LUND UNIVERSITY

# Bioinformatic approaches to gene expression in leukemia. Networks and deconvolution

Järvstråt, Linnea

2017

*Document Version:*
Publisher's PDF, also known as Version of record

Bioinformatic approaches to gene expression in leukemia

# Bioinformatic approaches to gene expression in leukemia

## Networks and deconvolution

by Linnea Järvstråt

Thesis for the degree of Doctor of Philosophy
Thesis advisors: Dr. Björn Nilsson, Prof. Urban Gullberg
Faculty opponent: Prof. Bengt Persson

To be presented, with the permission of the Faculty of Medicine of Lund University, for public criticism in I1245 lecture hall at the Department of Laboratory Medicine, Division of Hematology and Transfusion medicine on Thursday, the 9th of November 2017 at 13:00.

Organization
**LUND UNIVERSITY**
Department of Laboratory Medicine, Division of Hematology and Transfusion medicine

Document name
**DOCTORAL DISSERTATION**

Date of disputation
2017-10-09

Sponsoring organization

Author(s)
Linnea Järvstråt

Title and subtitle
Bioinformatic approaches to gene expression in leukemia: Networks and deconvolution

Abstract
The aim of this thesis is develop methods to extract information from high-dimensionality data and to apply in looking at gene-gene and gene-protein interactions in Acute Myeloid Leukemia (AML). The *in silico* methods developed can be used with data from other systems. The thesis gives an overview of the field for network inference and deconvolution methods, as well as a brief background on the biology concerning AML.

Paper I develops a method, Ultranet, for inferring Gaussian Graphical Models in an efficient manner. The models can created by solving the sparse inverse covariance selection (SICS) problem be used to identify gene networks from data containing a large number of variables but a proportionately low sample number. We apply Ultranet to data from several blood disorders and show that the models capture blood related gene interactions.

In paper II, we investigate how the cellular heterogeneity of many tissue samples can be taken into consideration when examining gene expression of said samples. Cell hierarchies are thought to be present in leukemia and some solid tumors, sustained by cancer stem cells (CSCs). We developed a computational approach that extracts gene expression patterns and cell type proportions *in silico*.

Paper III and IV looks at how the DEK and WT1 proteins, respectively, interact with DNA using chromatin immunoprecipitation followed by sequencing. The proteins are known to be altered in a range of different cancer forms, including Acute Myeloid Leukemia.

We find, in paper III, that DEK binds close to transcription start sites of actively transcribed genes as indicated by the presence of the histone markers. DEK binds to genes that are ubiquitously expressed across tissues, represented by presence of RNA polymerase II binding sites in cell lines. We also show that knockdown of DEK by shRNA results in both significant down- and up-regulations of DEK-bound genes, suggesting complex interactions.

In paper IV, we study how the binding patterns of the WT1 isoforms differ. There is an alternative splice site between zinc finger three and four which leads to inclusion or exclusion of three amino acids (+/-KTS). We show that WT1 -KTS binds to transcription start sites, while WT1 +KTS bind inside gene bodies.

Key words
bioinformatics, acute myeloid leukemia, AML, networks, deconvolution, microarray, ChIP-seq, DEK, WT1,

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1652 – 8220

ISBN
978-91-7619-521-5 (print)

Recipient's notes

Number of pages
56

Price

Security classification

# Bioinformatic approaches to gene expression in leukemia

## Networks and deconvolution

by Linnea Järvstråt

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarizes the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

**Cover illustration:** A part of the gene expression network in leukemia.

**Funding information:** The thesis work was financially supported by Swedish Foundation for Strategic Research, Kurt and Alice Wallenberg Foundation, Swedish Research Council, Swedish Childhood Cancer Foundation, and Svenska Läkaresällskapet.



Faculty of Medicine, Department of Laboratory Medicine, Division of Hematology and Transfusion medicine

ISBN: 978-91-7619-521-5 (print)
Lund University, Faculty of Medicine Doctoral Dissertation Series ISSN: 1652-8220

Printed in Sweden by Media-Tryck, Lund University, Lund 2017

*Dedicated to my siblings*
*Lotta – Klas*

# Contents

List of publications . . . . . ii
Acknowledgements . . . . . iv
Populärvetenskaplig sammanfattning på svenska . . . . . v

**Bioinformatic approaches to gene expression in leukemia: Networks and deconvolution** **1**
Abbreviations . . . . . 2

**Introduction** **3**

**Biological background** **5**
Blood formation . . . . . 5
Acute Myeloid Leukemia . . . . . 5
Leukemic stem cells . . . . . 7

**Computational background** **11**
Network inference . . . . . 11
Deconvolution . . . . . 17

**Main results of the research papers** **23**
Aim . . . . . 23
Laboratory methods . . . . . 24
Methods . . . . . 25
Results . . . . . 26
Discussion . . . . . 30

