of the  $i$ th column of  $\mathbf{P}^{\text{mp}}$

The significant difference between the approaches of MAS 5.0 and RMA is that Tukey's bi-weight is performed one chip at a time, whereas median polish uses information from all chips in estimating probe affinity effects. In this respect, too, RMA uses a method which is more powerful but also vulnerable for low-quality arrays. The summarization step is performed before normalization in MAS 5.0, and after in RMA. [32; 33]

### 2.7.3 Microarray databases

The rise in popularity of high-throughput measurements in biology has created a demand for means of effortless data sharing. The true power of microarray experiments has been suggested to be in multi-experiment studies. Furthermore, most

academic journals require the data used in research articles to be shared via public repositories accompanied with sufficient metadata. Several repositories are available for microarray data, the largest being Gene Expression Omnibus (GEO) curated by The National Center for Biotechnology Information (NCBI) [4]. Microarray experiments submitted to most public repositories must contain metadata complying with the Minimum Information about a Microarray Experiment (MIAME) standard created by The Functional Genomics Data (FGED) Society [35].

## 2.8 Machine learning

For decades, the cost of storing and processing data has fallen exponentially as predicted by the well-known Moore's law [36]. As a result, the amount of data gathered and piled in databases of all types has risen at an accelerating rate. In this age of amassing piles of data, there is more demand than ever for the ability to draw valuable information, or knowledge, from vast data sets. Due to the development of high-throughput biological measurement technologies, this data issue has not spared the fields of biology and medicine. DNA microarrays were one of the first technologies creating major data interpretation problems in biology and they were soon followed by the widespread utilization of massively parallel next-generation DNA- and RNA sequencing technologies [37].

Fortunately enough, before the pervasive data analysis problem hit all fields of research and business capable of producing big data, established theories of advanced data analysis had already been developed. Researchers, predominantly in computer science and applied mathematics, had created a range of algorithmic approaches to learn (i.e., acquire knowledge) from data, known commonly as the field of machine learning. Lying partially within the framework of artificial intelligence, machine learning relies on theories of statistics and optimization to develop computational methods giving humans and machines understanding of real-world problems from markedly different types of data [38].

Emphasizing its ability to discover novel knowledge from superficially uninformative data, machine learning is often viewed as knowledge mining. Yet another facet of this field is pattern recognition, as distinguishing patterns is in the very crux of all data interpretation, for both humans and machines. Recognizing patterns is especially apparent in such applications of machine learning as face or voice recognition, but it is also present in, for instance, recognizing complex patterns of gene expression, even though the patterns are not necessarily visible or comprehensible for humans. Here, in fact, lies the true power of machine learning: mimicking the human ability to learn, it extends the skill to amounts of data and complexity of patterns beyond the memory or processing capacity of a human brain. [38; 39]

Most methods of machine learning can be classified as either unsupervised or su-

pervised learning. In unsupervised learning, the data is unlabeled, meaning that the data points, or objects represented by a number of attributes, are of unknown class. This corresponds, for instance, to a set of images without any captions. The images may represent different objects, but no information of their content is provided. Yet, despite the lack of labels, the data may yield a mass of valuable information. This requires organizing and analyzing the data by methods emulating the human ability to understand and explain unprecedented input perceived by our sensory organs. The goal, simply, is to describe the data in a compact and, if possible, human-comprehensible way. Thus, unsupervised learning is an exploratory approach to data interpretation: the outcome, at best, is both useful and surprising. As opposed to its unsupervised counterpart, supervised learning requires labeled data. The objective is to learn to label the data points using their attribute values and to apply the learned ability to classify new, unlabeled objects.

### 2.8.1 Dimensionality reduction

Some times, the data encountered in machine learning is of high dimensionality. This means that the number of attributes used to represent the objects is high, for example several thousands. High dimensionality may cause several problems, including 1) inability to visualize the data, 2) excessive memory usage in storing it, 3) disproportionate processing time in analyzing it, and 4) the curse of dimensionality. The notorious curse of dimensionality is a notion based on the assumption that the number of data points (observations) sufficient to describe the patterns in data grows exponentially with the dimensionality. In the case of thousands of variables — not at all abnormal in high-throughput biology — the amount of observations needed could be astronomical. [38; 40]

Dimensionality reduction is an approach to evade the problems caused by high dimensionality by reducing the number of dimensions in a meaningful way. It relies on the notion of intrinsic dimensionality, according to which the data can be described in a lower dimensionality if it contains redundant dimensions (attributes) [41]. Fortunately, redundancy is often encountered in high dimensional data. In such cases, the dimensionality can be reduced — often to a fraction of the original — with only a marginal information loss. [38] Dimensionality reduction, thus, can be seen a method of data compressing.

A naïve approach to dimensionality reduction would be to discard the dimensions deemed uninformative by some pre-defined criteria. This is known as feature selection, if the dimensions preserved are considered features to be used in further analysis. Often, however, it is possible to combine the original dimensions in an appropriate way to reduce the dimensionality while preserving information from up to all of the original dimensions. Such approaches are methods of feature extraction.

One of the most used (and simplest) methods of dimensionality reduction is the principal component analysis (PCA). In PCA, the data is transformed by rotating the coordinates such that the first of the new dimensions, called principal components (PC), has the highest variance (and hence, information). The second PC contains the second most variance and so on. It may be appropriate to choose the number of PCs by adjusting the total amount of variance of the transformed data to, say, 90% of the original variance. If the data is to be visualized as a scatter plot, the first two principal components can be used (or three, in case of a 3D visualization).

The steps of PCA include 1) constructing a covariance matrix from the original data matrix, and 2) performing the eigenvalue decomposition of the covariance matrix. The eigenvectors sorted in descending order of their corresponding eigenvalues define the PCs. Each resulting PC is a linear combination of the original dimensions. Therefore, PCA is a linear dimensionality reduction method. If the dimensions hold nonlinear relations, a nonlinear method may be able to reduce the dimensions more efficiently. However, the theory of nonlinear dimensionality reduction as well as the interpretation of the resulting dimensions is more complicated. Due to its simplicity, PCA may be a reasonably useful method even in such cases. [40]

### 2.8.2 Distance measures

Any approach to group or classify data computationally requires an explicit or implicit means of measuring the similarity — or, alternatively, dissimilarity — between observations. As the dissimilarity of two observations is proportional to their distance in the attribute space (or at least indicative of it), these ways of determining dissimilarity are called distance measures. To define a distance measure, let the data consist of observations  $x_i$ ,  $i \in [1, n]$ . Intuitive properties for the distance between  $x_i$  and  $x_j$ , or,  $d(x_i, x_j)$ , include symmetricity [42]

$$d(x_i, x_j) = d(x_j, x_i), \forall i, j \in [1, n] \quad (2.3)$$

and non-negativity [42]

$$d(x_i, x_j) \geq 0, \forall i, j \in [1, n]. \quad (2.4)$$

These two conditions are natural defining properties of a distance measure as the concepts of asymmetrical or negative distance do not correspond to our intuitive understanding of distance. The measure is considered a distance metric, if it additionally satisfies the triangle inequality [42]

$$d(x_i, x_j) \leq d(x_j, x_k) + d(x_k, x_i), \forall i, j, k \in [1, n] \quad (2.5)$$

and reflexivity [42]

$$d(x_i, x_j) = 0 \iff x_i = x_j, \forall i, j, k \in [1, n]. \quad (2.6)$$

The conditions of triangle inequality ("the shortest distance is as the crow flies") and reflexivity ("zero distance means identical observation") are likewise rather intuitive, but in some cases, a distance measure not satisfying them might be used for faster or otherwise better performance [43].

Many machine learning algorithms require the user to select the most appropriate distance measure to be applied. The selection may significantly affect the performance of the algorithm. The main thing to take into account in distance measure selection is the type of the data. One has to know the types of the attributes (or features extracted from them). One way to classify attributes is the type of scale of their possible values. The four main scale types are [43; 44]:

1. Nominal scale. The attribute values are categories which cannot be ordered quantitatively. Example: species (*Homo Sapiens*, *Mus Musculus*).
2. Ordinal scale. The attribute values are quantitative categories, but the difference between two successive categories has no precise definition. Example: verbal expression of size (*small*, *medium*, *large*).
3. Interval scale. The attribute values are quantitative, and the difference (interval) between two values can be expressed numerically. However, the relation of two categories is not meaningful, i.e. the origin of the scale is arbitrary. Example: Celsius scale (in which value zero *does not* imply zero heat energy).
4. Relative scale. The attribute values are quantitative, and both the difference and relation between two values have meaningful interpretations. Example: Kelvin scale.

Nominal attributes are always qualitative, i.e., their value does not represent the amount of any property. The other three types are quantitative, but only attributes with interval and relative scales are numerical. Qualitative and numerical variables have separate distance measures, of which we will present some of the most important ones. [43]

### Distance measures for numerical variables

Numerical variables include both discrete and continuous variables, and the distance measures presented here apply to both types. Perhaps the most common distance measure, euclidean distance, which corresponds to physical distance based on orthogonal coordinates, is formulated for  $l$ -dimensional vectors as [45]

$$L_2(x_i, x_j) = \sqrt{\sum_{k=1}^l (x_{j,k} - x_{i,k})^2}. \quad (2.7)$$

Euclidean distance can be generalized to Minkowski distance, the  $L_p$ -norm [45]:

$$L_p(x_i, x_j) = \sqrt[p]{\sum_{k=1}^l (x_{j,k} - x_{i,k})^p}. \quad (2.8)$$

Setting parameter  $p$  to 2 yields euclidean distance, which, therefore, is also known as  $L_2$ -norm. A simpler distance measure,  $L_1$ -norm or Manhattan distance is defined as

$$L_1(x_i, x_j) = \sum_{k=1}^l |x_{j,k} - x_{i,k}|. \quad (2.9)$$

Letting  $p$  approach infinity gives the  $L_\infty$ -norm

$$L_\infty(x_i, x_j) = \max_{k=1}^l |x_{j,k} - x_{i,k}|. \quad (2.10)$$

All of the distance measures defined by  $L_p$ -norm fulfill the criteria for a metric [39].

Out of the  $L_p$ -family, other common measures for dissimilarity include correlation-based distances. They are defined as

$$d_{corr}(x_i, x_j) = 1 - r(x_i, x_j), \quad (2.11)$$

in which  $r(x_i, x_j)$  is the correlation coefficient of the two observations. The most common correlation coefficient is the Pearson's  $r$ :

$$r_{pearson}(x_i, x_j) = \frac{\sum_{k=1}^l (x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j)}{\sqrt{\sum_{k=1}^l (x_{i,k} - \bar{x}_i)^2 \sum_{k=1}^l (x_{j,k} - \bar{x}_j)^2}}, \quad (2.12)$$

where  $\bar{x}_i$  and  $\bar{x}_j$  denote the means of  $x_i$  and  $x_j$ , respectively. A more robust coefficient is Spearman's  $r$ :

$$r_{spearman}(x_i, x_j) = r_{pearson}(\hat{x}_i, \hat{x}_j), \quad (2.13)$$

in which  $\hat{x}_i$  and  $\hat{x}_j$  are  $x_i$  and  $x_j$  with their elements replaced by their ranks, i.e., integers between 1 and  $l$ . Spearman's correlation is able to reveal many non-linear correlations better than Pearson's correlation and it is more robust against outliers. The cost of robustness, however, is a decrease in statistical significance of the

coefficients as the data resolution is smaller. [40]

### Distance measures for categorical variables

A simple and popular distance measure for observations of categorical variables is Hamming distance. It is the count of attributes which have non-identical values in the two observations [42]:

$$d_{\text{hamming}}(x_i, x_j) = \sum_{k=1}^l S_k, \quad (2.14)$$

where

$$S_k = \begin{cases} 1 & \text{if } x_{i,k} \neq x_{j,k} \\ 0 & \text{if } x_{i,k} = x_{j,k}. \end{cases}$$

Hamming distance applies for any categorical attributes, including binary variables. For binary variables in which 1 signifies the presence of a property and 0 its absence, Jaccard index is a widely used similarity measure. It is defined as the ratio of shared properties to all properties present in either observation [45]:

$$s_{\text{jaccard}}(x_i, x_j) = \frac{\sum_{k=1}^l (x_{i,k} \wedge x_{j,k})}{\sum_{k=1}^l (x_{i,k} \vee x_{j,k})}. \quad (2.15)$$

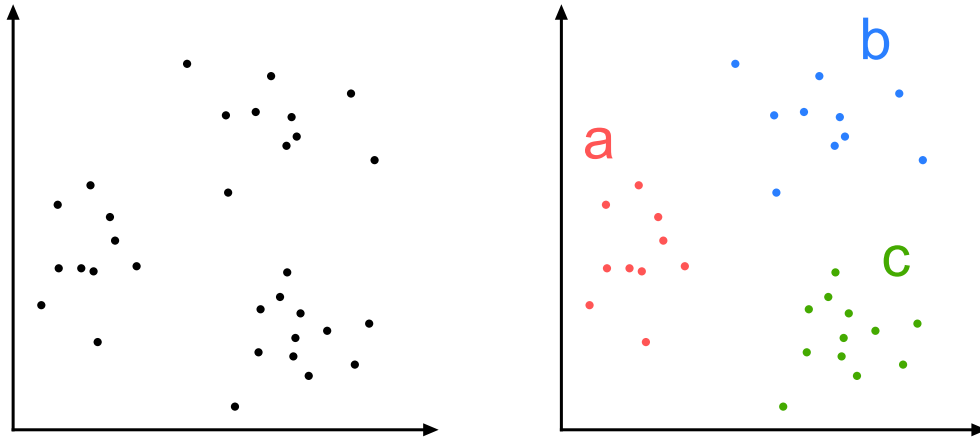
where the elements of the observations are interpreted as Boolean variables. Jaccard distance is defined as the complement of Jaccard index:

$$d_{\text{jaccard}}(x_i, x_j) = 1 - s_{\text{jaccard}}(x_i, x_j) = \frac{\sum_{k=1}^l (x_{i,k} \vee x_{j,k}) - \sum_{k=1}^l (x_{i,k} \wedge x_{j,k})}{\sum_{k=1}^l (x_{i,k} \vee x_{j,k})}. \quad (2.16)$$

Jaccard distance assumes properties present in either observation contribute to their dissimilarity, but properties absent in both observations have no effect. This is intuitive for rare properties (the absence of a rare genetic variant in a patient is a rather uninformative fact), but for common properties it works less well. [42]

### 2.8.3 Cluster analysis

In acquiring new knowledge, one important ability is grouping objects. Encountered by a set of objects unseen before, it is natural, easy and often useful for a human to



Kuva 2.6: Cluster analysis. The left panel shows two-dimensional data with three apparent groups. The right panel visualizes the result of a cluster analysis which was able to detect the groups.

divide the objects into groups based on their parent attributes: size, color, material and so on. In unsupervised machine learning, grouping unknown objects, called clustering or cluster analysis, is one of the most important ones. To detect meaningful groups among a set of observations is a central means of gaining information of the structure, or patterns, of the data. Clustering can also be seen as a data reduction method, in which the observations are replaced by artificial observations called prototypes representing the clusters. Furthermore, clustering attributes of a data set can be used to reduce the dimensionality. [43]

Figure 2.6 attempts to explain the idea of cluster analysis. The left panel of the figure shows a set of observations represented as two-dimensional data points. The two axes could correspond to numerical attributes such as height and weight. Inspecting at the data set visually, it is apparent that the points tend to three clusters. Thus, the data contains three natural clusters. The right panel of figure 2.6 shows the output of the mental clustering human brains automatically perform. Each observation has been assigned to one of the three groups, and it appears likely that these three groups represent three truly different types of objects. The objective of cluster analysis is to similarly detect the "true" clusters present in the data, especially when the dimensionality and complex cluster structure require the memory and processing capacity of a machine.