**References** **33**

# List of publications

This thesis is based on the following original articles, referred to by their Roman numerals:

I **Ultranet: efficient solver for the sparse inverse covariance selection problem in gene network modeling**

**Linnea Järvstråt**, Mikael Johansson, Urban Gullberg, Björn Nilsson
Bioinformatics. 2013 Feb 15;29(4):511-2.

II **Deconvolution of gene expression in cancer cell hierarchies**

**Linnea Järvstråt**, Ram Ajore, Anna-Karin Wihlborg, Urban Gullberg, Björn Nilsson
Manuscript

III **The DEK oncoprotein binds to highly and ubiquitously expressed genes with a dual role in their transcriptional regulation**

Carl Sandén, **Linnea Järvstråt**, Andreas Lennartsson, Per Ludvik Brattås, Björn Nilsson, Urban Gullberg
Molecular Cancer. 2014 Sep 12;13:215.

IV **Distinct global binding patterns of the Wilms tumor gene 1 (WT1) -KTS and +KTS isoforms in leukemic cells**

Tove Ullmark, **Linnea Järvstråt**, Carl Sandén, Giorgia Montano, Helena Jernmark Nilsson, Henrik Lilljebjörn, Andreas Lennartsson, Thoas Fioretos, Kristina Drott, Karina Vidovic, Björn Nilsson, Urban Gullberg
Haematologica. 2016 Sep, haematol.2016.148815.

Publications not included in this thesis:

**Anti-apoptotic quinolinate phosphoribosyltransferase (QPRT) is a target gene of Wilms' tumor gene 1 (WT1) protein in leukemic cells**

Tove Ullmark, Giorgia Montano, **Linnea Järvstråt**, Helena Jernmark Nilsson, Erik Håkansson, Kristina Drott, Björn Nilsson, Karina Vidovic, Urban Gullberg
Biochemical and Biophysical Research Communications. 482 (2017) 802-807

**The transcriptional coregulator NAB2 is a target gene for the Wilms' tumor gene 1 protein (WT1) in leukemic cells**

Helena Jernmark Nilsson, Giorgia Montano, Tove Ullmark, Andreas Lennartsson, Kristina Drott, **Linnea Järvstråt**, Björn Nilsson, Karina Vidovic, Urban Gullberg
Oncotarget. Accepted

# Acknowledgements

To Björn Nilsson, my supervisor, for programming advice, giving me extra chances and believing in my ability to get here.

To Prof. Urban Gullberg, my co-supervisor, for including me in his different projects and for trusting in my calculations.

To Carl Sandén and Tove Ullmark, for letting me be a part of their projects and for being absolutely essential for the ChIP-seq papers.

To Maurolitsa, Ram and Giorgia, for great collaborations and for patiently explaining their experimental set-up just one more time.

To past and present members of the BN group: Ellinor, Ildiko, Mikael, Magnus, Per-Ludvik, Neha, Bhairavi, Evelina, Angelica, Aitzkoa, Britt-Marie, Ludvig, Mina, Jenny and Anna-Karin, for stimulating discussions, fika, and laughter.

To everyone in the B13 crew, for making my time there enjoyable.

To my siblings, Lotta and Klas, for always being a reliable fall back line.

To Knut, for hikes, discussions on every subject, lots of tea and sharing life with me. I couldn't have done it without you.

## Populärvetenskaplig sammanfattning på svenska

Huvudsyftet med mitt avhandlingsarbete är att ta fram nya metoder för att tolka storskaliga datamängder, insamlade med storskaliga genetiska (genomiska) mätmetoder, med särskilt fokus på blodcancerforskning. Eftersom det finns cirka 20 000 proteinkodande gener i genomet, och antalet möjliga reglerförhållanden därmed är stort, så är det svårt att få en överblick över vilka gener som påverkar varandra. Genom att använda matematiska metoder och göra beräkningar med hjälp av datorer så går det att vaska fram en mindre mängd särskilt intressanta samband som en senare kan undersöka noggrannare.

Artikel I och II utgår från s k microarraydata, som kan visa hur starkt uttryckta enskilda gener är. I det här fallet kommer data från benmärg från patienter med akut myeloisk leukemia (AML), en allvarlig form av blodcancer. I benmärgen sker ständigt nytillverkning av blodceller, men hos AML-patienter tar cancercellerna över benmärgen och tränger undan de friska blodcellerna, vilket leder till blodbrist och utmattning.

I artikel I utvecklade vi en metod för att bättre förstå hur olika geners uttryck samvarierar. Genom att kombinera olika matematiska och beräkningstekniska metoder har vi utvecklat en effektiv metod, programmet Ultranet, för att indikera samband mellan olika gener. Vi visar att resultaten är rimliga genom att titta på benmärgsdata och på genen GATA1, en gen som är aktiv under blodbildningen. Resultaten visar att Ultranet på ett snabbt och effektivt sätt kan hitta troliga samband, eftersom GATA1 visar sig ha samband med en rad blodrelaterade gener (kända sedan tidigare).