Clustering, performed by a machine, requires the process to be formulated as an algorithm. Various types of algorithms have been proposed and used in clustering, each of which have their own strengths and weaknesses. All methods, explicitly or implicitly, are designed to perform a two-objective optimization by simultaneously 1) minimizing the variance of data within clusters, or cohesion, and 2) maximizing

the difference between clusters, or separation. Furthermore, measuring the cohesion and separation requires a distance measure to evaluate how dissimilar two observations are. Most clustering methods can be classified as centroid-based, model-based, hierarchical or density-based approaches. [43; 46]

### Centroid-based clustering

In centroid-based clustering, each cluster is defined by a centroid which is the — appropriately defined — most representative (true or artificial) observation of that cluster. The most simple centroid-based clustering algorithm,  $k$ -means clustering, is perhaps the most popular of all clustering methods. In  $k$ -means clustering, the observations are partitioned into  $k$  clusters, each of which are defined by its centroid. The algorithm consists of the following steps [40]:

1. Initialize  $k$  cluster centroids in the attribute space.
2. Assign each observation to the cluster whose centroid is closest.
3. Update centroids to new cluster means.
4. Return to step 2 if any centroid moved more than predefined  $\varepsilon$ .

As the centroids converge, the partition reaches a local optimum. The popularity of  $k$ -means clustering stems from its speed and simplicity. However, a notable drawback of the method is that  $k$ , the number of clusters, must be defined before clustering. This requires more insight of the structure of the data than is usually available. One way to avoid the problem is to perform the clustering with multiple values of  $k$  and choose the partition which performs best by an objective clustering performance measure. [42; 40]

### Density-based clustering

Density based algorithms approach the clustering problem by defining clusters as data point dense-areas in the attribute space, separated by sparse areas. The most well-known density based clustering algorithm is DBSCAN (density-based spatial clustering of applications with noise). DBSCAN uses two user-specifiable parameters to assign each data point to one of the following three types [47]:

1. A cluster core point, if it has at least  $p$  other points within a distance of  $\varepsilon$ .
2. An outlier, if it has no neighbors within  $\varepsilon$ .
3. A cluster edge point, if it has at 1 to  $p - 1$  core points within  $\varepsilon$ .

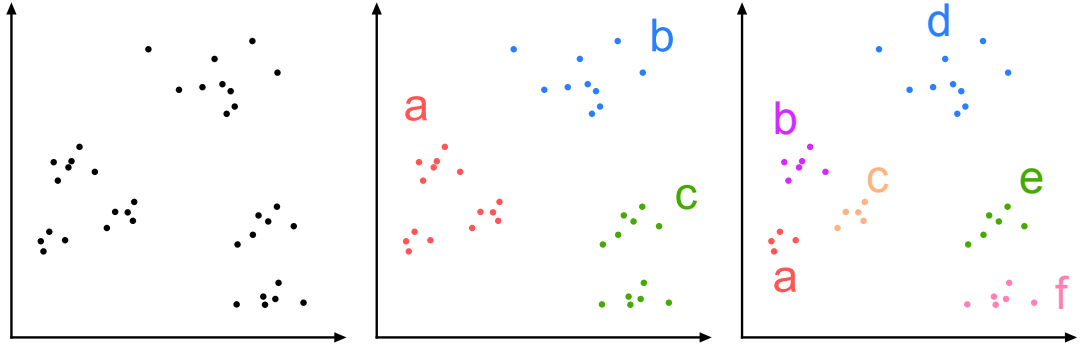
Any group of connected core points and the edge points directly connected to them is considered a cluster. The outliers are considered members of no cluster. The number of clusters, thus, emerges from DBSCAN and is not required as a parameter. Other advantages of the method include its ability to detect clusters of irregular and concave shapes and its explicit notion of outliers, making it a rather robust method. The main disadvantage, perhaps, is the need to specify the parameters  $p$  and  $\varepsilon$ . Also, detecting clusters which overlap or have variable densities has proven problematic with standard DBSCAN, leading to more adaptive — and complex — versions of the algorithm to be developed. [48; 49]

### Model-based clustering

As opposed to the somewhat heuristic  $k$ -means clustering and DBSCAN, model-based clustering offers a more statistically sound approach to grouping data. It assumes the data can be represented as a mixture model of consisting one probability density function per cluster. *A priori* knowledge is required in defining the number of clusters,  $k$ , and the type of distributions (e.g., Gaussian). The clustering algorithm will then seek for locally optimal parameters for the distribution of each cluster and assign each observation to the most likely cluster based on the distributions. Here, however, we encounter a chicken-or-egg-problem: to fit the distributions to the data the cluster members must be known, but, to define the cluster members, the distribution parameters must be known. This type of problem is solved using an iterative algorithm known as expectation maximization (EM). In constructing a mixture model using EM, the distribution parameters are first initialized and then the following two steps are iterated [39]:

1. Expectation, or E-step. Using the distribution parameters, calculate the cluster membership likelihoods for each observation.
2. Maximization, or M-Step. Update the distribution parameters are to fit the membership likelihoods. Unless the changes were insignificant, return to E-step.

With Gaussian mixture models (GMM), the algorithm converges to a local optimum relatively fast. The main benefits in model-based clustering are that prior knowledge of the clusters can be incorporated by selecting the model accordingly and that the predefined model defies overfitting to data, if the model complexity is restricted. However, if the selected model (i.e. distribution type) is inappropriate, model-based cluster analysis may produce useless results. If an educated model selection is not possible, a different clustering method, hierarchical clustering for instance, can be more suitable. As a side-note, it is interesting to notice that the  $k$ -means algorithm



Kuva 2.7: Cluster analysis of data with a hierarchical structure. The left panel shows two-dimensional data. The middle and right panels visualize the results of cluster analyses with three and six clusters, respectively. Both analyses yield apparently meaningful results, but at a different level of detail, or scale.

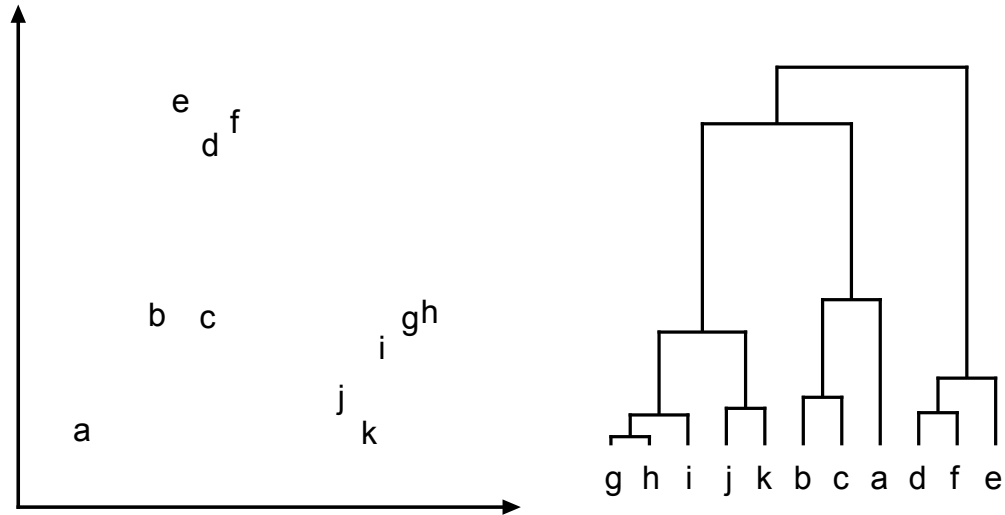
performs essentially EM for spherical Gaussian clusters. As such,  $k$ -means can also be seen as an instance of model-based clustering. [39; 38]

### Hierarchical clustering

Cluster analysis normally — especially when predefined  $k$  is used — yields information of the data only at a certain scale. Data can have both coarse and fine structure, but often it is up to the human interpreting the results which scale is most informative. Figure 2.7 shows a two-dimensional data set which has clustered structure at two different scales. The three main clusters can be further divided into smaller clusters. An approach to uncover the hierarchical structure of a data set is called hierarchical clustering.

In hierarchical clustering, the data is assumed to be composed of clusters with sub-clusters, which, in turn, may have sub-clusters and so on. Any algorithm performing hierarchical clustering will attempt to discover this hierarchy and produce a dendrogram, a tree-like graph visualizing the cluster structure. A flat (non-hierarchical) partition of the data can be obtained by using the dendrogram to define the appropriate level of detail and, hence,  $k$ . In other words, one benefit of using hierarchical clustering is that  $k$  can be defined in a rational way. [42; 39]

Figure 2.8 presents a two-dimensional data set and a dendrogram presenting its structure. The height at which the branches of the graph are joined is proportional to the distance between the clusters represented by the respective branches. In figure 2.8, we see that there are three natural clusters present: one consisting of observations a, b and c, another one consisting of d, e and f, and the last consisting of g, h, i and j. This can be seen from the dendrogram in which there is a significant gap between the second and third branch.

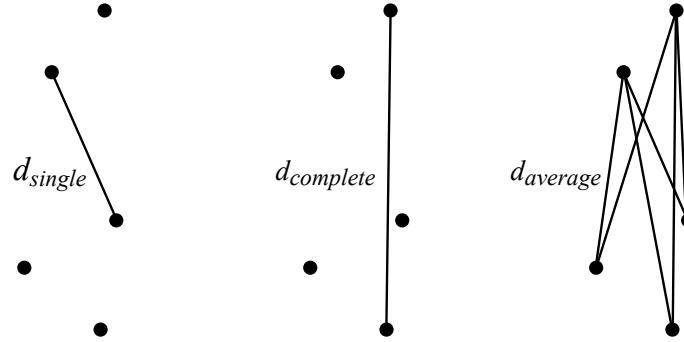


Kuva 2.8: A two-dimensional data set and a dendrogram presenting its structure. The dendrogram reveals three main clusters with varying numbers of subclusters.

Algorithms of hierarchical clustering can be divided into agglomerative and divisive methods. In agglomerative clustering, the cluster hierarchy is constructed from the bottom to the top, while divisive algorithms move from top to down. The agglomerative approach, thus, begins with unclustered data and combines two clusters (possibly consisting of a single data point) at a time. In divisive clustering, the data is assumed to compose one single cluster in the beginning, and is then divided into two sub-clusters in an optimal way, and the process is repeated until the bottom of the hierarchy is reached, i.e., all data points are in their own, terminal clusters. As there are normally significantly more ways to divide the data points than to combine them, the agglomerative algorithm is faster and therefore more popular. [39; 45] Its algorithm can be described in these four steps: [42]

1. Represent each observation as a cluster.
2. Calculate (or update) the distances between all possible pairs of clusters.
3. Combine the two nearest clusters.
4. Return to step 2 if there is more than one cluster left.

Each loop of the algorithm produces one level of the cluster hierarchy. The resulting dendrogram can be used in finding the best level to cut the dendrogram if a flat partition is wished. Often, however, the dendrogram itself is the most informative result of the clustering, at least if one is able to interpret it in order to learn about the way the data is structured. [42; 39]

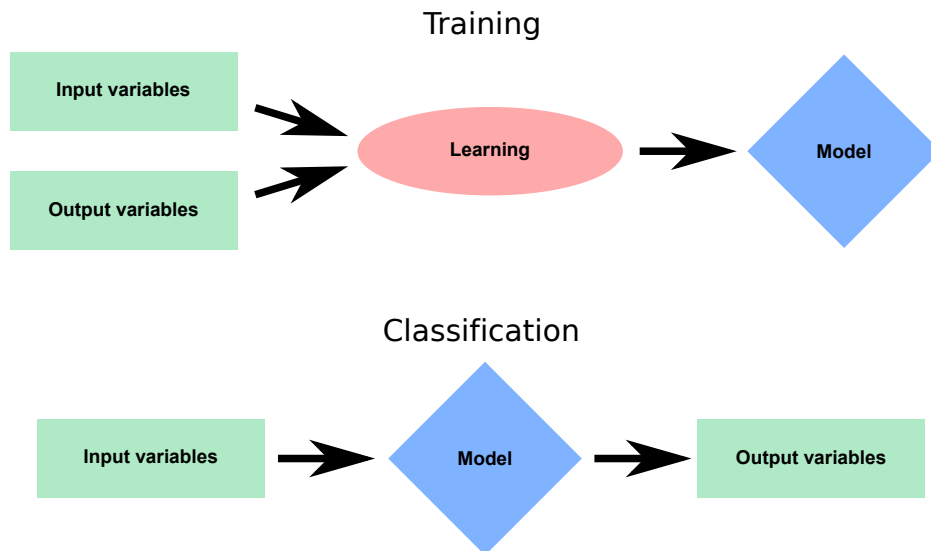


Kuva 2.9: Three common linkage methods used in hierarchical clustering. In single linkage (visualized on the left) the distance between two clusters is equal to the smallest pair-wise distance between the two clusters. In complete linkage (in the middle) it is equal to the largest pair-wise distance. Average linkage (on the right) uses the average pair-wise distance between the clusters. It is noteworthy that this is not equal to a the distance between the cluster sample means if dimensionality is over one.

Using hierarchical clustering requires choosing appropriate distance measures for the data points and for clusters, too. Distance measures between clusters are called linkage methods, and they all use the distance measure selected for data points. For this reason, calculating cluster distances requires a pair-wise distance matrix of all the data points. Four common linkage methods are: [45; 50]

1. Single linkage. In this method, the distance between two clusters is the smallest possible distance between an observation in one cluster to an observation in the other one. Vulnerable to outliers, single linkage does not work very well for large, overlapping clusters but applies to dense, irregularly shaped ones.
2. Complete linkage. In this method, the distance is the largest possible distance between observations of different clusters. Complete linkage is also vulnerable to outliers, but applies well for overlapping clusters.
3. Average linkage. In this method, the distance between two clusters is the average distance between all possible pairs-wise distances between the clusters. Average linkage is more robust than the two previous. Additional robustness is achieved by substituting average linkage to median linkage.
4. Ward linkage. This method defines the between-cluster distance as their within-cluster variance after merging them. This method tends to produce balanced clusters.

The first three methods are visualized in figure 2.9. Selecting a linkage method might require some knowledge about the cluster structure, which, of course, is what



Kuva 2.10: The two phases in classification. The classifier is trained using labeled training data. Then it can be used to label new, unlabeled samples.

cluster analysis is assumed to reveal. To overcome this chicken-or-egg-problem, it might be appropriate to try different linkage methods, study the results, and make an educated linkage selection for a final analysis based on them. [40]

## 2.8.4 Classification

Central to supervised machine learning are classification algorithms. Using a set of labeled observations, they are trained to learn the labels based on the attribute values of the observations. After the classifier is trained, it can be used to classify new, unlabeled observations. Figure 2.10 visualizes the training and classification phases. The observations can be anything ranging from images to expression profiles. In this section, we introduce some of the main types of classification algorithms.

### Instance-based classification

Instance-based classification algorithms use all or part of the observations in the training set to classify unlabeled observations. One of the most simple classification method is called nearest neighbor (NN) classification. In NN, the training phase is minimal: it consists simply of storing the training observations for use in the classification. It may be preceded by feature selection or extraction and under-sampling the observation set if necessary. Each unlabeled observation is then classified to the category of the training observation nearest to it, using any appropriate distance measure. An incrementally more advanced algorithm is KNN, or  $k$ -nearest neighbors, in which the class is based on a majority vote of the  $k$  nearest training

observations. [38]

Support vector machines (SVM) are a more complicated instance-based approach to classification. In SVM classification, a subset of the training observations, so called support vectors, are used to construct a decision boundary between the classes. In the classification phase, only the support vectors are required to categorize the unlabeled samples. The decision boundary is constructed in a way which requires linear separability of the classes, but to overcome this shortcoming, the observations can be (implicitly) mapped to a higher dimensionality in which linear separability is achieved. This adds to the computational complexity of SVMs, though it can be mitigated by using a so called kernel trick. Compared to KNN, however, the added complexity buys a advantage in memory-usage as only a minority of training observations have to be stored for classification. [38; 39]

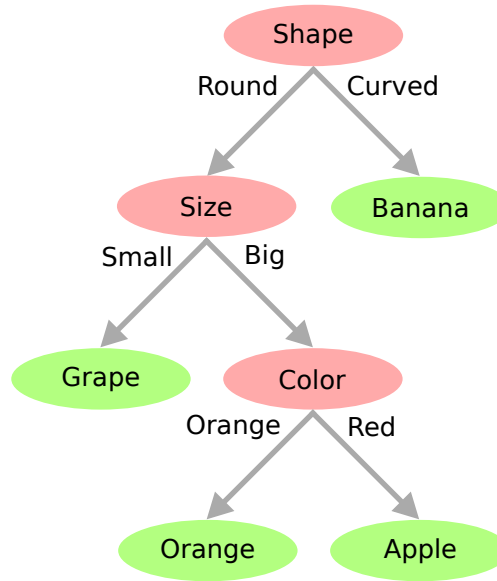
### Model-based classification

Similar to model-based clustering, a model-based classifier is trained by fitting probability density functions to the data. However, as the class labels are known, the fitting is simpler — EM is not required. Once each class has an appropriately selected and fitted PDF, unlabeled observations can be classified to the most likely class. A popular family of model-based classification methods is naïve Bayes classifiers. In naïve Bayes, each feature is considered independent, meaning that their contribution in defining the class of an observation can be determined separately. This naïve assumption is strong and potentially restricting, but it allows for a fast training using a closed-form solution, if maximum-likelihood (ML) estimation is used in fitting the PDFs. Unlike the name seems to suggest, a naïve Bayes classifier does not require a Bayesian approach to model estimation. Thus, it is not considered a proper Bayesian method. The final class probabilities, however, are calculated using Bayes' theorem, taking into account the prior class probabilities and class conditional likelihoods [39]:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}. \quad (2.17)$$

The classification result is obtained by a maximum a posteriori (MAP) estimate, i.e. selecting the class with the highest posterior probability.

Despite its inability to recognize dependencies between features, naïve Bayes classifiers have remained popular for decades. Besides fast training, their simplicity has another advantage: considering each feature separately restricts the model complexity, which, in turn, helps avoiding the issues of over-fitting and the curse of dimensionality. [39; 38]



Kuva 2.11: A simple classification tree. Composed of nodes representing binary decisions, the model classifies fruits to four categories based on three attributes. Such decision trees with a small number of nodes are easy to interpret even with very limited understanding of the theory of statistical classification.

### Decision tree classification

Predictive decision trees are perhaps the most intuitive class of classification methods. A decision tree can be visualized by a graph in which each branch point, or node, represents a decision based on a feature value. Classification of observations is performed by walking through the tree from the root node to a leaf, or terminal, node. At each node, the observation is passed on to one of the branches sprouting from the node based on the feature value. Figure 2.11 exemplifies a simple decision tree. Using features shape, size and color, it classifies fruits to four categories. The red nodes are decision nodes while green nodes are terminal nodes, each corresponding to one class.

Classification trees are trained by adding one node at a time, beginning from the root node. The feature and its threshold value for each node is determined by maximizing an optimization criterion. One commonly used criterion is Gini's impurity, which measures how well the node divides the observation sets. If the node divides the observations perfectly into separate classes, the impurities of the resulting groups are zero. If a group contains observations of more than one classes, its impurity is greater. Gini's impurity for a subset of observations is defined as [39]

$$G = 1 - \sum_{i=1}^k p^2(i) \quad (2.18)$$

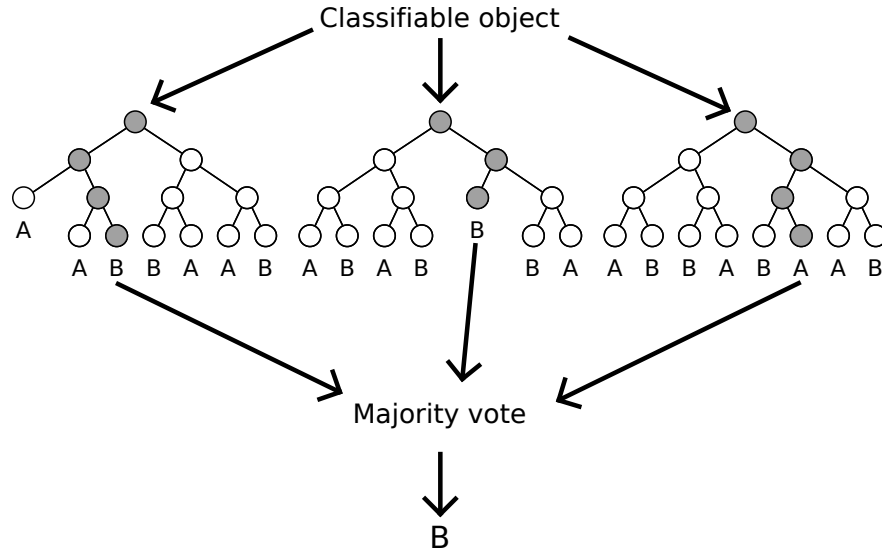
where  $p(i)$  denotes the fraction of observation in the subset belonging to the  $i$ :th class out of  $k$  classes in total. After the optimal feature and its threshold has been determined for the first node, the same is performed recursively for the following nodes. A node is considered a terminal leaf when its impurity is zero, the number of observations in the training set is smaller than a predefined threshold or the number of successive nodes reaches a predefined maximum. The tree is fully trained when All terminal leafs fulfill one of these criterions. [39; 38]

Unless the number of nodes in the trained classification tree is very high, the classifier is relatively easy to understand, visualize and communicate when compared to most other types of classifiers. If it is simple enough, it can be even used without any computation by simply walking through the tree manually and comparing the feature values to the node thresholds. This could be useful, for example, for a clinician in performing a diagnosis based on measurements and even possible categorical information (such as risk factors) of the patient. However, classification trees suffer from one disadvantage: they are notoriously prone to over-fitting. [38] To mitigate this downside, a resampling-based multi-tree approach has been developed, known as random forest (RF). [40]

### Random forest classification

RF is an example of ensemble classification. An ensemble classifier consists of multiple sub-classifiers, each trained with a subset (bag) of the entire training data set. Ensemble classification is used to avoid over-fitting by reducing the variance of the results. The result of the ensemble classifier is a combined result (often a majority vote) of the sub-classifiers. In RF, the sub-classifiers are classification trees. Each tree is trained by sampling, with replacement, a subset of predefined size from the entire training set. Furthermore, at each node in the training phase, a random feature selection takes place; only a subset of the features is used. This decreases the correlation between the trees, which, in turn, increases the generalizability and outlier tolerance of the forest. [51]

As RF can handle data of almost any type, does not require data normalization, and is considered rather generalizable, it has gained exceptional popularity within computational biology and bioinformatics, among other data-intensive disciplines, since being introduced by Leo Breiman in 2001 [52]. Besides training of RFs, Breiman introduced a method to rank the features by their predictive value. To do this, the RF is first trained and the out-of-bag (OOB) error, or classification error of the training observations not included in the bag of the tree is determined. Then, for each feature, its values are randomly permuted across the training set. The RF is re-trained with this modification and its OOB errors are calculated. The importance of the feature is then determined as the average increase in the OOB error, normalized by the



Kuva 2.12: Random forest classification. The classifier is an ensemble of decision trees, in which a random sample of features is used in defining each split within each tree. The final classification is a majority vote of the individual trees. Randomly restricting the number of features available for each split decreases the correlation between the trees and, as a result, enhances the generalizability of the ensemble. Compared to decision tree classification, the cost for avoiding over-fitting is reduced interpretability.

standard deviation of the tree-specific increases in OOB error. [51]

### 2.8.5 Classifier validation

Once a classifier is trained, its performance can be evaluated based on its ability to classify observations not used in the training. This data used to evaluate performance is called a validation set. Whenever a classifier is trained, it is advisable to use separate training and validation sets in order to ensure the classifier generalizes, i.e. performs well with unseen data. If the classifier is able to classify the training observations well but performs poorly on new data, it suffers from bad generalization and is over-fit to the training data. Over-fitting means that the classifier has learned "too much": in addition to the signal, it has modeled noise.

#### Confusion matrix-based performance measures

The most simple metric for classification performance is classification accuracy, defined as the fraction of correctly classified observations in a set. Alternatively, one might wish to use the classification error, defined as the complement of accuracy. Another commonly used and more informative presentation is a confusion matrix — a contingency table indicating the distributions of the observations over true and predicted classes. The left panel of figure 2.13 shows a confusion matrix visualizing

		Predicted class		
		Class 1	Class 2	Class 3
True class	Class 1	12	3	1
	Class 2	5	9	0
	Class 3	0	0	17

		Predicted class	
		Positive	Negative
True class	Positive	True positives	False negatives
	Negative	False positives	True negatives

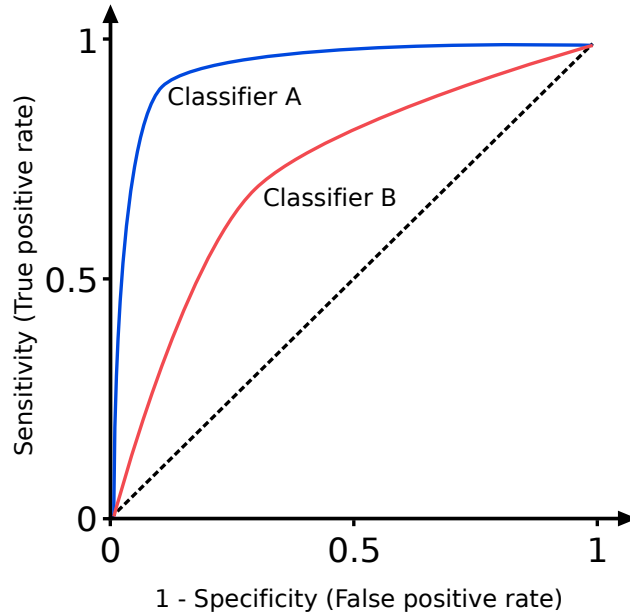
Kuva 2.13: An example confusion matrix presenting the results of three-class classification (left) and a confusion matrix template for binary classification (right).

the results of a three-category classification. The values of the diagonal elements tell that most of the observations are correctly classified. Each non-zero value on non-diagonal elements tells of confusion between classes. In this case, classes 1 and 2 are somewhat confused as observations of both classes are erroneously classified to the other class. Class 3, however, is well distinguished by the classifier: all of its observations are correctly classified and only one observation from another class is misclassified to class 3. Classification accuracy can be easily calculated from the confusion matrix by dividing the sum of its diagonal by the sum of all elements.

The right panel of figure 2.13 shows the confusion matrix template for binary classification, i.e., a two-class classification. In binary classification, the two classes are often seen as positive and negative, such as face recognized versus not recognized, or in reference to diagnosis of a disease. In the binary confusion matrix, there are only two elements corresponding to misclassifications: false positives and false negatives. A low number of false positives means that the classification has high specificity while low number of false negatives means high sensitivity. Both properties, of course, are desirable, but some times it may be useful to prefer one over the other. For example, in diagnosing a serious disease requiring immediate treatment, false positives are not as harmful as false negatives. Specificity is defined as

$$Specificity = \frac{n_{TP}}{n_{TP} + n_{FP}} \quad (2.19)$$

and sensitivity as



Kuva 2.14: Receiver operator characteristic curves of two classifiers. Classifier A has a higher sensitivity and specificity as classifier B for all values of a parameter used here to study the tradeoff between sensitivity and specificity.

$$\text{Sensitivity} = \frac{n_{TP}}{n_{TP} + n_{FN}}, \quad (2.20)$$

where  $n_{TP}$  is the number of true positives,  $n_{FP}$  false positives and  $n_{FN}$  false negatives. [40]

A useful tool in adjusting the tradeoff between sensitivity and specificity is a receiver operator characteristic (ROC) curve. Figure 2.14 shows ROC curves of two classifiers. A ROC curve is obtained by varying the value of a classifier parameter and calculating the sensitivity and specificity for every value. The sensitivity is then plotted against  $1 - \text{specificity}$ . Looking at the curve, a suitable tradeoff between sensitivity and specificity can be obtained by selecting an appropriate parameter value. The dashed line from  $(0,0)$  to  $(1,1)$  represents the most likely performance of a random, untrained binary classifier. The red curve represents a classifier somewhat better than a random binary class assigner, and the blue curve represents an even better classifier. ROC curves can also be used to determine a robust performance metric known as area under curve (AUC) by integrating the ROC curve (i.e., calculating the area under it) from 0 to 1. The resulting metric is independent of the parameter value. [40]

### Resampling-based performance estimation

When dividing the data into training and validation sets, both should be sufficiently representative of the signal, i.e., patterns present in the data. Normally, classifier performance is weighted more important than having an accurate estimate of the performance and. As a result, the split between training and validation sets is asymmetrical, favoring the training set. If, however, the entire observation set is very small (for instance, if the number of observations is smaller than that of features), resampling-based approaches can be used to overcome the problem. [38]

The perhaps most popular resampling-based method of classifier performance estimation is cross-validation (CV). In cross-validation, the classifier is trained multiple times with subsets of the data, with the left-out observations used to validate the classifier. In  $k$ -fold CV, the data is split  $k$  times such that the validation set comprises one  $k$ th of the data and is entirely separate in each fold. In 10-fold CV, for instance, a 90 % of the observations is used to train the classifier ten times, with separate 10 % validation sets. The final classifier can then be trained with the entire set and its accuracy estimated as the average accuracy of the folds (and similarly for other performance measures). It is important to remember, however, that the estimated accuracy is not precisely that of the final classifier, but rather a somewhat artificial estimate. For this reason, cross-validation should be used in performance estimation only if splitting the data into separate training and validation sets really risks compromising the signal-to-noise ratio of the sets. [38]

A method similar to CV used in classifier training, not performance estimation, is bagging. It means involves training multiple classifiers with different samples of the training data, and constructing the final classifier as an ensemble of the sub-classifiers. The goal is to reduce over-fitting, and, thus, enhance the classifiers generalizability. [39] Random forest, discussed in the previous section, is an example of a resampling-based ensemble classifier.

### 3. MATERIAL AND METHODS

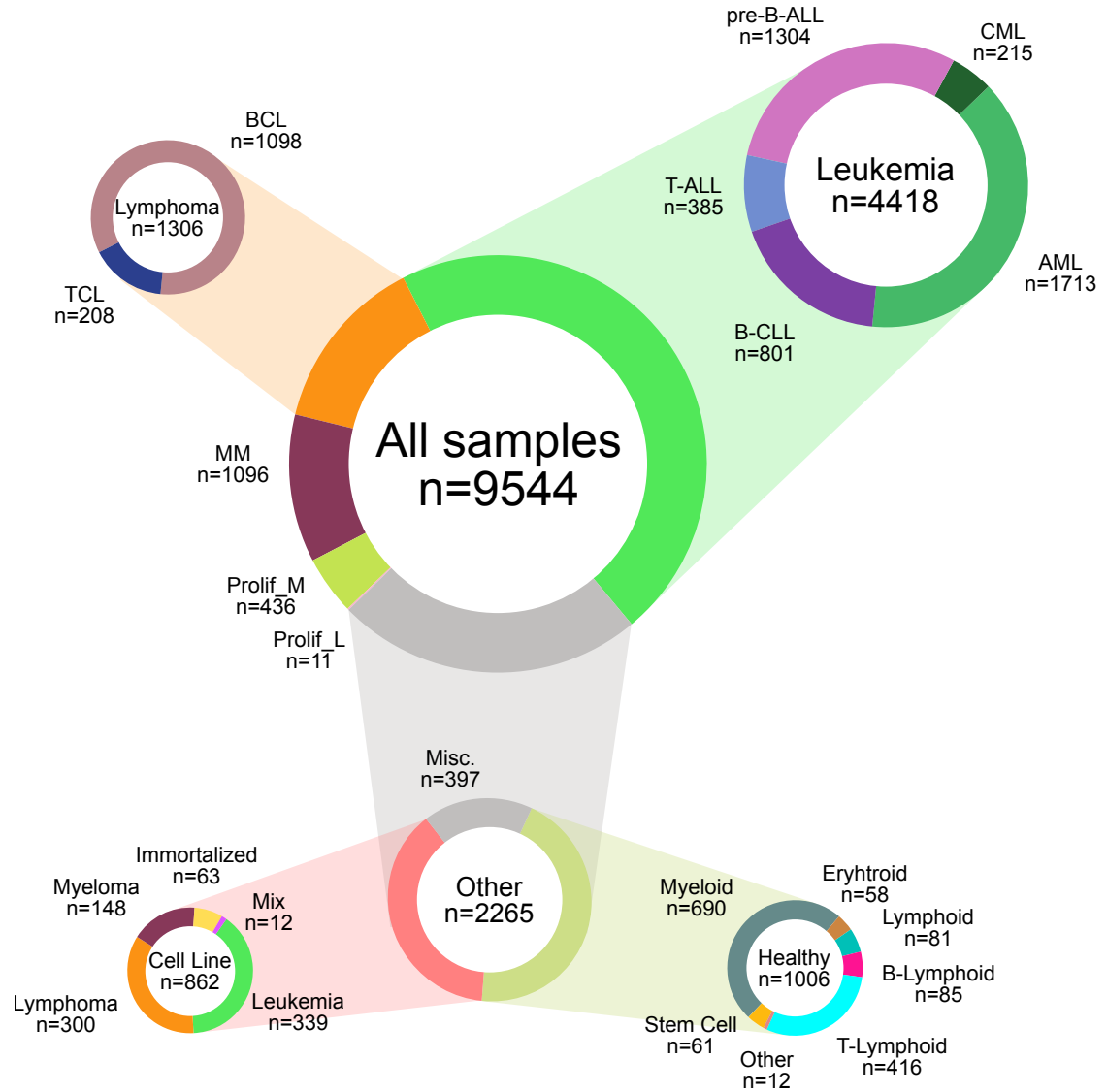
#### 3.1 Data

The data used in this thesis is publicly available gene expression microarray data from the GEO repository. Over the years, researchers around the world have conducted microarray experiments and uploaded the resulting data to public, resulting in a impressive collection of hundreds of thousands of entries from the entire spectrum of bio-medical fields. The true value in such a database is that it allows for an abundance of data-driven analyses across multiple conditions — both healthy and disease — with sample sizes great enough to produce results of extremely high statistical significance. Merging and analyzing large sets requires at least

1. careful selection of the data sets to include in analysis,
2. manual curation of the entries to ensure congruent annotations, or meta data,
3. computational methods to normalize the data to mitigate the variety of technical differences, or batch effects, between data sets and
4. computational methods and resources capable of handling the down-stream analyses of high-dimensional data.

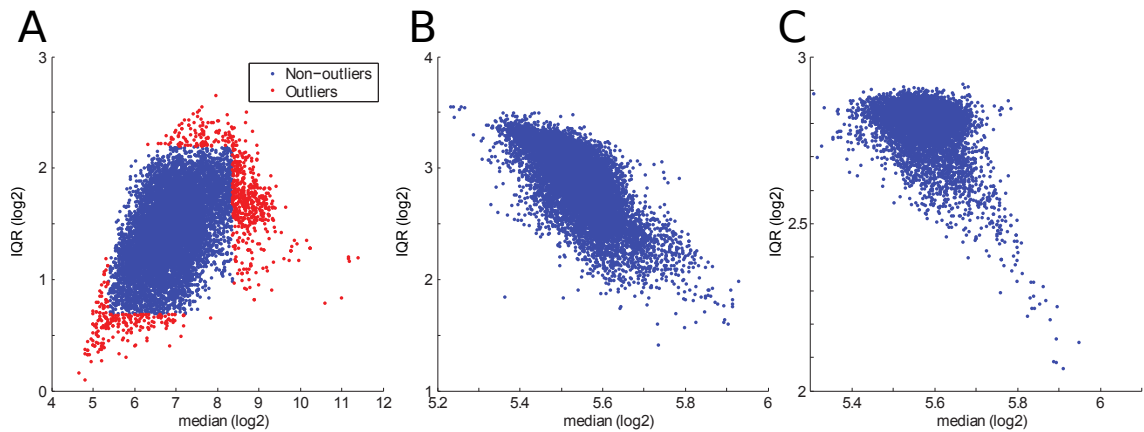
This thesis centers on approaches dealing with the latter two issues. The aim was to analyze a comprehensive hematological gene expression set. Toward this end, a collection of data sets (known in GEO as series and identified by a unique ID called GSE numbers) collectively comprising raw intensity data of over 10,000 individual microarrays (identified in GEO by IDs called GSM numbers). GEO hosts data from various microarray platforms, but only entries produced by the Affymetrix GeneChip v 2.0 plus was chosen since it is the most common platform among GEO entries and cross-platform analyses face major technical disadvantages. [53; 4]

After preliminary quality control, 9544 samples remained. The composition of the set is visualized in the charts of figure 3.1. The majority of samples (7,279) are biopsies from hematological tumors or proliferative disorders. The rest consists of healthy cells of hematological origin (1,006), blood cancer cell lines (862) and miscellaneous samples (397), such as normal cell samples from cancer patients.



Kuva 3.1: Composition of the hematological gene expression data set.

Metadata of the samples were collected and merged in a single table in a standardized fashion (outside of the scope of this thesis). This included further disease or cell type categories, genetic features, clinical parameters, basic information of the patient (such as sex or age) and information on the study for which the data was originally produced. Metadata was not available in the same detail for all samples. For example, for some patients the age is unknown, and cytogenetical information is known for only a minority of the tumors. Thus, any approach utilizing the metadata will have to handle its sparsity or the samples with insufficient annotations have to be left out of that particular analysis.



Kuva 3.2: Data dispersion as a function of its location in the three pre-processing steps. In the scatter plots, each point represents a single array of the data set. Dispersion is measured by the interquartile range (IQR) of the  $\log_2$ -intensities of an array and location as its median  $\log_2$ -intensity. Panel A shows the raw data. Arrays with deviating IQR-median positions were discarded as outliers. B shows RMA-normalized data which, due to normalization, occupies a smaller area in the IQR-median space. In C, the bias corrected data occupies an even smaller area. Notice the different scales of the plots.

### 3.2 Pre-processing

The raw intensity data for each microarray were downloaded from GEO as CEL-files. They contain intensities in numerical form of each probe on the array. To ensure the data quality, the IQR of the intensities of each array was plotted against their median, both on a  $\log_2$ -scale. Cut-off values for discarding outlier arrays were determined by visually inspecting the IQR-median scatter plot of all 11,597 arrays. As figure 3.2 illustrates, only arrays with  $\log_2$ -IQR within (5.5, 8.5) and  $\log_2$ -median within (0.75, 2.25) were selected for further pre-processing. Also, as some arrays have been submitted to GEO under different GSM IDs, the duplicates were removed at this stage.

RMA was chosen for normalization and summarization of the probe intensities to probe set (gene) expression level, since RMA has an implementation for extraordinarily large data sets. It is implemented as an R-function *justRMA Lite* in the *affyExtensions* R-package, and is designed by RMA developers to normalize extremely large data sets [54]. To further mitigate batch effects, a bias correction tool, likewise implemented as an R-function, was used in gene-wise elimination of the linear relationship between four bias metrics and gene expression [55]. The metrics include 1) the IQR of the array raw intensities, 2) the median of the raw intensities, 3) the IQR of the RMA-normalized array expression values and 4) RNA degradation values estimated for each probe set based on its probes. This bias correction step was performed after RMA-normalization.

After pre-processing, the 62 control probe sets included in the array were discarded. Their expression varied greatly across the arrays, suggesting that RNA-spike-in was performed in only a part of the arrays [32]. Therefore, they were deemed useless in the context of this data set. All of the pre-processing was performed in R using applicable R-packages. All subsequent analysis was performed in MATLAB environment.

### 3.3 Dimensionality reduction

PCA was used to visualize the data set and examine the source of its variance. Although not necessarily the best method of carrying the most essential information of the data set into lower dimensions, the simplicity of the method and the interpretability of PCs favored the choice. PCA was performed at three levels, or disease-specific sample subsets, of the data set to demonstrate the different levels of information carried by the data and to verify if the multiple-source data is comparable at different levels of detail. The three selected levels were 1) all cancer samples, 2) all leukemia samples and 3) all pre-B-ALL samples. Since pre-B-ALL samples had detailed cytogenetical annotations available, they were selected as the most detailed level in the analysis.

For the further analyses, the dimensionality was reduced by leaving out all genes with low variance across the data set, as this has been shown to enhance the performance of down-stream analyses [56]. The threshold was determined separately in each analysis by plotting the sorted variances and selecting an appropriate value to obtain only the genes with significant variance and, presumably, containing or signal instead of mere noise. The number of genes with high variance depends on the number of phenotypes represented by the data. Therefore analyses with a restricted set of disease subtypes use a lower number of genes.

### 3.4 Cluster analysis

Cluster analysis was performed to study how well disease groups can be distinguished from the data in an unsupervised manner. To test the separability in a phenotypically restricted subset, pre-B-ALL was selected for the analysis. Performing cluster analysis in a group with low biological variance facilitates identifying technical artifacts such as the array effect.

Agglomerative hierarchical clustering was selected to avoid the problem of pre-selecting the number of clusters, distribution types (in model-based clustering) and other parameters (in, for example, DBSCAN). Moreover, the additional information on possible patterns present in the data provided by the resulting dendrogram supported the decision. Ward linkage was used since it tends to produce robust, ba-

lanced clusters [57; 58]. Euclidean distance was selected as the distance measure. To reduce the amount of noise, only genes with significant variance across pre-B-ALL samples were selected.

### 3.5 Subtype prediction

Supervised classification of the cytogenetical subtypes of pre-B-ALL samples was performed to further assess the separability of the subgroups and to characterize their transcriptomic differences. RF classification was selected for this task as it has proven suitable, and popular, for gene expression studies, especially because of its ability to avoid over-fitting. Also, it has been proposed as a method to identify gene sets relevant in characterizing the differences between multiple phenotypes. [60; 59]

Before classifier training, the data set of 664 pre-B-ALL samples with cytogenetical annotations available were divided to a training set of 399 and validation set of 265 samples (a 60/40 split). The number of samples per class was assumed to be sufficient in both the training and validation set. Variance filtering of genes was used before the classification, resulting in a dimensionality of 510. The fraction of training samples to be used in training of a single tree was set to 80% and the fraction of genes to be used per tree was set to 20 (rounded square root of dimensionality).

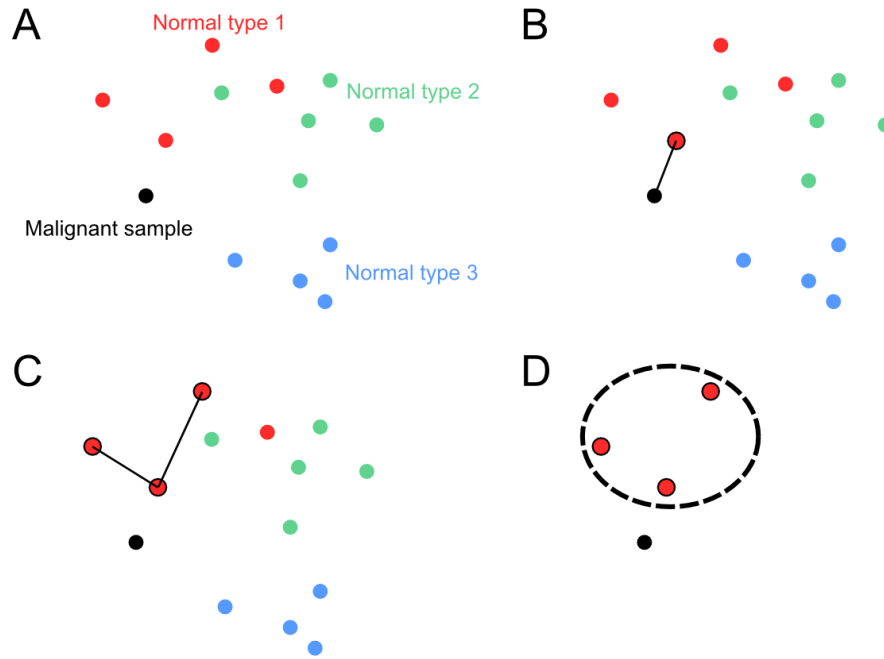
The number of trees was defined by observing the out-of-bag error rate for forest sizes ranging from 1 to 100 trees. The OOB error appeared to decrease until approximately 20 to 25 trees, and therefore, the forest was set to 25. The final validation of the model was performed by classifying the unlabeled validation set. The increase in OOB error when the expression values of a gene were randomly permuted was used to assess the its predictive value.

### 3.6 Lineage identification

To uncover the nearest healthy gene expression state of a tumor, euclidean distances (as defined by equation 2.7) between each cancer-normal sample pair were calculated. Before calculating the distances, the genes were variance-filtered, resulting in a feature dimensionality of 1,720 genes.

Figure 3.3 visualizes the process of defining the nearest normal state. Once the distances were determined, the nearest normal sample for each cancer sample was then used to define its nearest normal expression state. This was done by selecting the two nearest samples of the previously found nearest normal sample, within its cell type. The resulting nearest normal state is defined by the three normal samples, all of which represent the same cell type.

The method falls into the mid-ground between two more naïve approaches. The first would be to select the nearest normal neighbor and treat it as the nearest



Kuva 3.3: Finding the nearest healthy gene expression state for a tumor. Panel A visualizes a cancer sample in the same gene expression space as samples from three different normal cell types. First, the nearest normal sample is identified (B). Then, the two nearest samples to the one found in the first step are identified, with the criterion that they must all represent the same cell type (C). Finally, the resulting three normal samples are used to model the nearest stable healthy state of the cancer sample (D).

normal state. This approach, although simpler, is not very robust: defining a stable expression state based on a single microarray measurement is vulnerable to biases and does not appreciate the intrinsic biological variability within the state. The other approach would be essentially NN classification: defining the nearest normal state as the entire sample population of the cell type represented by the nearest normal individual sample. This method would overcome both of the issues of the first approach. However, it does not appreciate the fact that samples of the same cell type might comprise multiple sub-populations, each representing a separate stable state within the cell type. Also, the expression states of different cell types may inhabit spaces of different sizes and shapes. As we are interested in finding the closest stable state, a part of a larger normal state, or attractor, is sufficient, and even desirable. Moreover, defining the normal states as a fixed amount of samples facilitates comparing the divergences of different cancer samples.

The number of normal samples defining a stable normal state was set at three because the smallest number of samples within a normal cell the in the data set is three. If it was higher, a larger amount could have been chosen, but it is not clear which number would be an optimal tradeoff between robustness and biological

meaningfulness. The selected method does suffer from some vulnerability to outliers, though, as a single biased sample may have effect on the defined normal state.

After the nearest normal state of each cancer sample had been identified, their regulatory divergence from the normal state was quantified by estimating the number of differentially expressed TFs between the malignant and normal state. A set of 2,118 TF and other regulatory genes were selected from the entire set of 17,612 genes. Low-variance TF genes were further excluded by thresholding in a similar way as in the previous step. As a result, 440 TF genes were deemed informative of the gene-regulatory state.

Three criteria were placed for differential expression. First, an absolute expression difference of 100 was required in order to exclude noise present in very low expression values. Secondly, an expression fold-change of 1.5 was required to ensure biological relevance of the difference. Lastly, a two-tailed t-test was performed to ensure that the difference between the malignant and normal state expression values is likely not to be zero-centered. The resulting p-values were corrected for multiple testing using a method proposed by Storey [61]. The significance threshold for corrected p-values was set at 0.05.

Any TF satisfying the three above-mentioned criteria was considered differentially expressed between the malignant and normal state. The regulatory divergence of the cancer sample in question was defined as the number of differentially expressed TFs. As each malignant state is modeled by a single sample in this method, they are likely to suffer from technical biases. However, the number of samples within any cancer subtype is presumably high enough to reveal the true pattern and mitigate the noise. Naturally, the signal-to-noise ratio is best with the largest cancer subtypes.

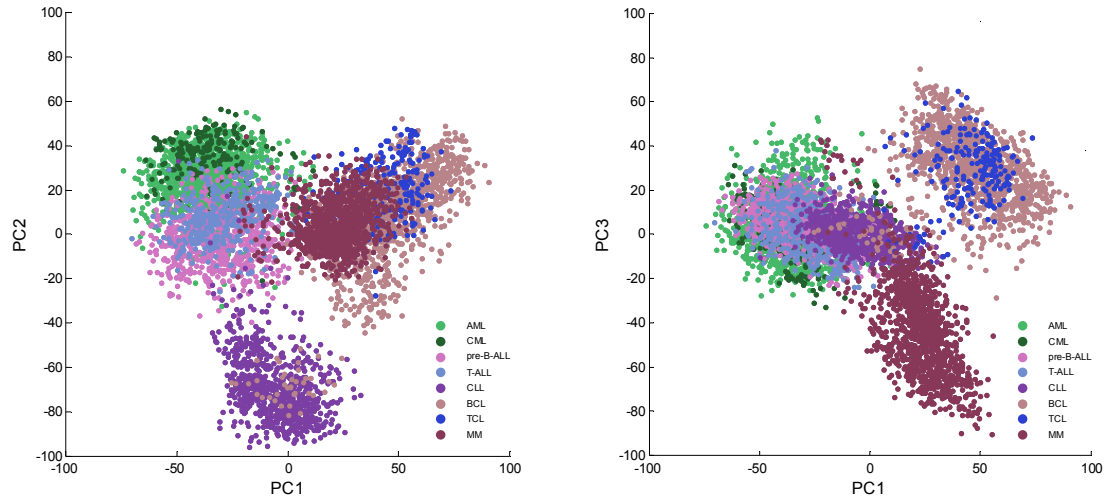
## 4. RESULTS AND DISCUSSION

### 4.1 Gene expression landscapes of cancer

Principal component visualizations attempt to preserve as much information of the difference between samples as possible. The PC coordinate system is orthogonal, meaning that the information carried by a PC is independent to all other PCs, and likely conveys its own biological interpretation. The PCs can be interpreted one at a time based on how the different disease groups are projected to it. If a PC separates to different phenotypes, it carries information on their difference. In this section, we interpret PCA results by characterizing the PCs on their ability to separate groups. The biological interpretation can be further characterized and validated based on the genes which contribute most to the PC, i.e. have the highest (positive or negative) loadings. Some PCs give no apparent information on when viewed in the context of few disease groups. In such cases, the information is more detailed and requires taking a look at the genes behind the PC. They can reveal a true biological phenomenon independent of disease groups, or, possibly, be artifacts due to batch effects.

Visualizing all the cancer samples in the dataset using PCA yields a unique birds-eye view on the transcriptomic landscape of hematological malignancies. Figure 4.1 shows all of the cancer samples along the first two PCs (left panel) and first and third PCs (right panel). Coloring the samples according to their type of hematological malignancy and lineage reveals how the samples of similar cancer tend to group together. Most notably CLL samples are grouped together, separate from all other groups. ALL and AML lay close, partially overlapping, while AML and CML, both malignancies of the myeloid lineage, occupy nearly the same space in this two-dimensional PCA plane. Lymphomas and multiple myeloma, both malignancies of more differentiated lymphocytes, also occupy nearly the same space. The first two PCs, however, clearly separate the malignancies of blast cells (leukemias) and those of differentiated hematopoietic cells (lymphocytes and multiple myeloma) from each other. Furthermore, they organize leukemias into a lineage-based order, with acute leukemias in the center and their corresponding chronic counterparts to their sides.

The right panel of figure 4.1 reveals the information carried by the third PC: it separates lymphomas from myelomas, which appeared transcriptomically similar in the context of the first two PCs. Similarly, each following PC carries its own,



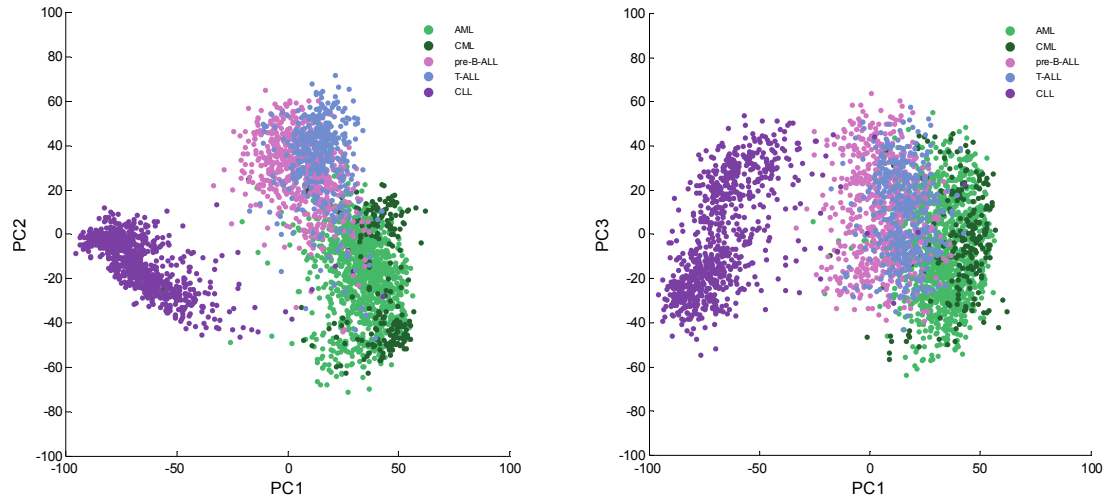
Kuva 4.1: PCA of all cancer samples. Coloring the samples based on the eight main categories of malignancies within the data set reveals that the three first PCs clearly separate four groups: 1) multiple myeloma, 2) lymphoma, 3) CLL and 4) AML, ALL and CML.

independent level of information of the transcriptomic variance within all cancers of hematopoietic origin. The fourth and fifth principal components of this pan-cancer analysis are found in the appendix (A.1) The amount of information, however, becomes more and more subtle with each following PC. The first two PCs explain 23.6 % of all variance, which is a disproportionally high fraction considering the total amount of PCs (17,612). The number of PCs needed to explain 99 % of the variance is 521, meaning that the intrinsic dimensionality is significantly lower than the actual number of features (genes). However, the finding is not surprising in the context of high-throughput measurements, where highly correlating features are commonly found within the data. Table 4.1 summarizes the interpretation, main contributing genes and variance of the first five PCs.

Taulukko 4.1: Characterization of the top principal components of all hematological cancers.

PC	Interpretation	Top gene (+)	Top gene (–)	Variance
1	LYM & MM vs. LEU	GPNMB	AZU1	12.6 %
2	Myel. vs. lymph. LEU	RRM2	KIAA0226L	11.0 %
3	LYM vs. MM	CXCL13	SDC1	7.9 %
4	Lymph. vs. myel. LEU	DNTT	FCN1	5.2 %
5	Myeloid vs. lymphoid	ATP8B4	TUBB2A	3.6 %

Performing PCA for a subset of cancer types reveals more detailed information on the specific set while compromising information of the wider context. Figure 4.2 visualizes the PCA of leukemia samples. It clearly gives more leukemia-specific infor-



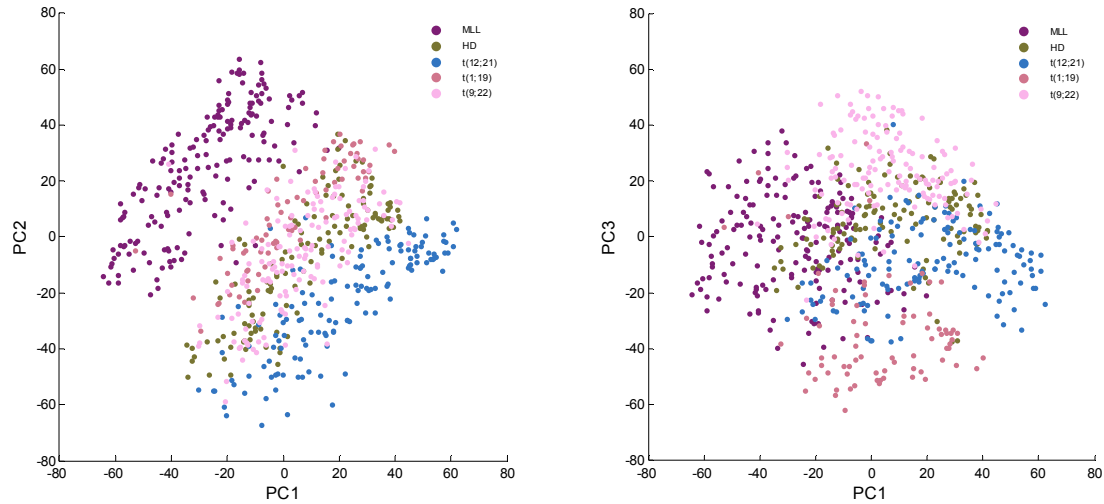
Kuva 4.2: PCA of all leukemia samples. The greatest independent source of variance within the data is explained by the transcriptomic differences between CLL and all other leukemias. The second PC explains differences between ALL and myeloid leukemias, but the third PC carries no apparent information regarding the subtypes of leukemia.

mation. The first PC separates CLL from other leukemias while placing ALL, specifically pre-B-ALL closer to CLL than other leukemias. This reflects the fact that ALL and CLL are both cancers of the lymphoid lineage. ALLs of B- and T-lineage are somewhat overlapping. The second PC separates ALL from myeloid leukemias. Interestingly, it also separates CML into two distinct groups, both overlapping with AML. The third PC, however, is more difficult to explain by this five-class leukemia grouping. It disperses all of the groups approximately equally. Further study is required to determine if this variance is explained by a true biological phenomenon instead of a technical artifact. The fourth PC separates T- and B-ALL while the fifth divides CML into two subgroups similarly to the second PC (A.2). Table 4.2 summarizes the leukemia-specific PCs.

Taulukko 4.2: Characterization of the top principal components of leukemias.

PC	Interpretation	Top gene (+)	Top gene (–)	Variance
1	Other LEU vs. CLL	AZU1	POU2AF1	19.7 %
2	ALL vs. myeloid LEU	DNTT	CSTA	8.8 %
3	?	TUBB2A	GAPT	6.2 %
4	pre-B-ALL vs. T-ALL	CTGF	ITM2A	4.6 %
5	Divides CML	S100A12	CPA3	3.9 %

Moving to an even more deeper level of information, figure 4.3 and table 4.3 show the results of PCA for pre-B-ALL. Only the samples belonging to one of the common cytogenetic subgroup were selected in order to study the transcriptomic differences



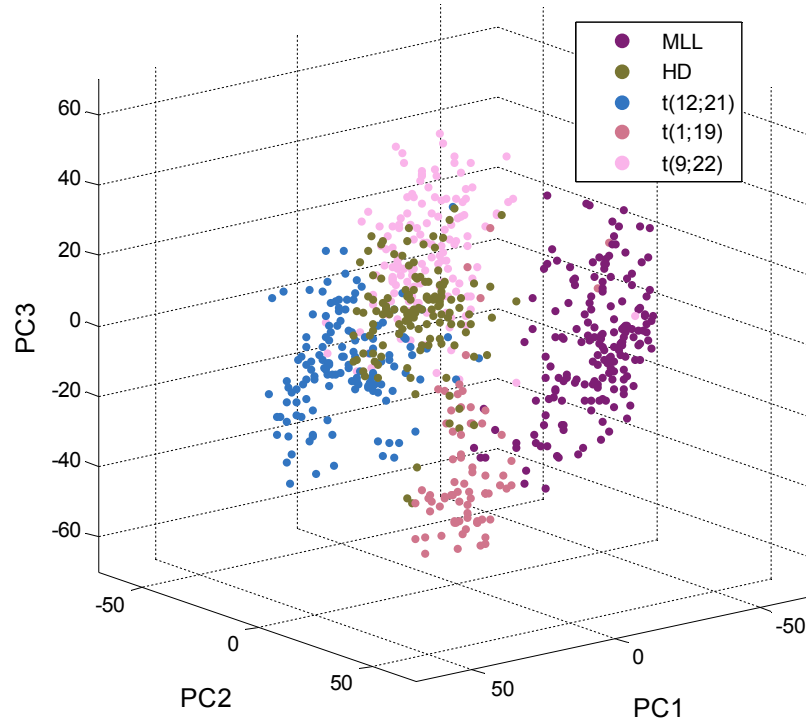
Kuva 4.3: PCA of pre-B-ALL samples. Within pre-B-ALL, PCA reveals differences between the cytogenetical subtypes, though the separation is not as clear as in main types of leukemia. Noise and artifacts are likely to cause more variance in sample sets of specific disease subtypes as the biological variance decreases.

between these groups. Three of the groups represent recurrent chromosomal translocations ( $t(12;21)$ ,  $t(1;19)$  and  $t(9;22)$ ), one represents MLL-rearrangements and one hyperdiploid karyotype. Interestingly, the first two PCs carry essentially the same information: they separate MLL and  $t(12;21)$  groups, leaving the remaining three in between. The first two PCs have also high variance in a direction perpendicular to this separation. In this direction, however, no separation between the classes is seen. The third PC separates  $t(9;22)$  from  $t(1;19)$  with HD group overlapping  $t(9;22)$ , suggesting gene-expression similarity with the two groups. The fourth and fifth PCs further separate the groups (A.3).

Taulukko 4.3: Characterization of the top principal components of pre-B-ALL.

PC	Interpretation	Top gene (+)	Top gene (–)	Variance
1	$t(12;21)$ , vs. MLL	MME	MEIS1	10.0 %
2	MLL vs. $t(12;21)$	LAMP5	SHANK3	9.7 %
3	$t(9;22)$ vs. $t(1;19)$	S100A12	PRKCZ	7.0 %
4	HD vs. $t(9;22)$	IRX1	IGJ	4.8 %
5	$t(12;21)$ vs. HD	S100A12	S100A16	4.1 %

The first three PCs suggest that the cytogenetical subtypes have distinct gene expression patterns. Even though they do not separate linearly in the two-dimensional plots, the full-dimensional representation of the data is likely to distinguish the classes more clearly. Figure 4.4 shows a two-dimensional projection of the three first PCs of pre-B-ALL samples. This projection was manually selected to



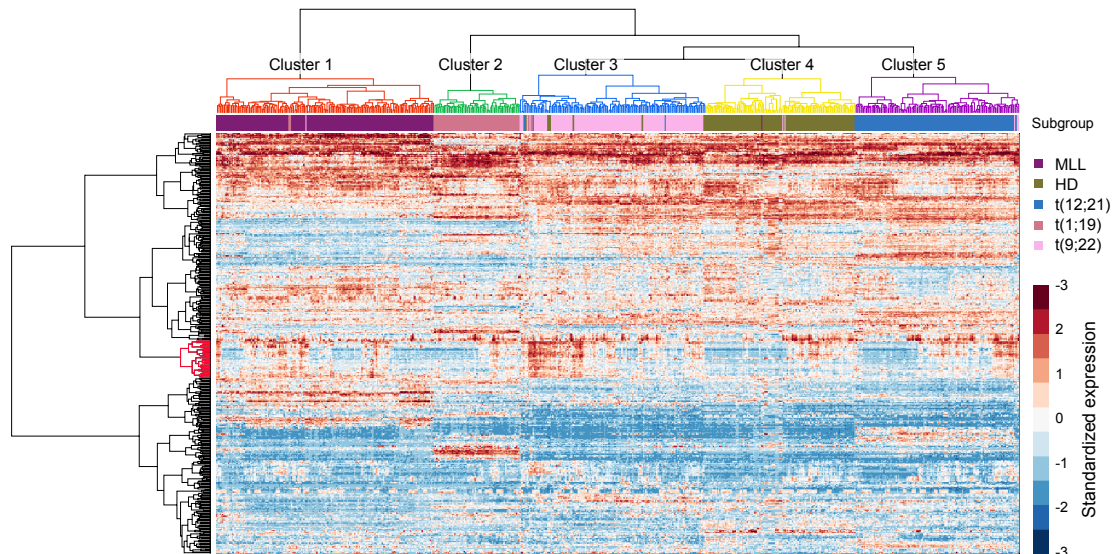
Kuva 4.4: Three-dimensional PCA of pre-B-ALL samples. Appropriately selecting a two-dimensional projection of the first three PCs reveals that the pre-B-ALL subtypes represent distinct gene expression profiles, though some overlap between subtypes is present in this PC space.

separate the classes, and it indeed separates them better than either two-dimensional plot of figure 4.3.

## 4.2 Cluster analysis

PCA hinted that the the cytogenetical subtypes of pre-B-ALL are separable in the gene expression space. To further assess this, hierarchical clustering was performed for all pre-B-ALL samples belonging to one of the five most common cytogenetical subtypes. Although pre-B-ALL is generally considered a single disease, the cytogenetics have been shown to be indicative of survival and, thus are likely to represent slightly different phenotypes.

Figure 4.5 shows a heat map of the expression profiles of pre-B-ALL subtypes with both genes (rows) and samples (columns) clustered. Cutting the dendrogram at the level of five clusters yields clusters with high concordance with the five cytogenetical subtypes. Cross-tabulation of the cytogenetical annotations and cluster assignments of the pre-B-ALL samples are shown in table 4.4. Over 90 % of samples in each cluster belong to a single subtype. Cluster 2, consisting of only t(1;19)-samples,



Kuva 4.5: Hierarchical clustering of pre-B-ALL. Selecting the cluster partition at the level of six clusters reveals that the six cytogenetical subtypes are the most defining source of variance at this scale of the data. Apart from a few outliers, all clusters correspond to a distinct subtype. The expression profiles reveal subtype-specific patterns, but one gene cluster (marked in red) appears to have high variance within each cluster. Thus, it is likely to explain a biological phenomenon unrelated to the cytogenetics. This cluster is visualized in the appendix with its gene namesA.4.

is the purest, meaning that the fraction of its most representative class is 100 %. The most impure is cluster 3, with 90.8 % of t(9;22)-samples. Considering the gene expression-wise separability of the classes shown in 4.4, this appears to reflect the fact that t(9;22) group overlaps with other groups, at least in PCA. Subtype purities, i.e. fraction of samples in a subtype in the same cluster, range between 93.4 % and 99.4 %. The overall purity, i.e. fraction of samples in the most representative cluster, is 96.4 %.

Taulukko 4.4: Cross-tabulation of the subtypes and cluster assignments of pre-B-ALL samples.

	Clust. 1	Clust. 2	Clust. 3	Clust. 4	Clust. 5	Purity
<b>MLL</b>	177	0	0	1	0	99.4 %
<b>t(1;19)</b>	2	71	3	0	0	93.4 %
<b>t(9;22)</b>	1	0	139	2	3	95.9 %
<b>HD</b>	0	0	7	121	0	94.5 %
<b>t(12;21)</b>	0	0	4	1	132	96.4 %
<b>Purity</b>	98.3 %	100.0 %	90.8 %	96.8 %	97.8 %	<b>96.4 %</b>

The fact that the five clusters produced by hierarchical clustering so clearly correspond to the five cytogenetical subtypes reveals that the subtypes represent separate

states in the gene expression space. Furthermore, at this level of detail, cytogenetics explains the gene expression differences better than any other parameter such as age, gender, cancer stage or the possible batch effect. The heat map reveals, however, a cluster of genes with significant variance in expression within pre-B-ALL, yet no correlation to the subtype. In figure 4.5, this gene cluster is marked in red in the dendrogram to the left of the heat map. In each subtype, there are samples in which the expression of the genes in this cluster are high and others in which it is low. A heat map with these genes (and their names) across pre-B-ALL samples is shown in A.4. Interestingly, the genes appear to be neutrophil specific. This suggests that some of the pre-B-ALL samples have been impure, containing neutrophils, or that some of the pre-B-ALL tumors have activated neutrophil pathways. Finding the biological reason for the phenomenon would require some further research.

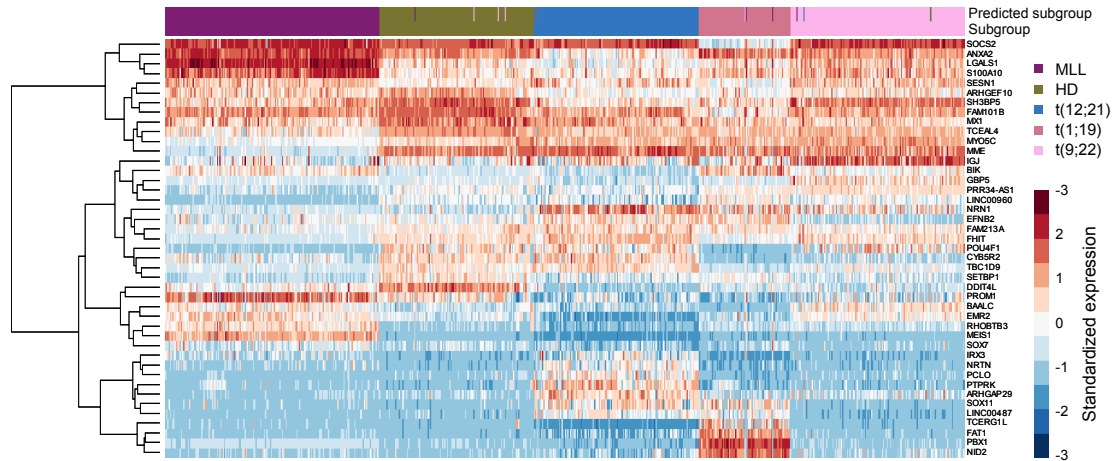
### 4.3 Subtype prediction

Unsupervised methods of dimensionality reduction and clustering revealed that the cytogenetical subtypes of pre-B-ALL have distinct gene expression profiles. To further study their separability and to determine the genes most responsible for the differences between the subtypes, supervised random forest classification was applied to pre-B-ALL samples. The classification confusion matrix for the test data set of 267 samples is shown in table 4.5. The total classification accuracy is 95.9 %, meaning that the RF classifier performs on new data essentially as well as unsupervised clustering in separating the five cytogenetical subtypes. Subtype-specific sensitivity ranges from 90.6 % in t(1;19) to 100 % in MLL. The specificity of classes is somewhat higher, ranging between 93.8 % and 100 %. The subtype-specific sensitivities and specificities correspond to the cluster- and subtype specific purities of the cluster analysis.

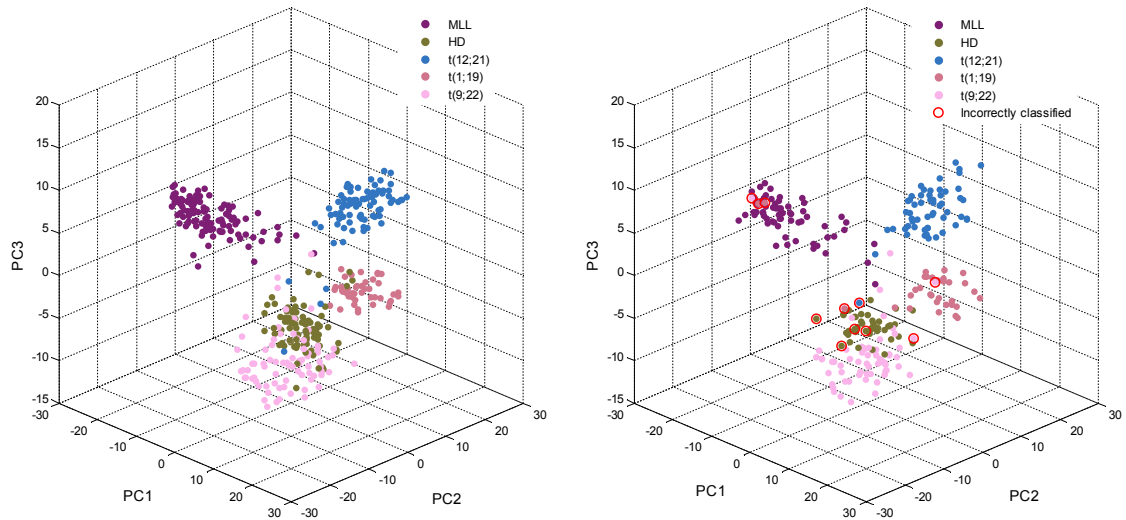
Taulukko 4.5: Confusion matrix of pre-B-ALL random forest classification.

		Predicted class					
		MLL	HD	t(12;21)	t(1;19)	t(9;22)	Sensitivity
True class	MLL	70	0	0	0	0	100.0 %
	HD	1	39	0	0	3	90.7 %
	t(12;21)	0	1	55	0	0	98.2 %
	t(1;19)	2	0	0	29	1	90.6 %
	t(9;22)	1	1	1	0	61	95.3 %
Specificity		94.6 %	95.1 %	98.2 %	100.0 %	93.8 %	

Figure 4.6 shows a heat map of the genes with the highest predictive value in classifying pre-B-ALL samples. It reveals clusters of genes with clearly subtype-specific expression patterns. For example, one can see genes which are specifically



Kuva 4.6: A gene expression heat map representing subgroups of pre-B-ALL and their RF-predicted classes. The expression profiles include only the genes deemed predictively important in RF classification. Subtype-specific patterns are strikingly clear.



Kuva 4.7: The training set (left) and validation set (right) of RF classification. All of the training samples were classified correctly by the trained RF, but eleven validation samples were misclassified. The incorrectly classified samples are marked with red circles in the right panel. Some of them appear to be clear outliers and possible misannotations while others are due to non-separability between classes. The PCA used to create this projection was performed by using only the predictive genes listed in 4.6 to obtain class separability. Thus, the PCA cannot be considered fully unsupervised. It does, however, reveal the separabilities of different classes well.

up-regulated in one subtype: *MEIS1* in MLL, *DDIT4L* in HD, *PTPRK* in t(12;21), *PBX1* in t(1;19), and *GBP5* in t(9;22). Most genes in 4.6, however, are up-regulated in several subtypes and down-regulated in others. All of them nevertheless contribute to the classification accuracy.

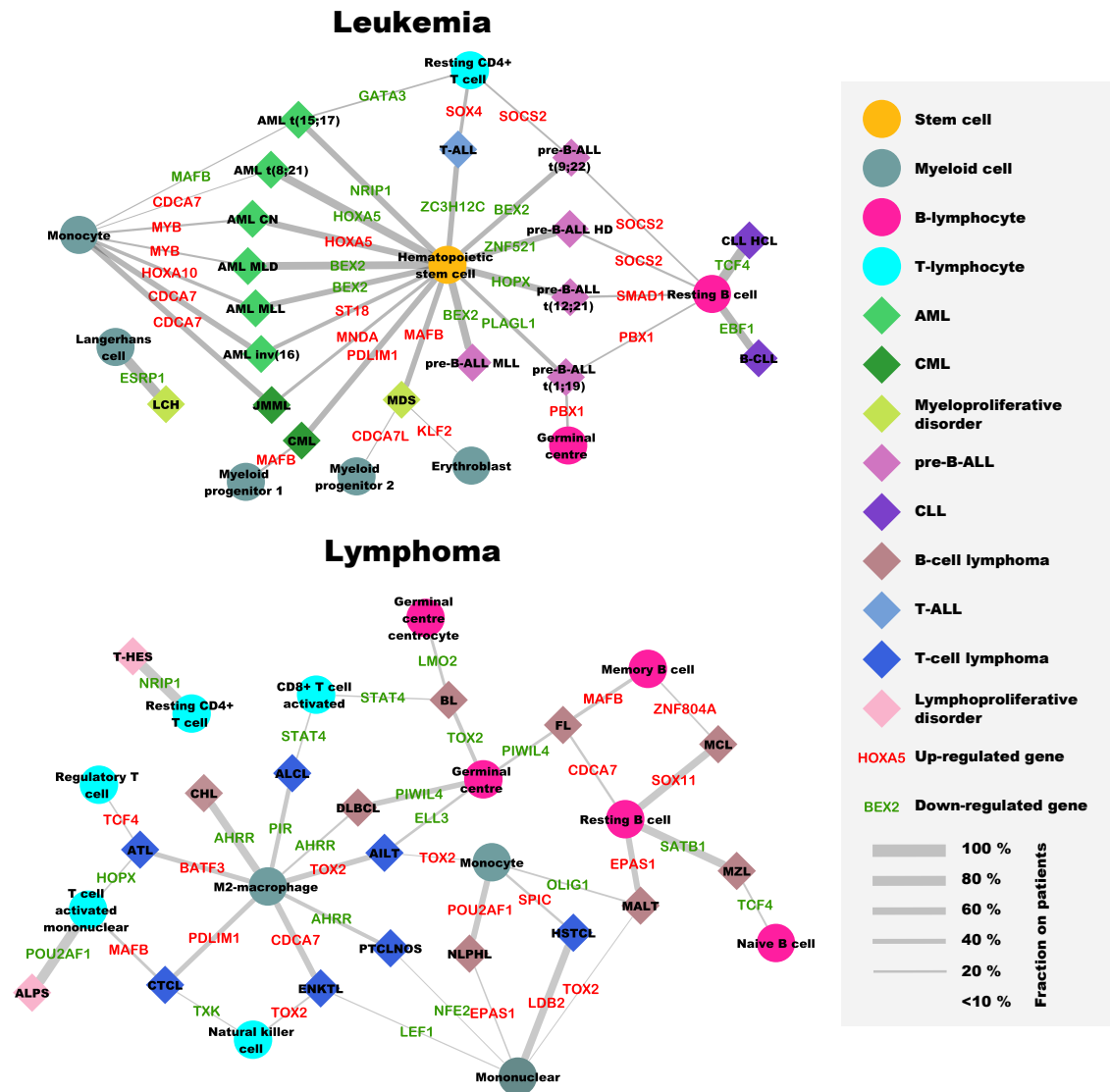
The genes with predictive value in RF classification were used to visualize the subtype separability. Figure 4.7 shows a PCA of the predictive genes. This PCA is not unsupervised in that the features (genes) were preselected to yield class separability. However, it reveals how well the classes separate. Apart from a few outliers, the MLL, t(12;21) and t(1;19) subtypes form clearly distinct clusters in the PC-space. HD and t(9;22) samples, however, partially overlap. The left panel of figure 4.7 visualizes the samples used to train the RF classifier and the right panel shows the test set. All of the training samples are correctly classified by the classifier, but eleven test samples, marked by red circles in the PCA plot, out of 265 were misclassified. Besides obvious outlier samples, possibly mislabeled, the wrongly classified samples include HD and t(9;22) samples which, apparently, are misclassified because of class the overlap. This suggests that the areas of the gene expression space for these two classes, i.e., their gene-regulatory attractors, are closer to each other than to any other pre-B-ALL subtype, and might even overlap.

#### 4.4 Cancer lineage identification

Figure 4.8 visualizes the results of lineage identification of hematological malignancies and proliferative diseases. The graphs show the closest normal cells of the patient samples in each disease group. The width of the edges between a cancer and normal cell node is proportional to the fraction of samples in the cancer subtype which were gene expression-wise closest to the respective normal cell. All fractions less than 10 % of all samples in the respective cancer subtype are omitted from the graph for clearer visualization and to avoid modeling noise. The graphs reveal that most cancers have two nearby normal states. In very few subtypes of hematological malignancies do over 90 % of the patient samples have the same closest normal state. These include pre-B-ALL with MLL rearrangement (closest normal cell HSC) and CLL (resting B-cell).

Especially in the case of leukemias, the graph suggests that the malignant state is somewhere between HSC and a mature hematopoietic cell. All AML subtypes, for example, are linked to both HSC and monocyte, a mature myeloid cell. Most pre-B-ALL subtypes, in turn, are linked to HSC and a B-lymphocyte while T-ALL is linked to HSC and a T-lymphocyte. The strong association to HSC in all acute leukemias reflects the fact that leukemic cells are immature blasts which have retained, and possibly even regained, stem cell-like properties such as replicative potential.

Unlike leukemias, lymphomas do not show any association to the HSC but, rather,

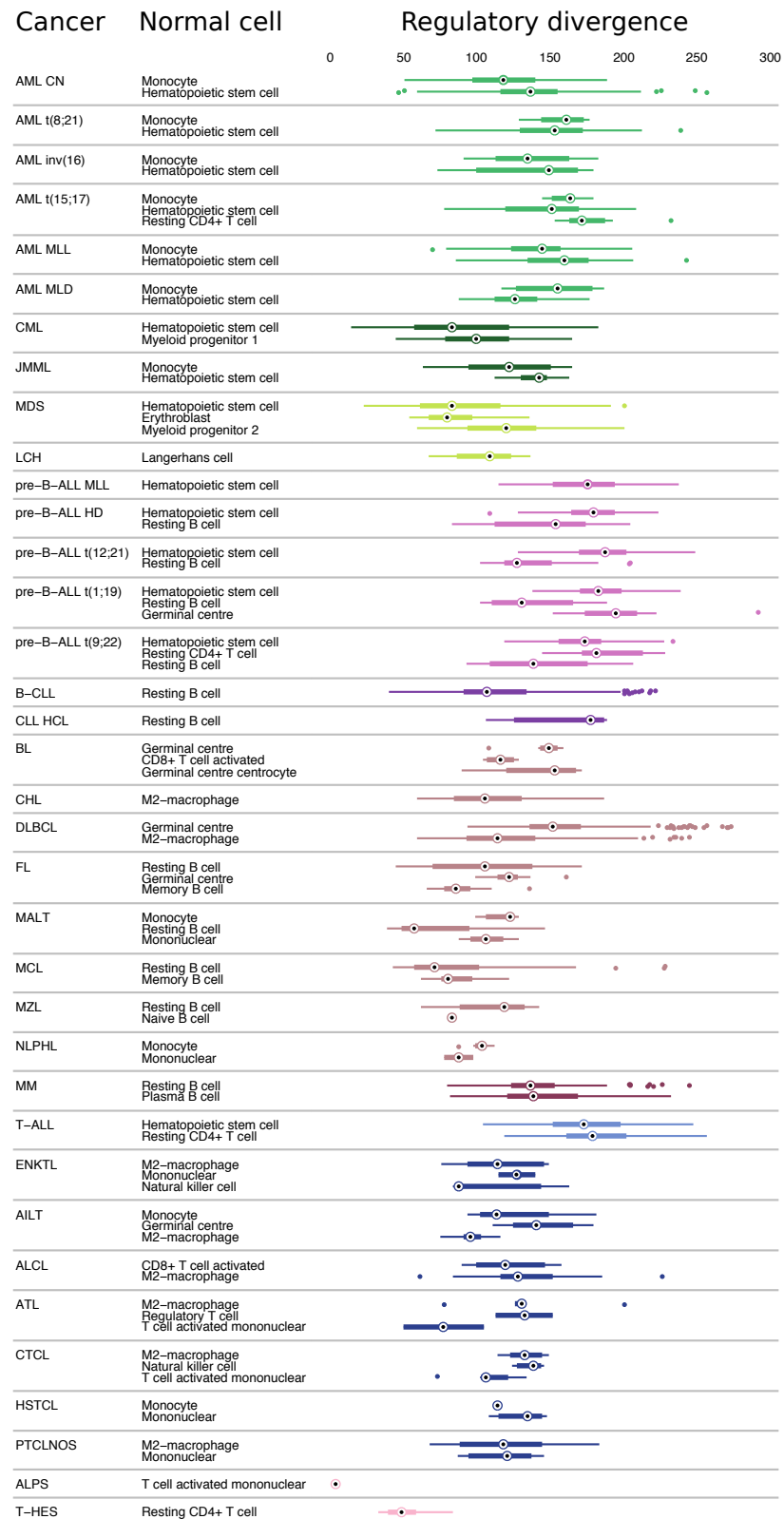


Kuva 4.8: The closest normal gene expression states of hematological malignancies. The graphs indicate which normal cell types are gene expression-wise nearest to the cancer samples of different subtypes of leukemia and lymphoma. Only in rare cases are all the patient samples of a cancer type associated to a single cell type. AMLs for example, are composed of HSC-associated as well as monocyte-associated samples. This suggests the regulatory state of AML lies between those of the stem cell and mature myelocyte. The most significant TF explaining the regulatory difference between each cancer-normal cell link is named — green indicates down-regulation in cancer while red indicates up-regulation.

mature hematopoietic cells. Interestingly, several lymphomas are strongly linked to macrophages and monocytes, even though they are cells of the myeloid, not lymphoid lineage. This seems to suggest that the solid lymphoma samples suffer from purity issues. In other words, the lymphoma samples seem to contain significant amounts of myeloid cells. The fact that these cells are specifically macrophages and monocytes might be due to 1) them being activated in fighting against the tumor cells, 2) them being recruited by the tumor or, perhaps most likely 3) them being common cells in the locations where lymphomas tend to form, i.e., lymph nodes.

Also present in figure 4.8 is the top TFs for each cancer-normal link defined as the TF which is most commonly differentially expressed within the patients and normal their closest cell type. If multiple TFs were equally frequent in this respect, the one with highest absolute expression difference was selected. The top TFs, not surprisingly, represent many of the central hematopoietic regulatory genes implicated in tumorigenesis. However, to appreciate the complexity of the regulatory aberrations, one TF likely to not be sufficient.

Figure 4.9 shows the regulatory divergences of cancer subtypes to their closest normal state as measured by differentially expressed TFs. In most cases, the divergences within a subtype are widely distributed. Some systematic patterns are, however, visible. Acute leukemias, in general, are farther from their nearest normal states than lymphomas. Specifically, cancer samples whose nearest normal state is the HSC, tend to have greater regulatory divergences. Interestingly, most pre-B-ALL subtypes are divided into samples associated to the HSC and those associated to B-lymphocytes. As the samples associated to B-cells are systematically closer to their nearest normal state, it seems that there is two pre-B-ALL sub-populations: the ones closer to HSC and those closer to B-cells. This could, however, result from impure samples containing normal B-cells in addition to leukemic blasts.



Kuva 4.9: Regulatory divergences of hematological malignancies from healthy cell types. The divergence is defined as number of TFs which are differentially expressed between the cancer samples and their nearest normal cell type.

## 5. CONCLUSIONS

The work presented in this thesis is centered around a hematological gene expression data set downloaded from a public repository. Although available for anybody with an internet connection, this type of data is next to useless without an understanding of the technical and biological biases present in multiple-source data and a means of addressing them properly. Furthermore, analyzing high-dimensional biological data spanning a plethora of hierarchically organized phenotypes requires state-of-the-art approaches of data mining, and even developing novel computational methods.

The main challenge in integrating data produced by hundreds of different laboratories around the world is to ensure that the data is comparable. This issue was addressed in multiple steps: discarding low-quality measurements before integration, collective normalization the measurements, correcting for four separate sources of bias across the entire data set, and performing down-stream analyses to validate the comparability. PCA revealed that the variance in the data set — whether analyzing thousands of cancer samples or a restricted set representing a specific disease subtype — was explained by the phenotype, not the producer of the data. In the exemplary case of pre-B-ALL, both unsupervised and supervised machine learning approaches were able to separate the cytogenetical subtypes with high sensitivity and precision, confirming that the integrated and bias-corrected expression data indeed enables cross-study analyses in highly specific settings.

Undoubtedly the good comparability of the integrated data stems, at least in part, from the sheer size of the data set: the high number of both phenotypes and instances thereof allows for 1) high-confidence detection of failed measurements as outliers, 2) a robust, low-bias estimation of the intensity distribution in quantile normalization as well as 3) high-quality estimates of the linear dependency between the four bias metrics and probe set expressions. Moreover, the number of instances in the data set enables drawing conclusions with a higher statistical significance than in single studies with a limited patient cohort. Thus, the results suggest that assembling similar data sets in the context of other diseases and healthy conditions likewise could benefit especially in analyses involving multiple phenotypes.

Characterizing leukemias, lymphomas and multiple myeloma as aberrant states of the gene regulatory system provides a novel systems biological birds-eye view to the family of cancers arising from the hematopoietic lineages. The characteriza-

tion is unique in its comprehensiveness of hematological diseases and the size of the cohort used to generate it. Studying the cancer-normal state associations and the corresponding quantified regulatory divergences yields a pan-hematological organization of myeloid and lymphoid malignancies as abnormal, immature cellular states, gene regulatory-wise somewhere between the hematopoietic stem cell and fully differentiated cells. Also, it highlights the issue of sample purity, presumably a more significant problem in solid tumors than liquid ones. Fortunately, several computational methods have been developed to purify samples *in silico* utilizing the known expression profiles of different tissues. They could prove to be crucial in saving a large proportion of gene expression data available in public repositories, possibly suffering from poor sample purity.

Revealing and quantifying the regulatory deviations of malignancies provides a new framework for cancer drug discovery. Finding any means to nudge the gene regulatory system to change its attractor away from the malignant state would cure cancer. Knowing the goal, or the closest normal attractor, and the regulatory divergence from it, is useful in determining a rational approach to push the regulatory system of a cancer cell to the right direction. Conventional ways to treat cancer — surgery, x-ray and chemotherapy — do not cure the disease in many cases. For this reason, it is fruitful to study the system-level properties of cancer cells in order to detect specific types of malignancies which might have a healthy state within a surprisingly short regulatory distance.

This thesis manages to grasp only a small, yet promising, sliver of the potential in combining and re-using data constantly produced by the worldwide biomedical research community and stored in massive repositories. Further potential lies in integrating the microarray-based expression data to that of newer, next-generation sequencing-based technologies. Even though RNA-sequencing provides valuable additional information to the expression profiles, the number of microarray measurements available is likely to outnumber that of RNA-sequencing for years. Therefore, user-friendly methods to render data from different measurement systems comparable will hold their value.

## REFERENCES

- [1] Eric Davidson, Michael Levin, "Gene regulatory networks", *Proc. Natl. Acad. Sci.*, 102, 2005
- [2] Robert Weinberg. "The Biology of Cancer". 1st ed. New York: Garland Science, 2007
- [3] Sui Huang, Gabriel Eichler, Yaneer Bar-Yam, and Donald E. Ingber, "Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network", *Phys. Rev. Lett.*, 94, 2005
- [4] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis and Alexandra Soboleva, "NCBI GEO: archive for functional genomics data sets—update", *Nucleic Acids Res*, 41, 991–995, 2013
- [5] Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, Victor Robles, "Machine learning in bioinformatics", *Brief Bioinform*, 7, pp. 86–112, 2006
- [6] Merja Heinäniemi, Matti Nykter, Roger Kramer, Anke Wienecke-Baldacchino, Lasse Sinkkonen, Joseph Xu Zhou, Richard Kreisberg, Stuart A Kauffman, Sui Huang and Ilya Shmulevich, "Gene-pair expression signatures reveal lineage control", *Nature Methods*, 10, pp. 577–583 , 2013
- [7] Thomas Liuksiala, Kaisa Teittinen, Kirsi Granberg, Merja Heinäniemi, Matti Annala, Markku Mäki, Matti Nykter and Olli Lohi, "Overexpression of SNORD114-3 marks acute promyelocytic leukemia", *Leukemia*, 28, pp. 233–236, 2014
- [8] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and Peter Walter, "Essential Cell Biology", 3rd ed. New York: Garland Science, 2010
- [9] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and Peter Walter, "Molecular Biology of the Cell", 5th ed. New York: Garland Science, 2008
- [10] Robert Weinberg and Douglas Hanahan. "The Hallmarks of Cancer", *Cell*, 100, pp. 57–70, 2000

- [11] Robert Weinberg and Douglas Hanahan. "Hallmarks of Cancer: The Next Generation", *Cell*, 144, pp. 646–674, 2011
- [12] Nancy Lee Harris, Elaine S. Jaffe, Jacques Diebold, Georges Flandrin, H. Konrad Muller-Hermelink, James Vardiman, T. Andrew Lister and Clara D. Bloomfield, "The World Health Organization Classification of Hematological Malignancies Report of the Clinical Advisory Committee Meeting, Airlie House, Virginia, November 1997", *Mod Pathol*, 13, 193–207, 2000
- [13] James W. Vardiman, Nancy Lee Harris and Richard D. Brunning, "The World Health Organization (WHO) classification of the myeloid neoplasms", *Blood*, 100, 2002
- [14] Alberto Orfao, Gerd Schmitz, Bruno Brando, Alejandro Ruiz-Arguelles, Giuseppe Basso, Raul Braylan, Gregor Rothe, Francis Lacombe, Francesco Lanza, Stefano Papa, Paulo Lucio and Jesus F. San Miguel, "Clinically useful information provided by the flow cytometric immunophenotyping of hematological malignancies: current status and future directions", *Clin Chem*, 45, pp. 1708–1717, 1999
- [15] Fiona E. Craig and Kenneth A. Foon, "Flow cytometric immunophenotyping for hematologic neoplasms", *Blood*, 111, 2008
- [16] Esteban Braggio, Jan B. Egan, Rafael Fonseca and A. Keith Stewart, "Lessons from next-generation sequencing analysis in hematological malignancies", *Blood Cancer Journal*, 3, e127, 2013
- [17] R. Coleman Lindsley and Benjamin L. Ebert, "The biology and clinical impact of genetic lesions in myeloid malignancies", *Blood*, 112, 2013
- [18] Chun Yew Fong, Jessica Morison, Mark A. Dawson, "Epigenetics in the hematologic malignancies", *Haematologica*, 99, pp. 1772–1783, 2014
- [19] John M. Bennett, Daniel Catovsky, Marie T. Daniel, George Flandrin, David A. G. Galton, Harvey R. Gralnick and Claude Sultan, "Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group", *Br. J. Haematol.* 33, 451–458, 1976
- [20] James W. Vardiman, Jürgen Thiele, Daniel A. Arber, Richard D. Brunning, Michael J. Borowitz, Anna Porwit, Nancy Lee Harris, Michelle M. Le Beau, Eva Hellström-Lindberg, Ayalew Tefferi, and Clara D. Bloomfield, "The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes", *Blood*, 114, 2009

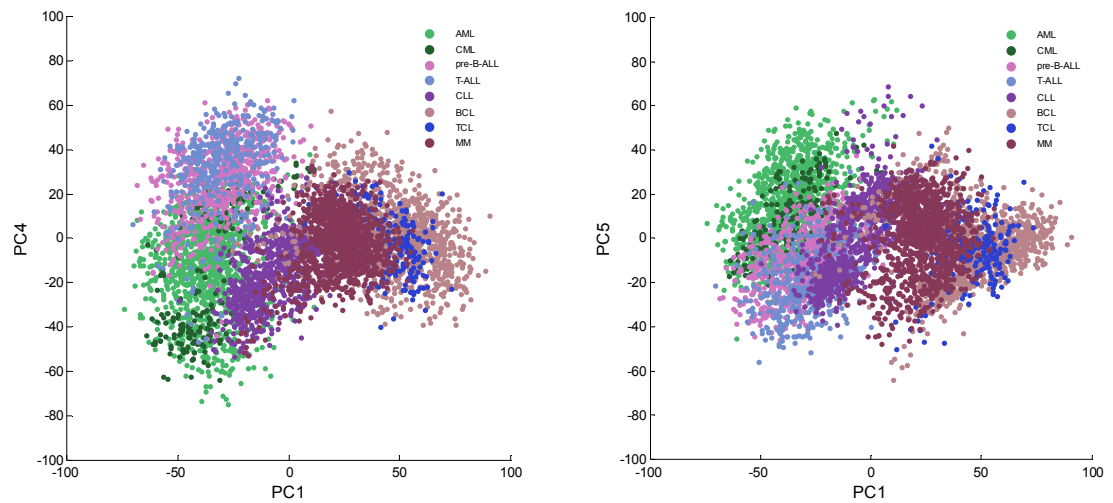
- [21] Elias Campo, Steven H. Swerdlow, Nancy L. Harris, Stefano Pileri, Harald Stein, and Elaine S. Jaffe, "The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications", *Blood*, 117, 2011
- [22] Charles G. Mullighan, "Genome sequencing of lymphoid malignancies", *Blood*, 122, 3899–9307, 2013
- [23] Conrad H. Waddington, "The Strategy of the Genes", 1st ed. London: George Allen & Unwin, 1957
- [24] Aaron D. Goldberg, C. David Allis and Emily Bernstein, "Epigenetics: A Landscape Takes Shape", *Cell*, 128, 935–938, 2007
- [25] Lorraine Robb, "Cytokine receptors and hematopoietic differentiation", *Oncogene*, 26, 6715–6723, 2007
- [26] Stuart Kauffman, "Homeostasis and Differentiation in Random Genetic Control Networks", *Nature*, 5215: 177–178, 1967
- [27] Alexandre Haye, Jaroslav Albert and Marianne Rooman, "Robust non-linear differential equation models of gene expression evolution across *Drosophila* development", *BMC Research Notes*, 46, 2012
- [28] Sui Huang, Ingemar Ernberg, and Stuart Kauffman, "Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective", *Semin Cell Dev Biol.*, 20, 869–876, 2009
- [29] Jakob Lovén, David A. Orlando, Alla A. Sigova, Charles Y. Lin, Peter B. Rahl, Christopher B. Burge, David L. Levens, Tong Ihn Lee and Richard A. Young, "Revisiting Global Gene Expression Analysis", *Cell*, 151, pp. 476–482, 2012
- [30] John Quackenbush, "Computational analysis of microarray data", *Nat Rev Genet.*, 6, 418–427, 2001
- [31] Atul Butte, "The use and analysis of microarray data", *Nature Reviews Drug Discovery*, 1, 951–960, 2002
- [32] Affymetrix, "Affymetrix Microarray Suite Guide", version 5.0, Affymetrix Inc, Santa Clara, CA, 2001.
- [33] Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed, "Summaries of Affymetrix GeneChip probe level data", *Nucleic Acids Research*, 31:e15, 2003

- [34] Bettina Harr and Christian Schlötterer, "Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons", *Nucleic Acids Res.*, 34, e8, 2006
- [35] Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A. Ball, Helen C. Causton, Terry Gaasterland, Patrick Glenisson, Frank C.P. Holstege, Irene F. Kim, Victor Markowitz, John C. Matese, Helen Parkinson, Alan Robinson, Ugis Sarkans, Steffen Schulze-Kremer, Jason Stewart, Ronald Taylor, Jaak Vilo and Martin Vingron, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data", *Nat Genet.*, 29, 365–71, 2001
- [36] David Brock, "Understanding Moore's Law: Four Decades of Innovation", *Chemical Heritage Foundation*, pp. 67–84, 2006
- [37] Edda Klipp, Wolfram Liebermeister, Christoph Wierling, Axel Kowald, Hans Lehrach and Ralf Herwig, "Systems Biology", 2nd ed. Weinheim: Wiley VCH, 2012
- [38] Christopher R. Bishop, "Pattern Recognition and Machine Learning", 8th ed. New York: Springer, 2009
- [39] Richard Duda, Peter Hart and David Stork, "Pattern Classification", 2nd ed. New York: John Wiley & Sons, 2001
- [40] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, "An Introduction to Statistical Learning with Applications in R", 4th ed. New York: Springer 2014
- [41] Richard Bennett, "Representation and analysis of signals—Part XXI: The intrinsic dimensionality of signal collections", 1st ed. Baltimore, MD: The Johns Hopkins University, 1965
- [42] Rui Xu and Donald C. Wunsch II, "Clustering", 1st ed. London: Chapman & Hall/CRC, Hoboken: John Wiley & Sons, 2009
- [43] Anil Jain and Richard Dubes, "Algorithms for Clustering Data", 2nd ed. Englewoods Cliffs: Prentice Hall, 1988
- [44] Helmut Späth, "Cluster Analysis Algorithms for Data Reduction and Classification of Objects", 4th ed. Chichester: Ellis Horwood, 1980
- [45] Leonard Kaufman and Peter J. Rousseeuw, "Finding Groups in Data", 1st ed. New York: John Wiley & Sons, 1990

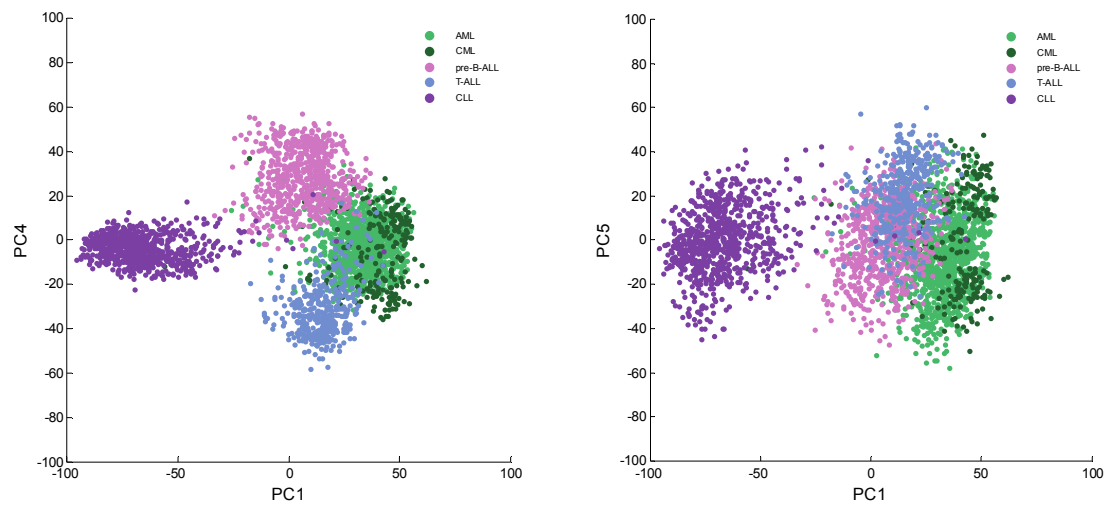
- [46] Brian Everitt, "Cluster Analysis", 3rd ed. Bristol: J W Arrowsmith, 1993
- [47] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, Evangelos Simoudis, Jiawei Han, Usama M. Fayyad, "A density-based algorithm for discovering clusters in large spatial databases with noise", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231, 1996
- [48] Jörg Sander, Martin Ester, Hans-Peter Kriegel and Xiaowei Xu, "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications", *Data Mining and Knowledge Discovery* 2, pp.169–194, 1998
- [49] Jörg Sander, "Generalized Density-Based Clustering for Spatial Data Mining", 1st ed. München: Herbert Utz Verlag, 1998
- [50] Joe H. Ward Jr., "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, 58, pp. 236–244, 1963
- [51] Leo Breiman, "Random Forests", *Machine Learning*, 45, pp. 5–32, 2001
- [52] Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa and Inke R. König, "Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics", University of Munich Department of Statistics, Technical Report Number 129, 2012
- [53] Ron Edgar, Michael Domrachev and Alex E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository" *Nucleic Acids Res*, 30, pp. 207–10, 2002
- [54] Laurent Gautier, Leslie Cope, Benjamin M. Bolstad and Rafael A. Irizarry, "affy—analysis of Affymetrix GeneChip data at the probe level", *Bioinformatics*, 20, pp. 307–315, 2004
- [55] Aron C. Eklund and Zoltan Szallasi, "Correction of technical bias in clinical microarray data improves concordance with known biological information", *Genome Biology*, 9:R26, 2008
- [56] Richard Bourgon, Robert Gentleman, and Wolfgang Huber, "Independent filtering increases detection power for high-throughput experiments", *Proceedings of the National Academy of Sciences*, 107, pp. 9546–9551, 2009
- [57] Harry Clifford, Frank Wessely, Satish Pendurthi, and Richard D. Emes, "Comparison of Clustering Methods for Investigation of Genome-Wide Methylation Array Data", *Front Genet.*, 2: 88, 2011

- [58] Jeremy J. Jay, John D. Eblen, Yun Zhang, Mikael Benson, Andy D. Perkins, Arnold M. Saxton, Brynn H. Voy, Elissa J. Chesler and Michael A. Langston, "A systematic comparison of genome-scale clustering algorithms", *BMC Bioinformatics*, 13(Suppl 10):S7, 2012,
- [59] Huey-Miin Hsueha, Da-Wei Zhoua and Chen-An Tsaib, "Random forests-based differential analysis of gene sets for gene expression data", *Gene*, 518, pp. 179–186, 2013
- [60] Ali Anaissi, Paul J. Kennedy, Madhu Goyal and Daniel R. Catchpoole, "A balanced iterative random forest for gene selection from microarray data", *BMC Bioinformatics*, 261, 2013
- [61] John D. Storey. "A direct approach to false discovery rates", *Journal of the Royal Statistical Society*, 64, pp. 479–498, 2002

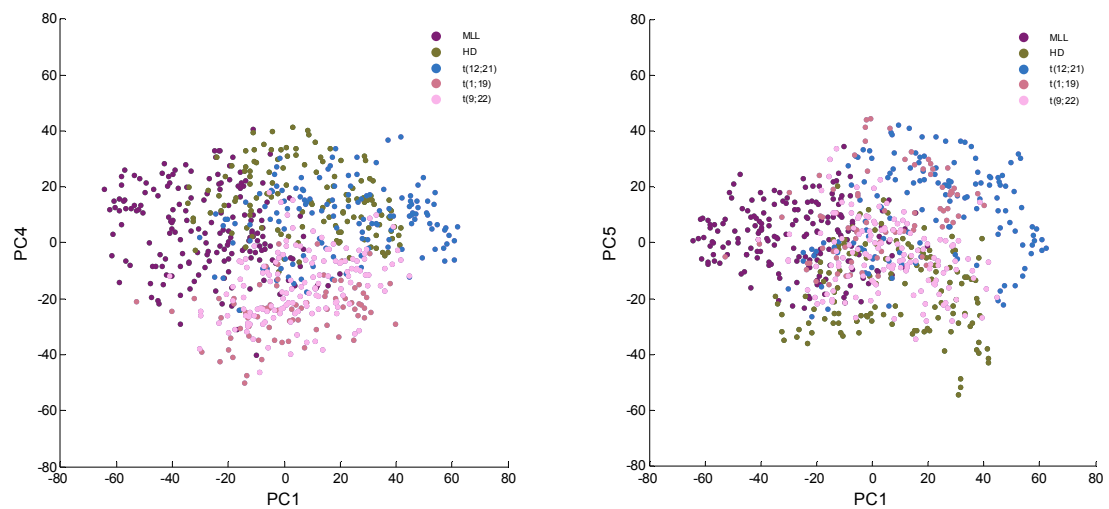
## A. APPENDICES



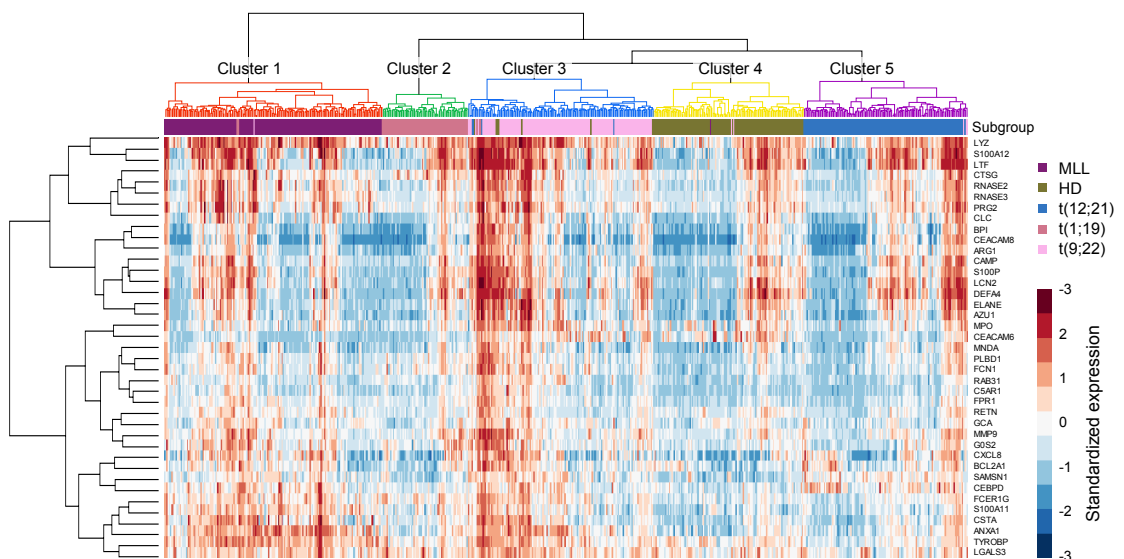
Kuva A.1: The fourth and fifth principal components of all cancer samples against the first principal component.



Kuva A.2: The fourth and fifth principal components of all leukemia samples against the first principal component.



Kuva A.3: The fourth and fifth principal components of all pre-B-ALL samples against the first principal component.



Kuva A.4: Expression profiles of genes of a cluster with high intra-cluster variance in cluster analysis of pre-B-ALL. The expression profiles of all genes involved in the clustering are visualized in figure 4.5