Artikel II fokuserar på hur en kan analysera genuttrycksdata från patientprover som inte består av en typ av celler, utan av en blandning av olika celltyper, vilket oftast är fallet. När en vill studera hur cellerna beter sig kan detta ställa till problem, eftersom det är svårt att veta vilka celler som ger upphov till vilken signal. Det blir lite som en orkester där varje instrument spelar sin egen melodi, men det är svårt att urskilja exakt vad cellon gör eller om det är tre eller fyra flöjtister (eller om de bara spelar ovanligt starkt). På cellnivå är varje stämma (instrument) en typ av cell medan melodin är vilka gener cellen använder. Vi har utvecklat en metod som utnyttjar att vi vet något om hur cellerna vi är intresserad av beter sig för att samtidigt bestämma vilka celltyper som ett prov består av och vad de uttrycker för gener. Genom att analysera sådana komplexa cellprover kan vi återskapa resultat från experiment där de har sorterat ut cellerna laborativt.

Artikel III och IV tittar närmare på hur två olika protein interagerar med DNA under blodbildning. Vi undersöker var i genomet som proteinet binder genom att med olika laborativa metoder rena fram små bitar av DNA som proteinet har bundit till. Sedan sekvenserar vi dessa fragment och letar med datorns hjälp upp var i genomet som de kommer ifrån.

I artikel III tittar vi på proteinet DEK och kommer fram till att det ofta binder till starten

av gener, vilket tyder på att det påverkar hur dessa gener uttrycks. I artikel IV tittar vi på hur två olika varianter av proteinet WT1 skiljer sig från varandra i hur de binder DNA. Det som skiljer dem åt strukturmässigt är att den ena varianten har en lite längre loop mellan två zinkfingrar som binder DNA. Detta gör att den ena formen, den med kort loop, binder till början av gener och tros reglera uttrycksnivån, medan den andra formen, med en längre loop, istället binder inuti genen.

# Bioinformatic approaches to gene expression in leukemia: Networks and deconvolution

# Abbreviations

| | |
|---|---|
| AML | Acute Myeloid Leukemia |
| ChIP-seq | Chromatin immunoprecipitation followed by sequencing |
| CMP | Common myeloid progenitor |
| CSC | Cancer stem cells |
| DEK | DEK proto-oncogene |
| DNA | Deoxyribonucleic acid |
| FACS | Fluorescence-activated cell sorting |
| GGM | Gaussian graphical model |
| GMP | Granulocyte-monocyte progenitor |
| HSC | Hematopoetic stem cells |
| K562 | Myeloid cell line |
| LIC | Leukemia-initiating cells |
| LSC | Leukemic stem cells |
| MEP | Megakaryocyte-erythroid progenitor cell |
| mRNA | Messenger RiboNucleic Acid |
| NNLS | Non-Negativity-constrained Least Squares |
| NOD | Non-Obese Diabetic (mouse strain) |
| *scid* | Severe combined immunodeficiency (mouse strain) |
| RNA | Ribonucleic acid |
| RT-PCR | Reverse transcriptase polymerase chain reaction |
| SICS | Sparse inverse covariance selection |
| U937 | Myeloid cell line |
| WHO | World Health Organization |
| WT1 | Wilms tumor gene 1 |

# Introduction

In recent years, genomic technologies like microarrays and next-generation sequencing have transformed the study of human malignancies. Now standard procedures in both research and in the clinic, genomics technologies generate enormous amounts of data at a rapidly increasing pace. Computational methods are needed to translate these into biologically and clinically relevant information.

This thesis focuses on novel methods to interpret complex genomics data, with an emphasis on methods to identify regulatory networks and cell markers from genome-wide gene expression data (Papers I and II). Related to this aim, a second theme has been the application of bioinformatics methods to analyze gene regulation using chromatin immunoprecipitation and massively parallel sequencing (ChIP-seq; Papers III and IV).

In terms of diseases, the focus has been on hematologic malignancies, particularly Acute Myeloid Leukemia (AML). This class of diseases is characterized by a rapid, uncontrolled growth of abnormal white blood cells (blasts) in the bone marrow. These cells outcompete normal blood cell formation, leading to anemia, thrombocytopenia, and compromised immunity. There are several different subtypes, defined by morphology and acquired genetic changes. While therapy has improved, the prognosis for most subgroups of AML remains poor with current therapies.

# Biological background

## Blood formation

Blood is one of the most actively replenished tissue types in the body. In healthy people, blood cells are continuously replenished from proliferation and successive maturation of hematopoietic stem and progenitor cells. Each day, somewhere in the order of 1 trillion new blood cells are produced[16]. Hematopoeisis is organized as a hierarchy with hematopoietic stem cells (HSCs) at the apex. HSCs have the ability to regenerate all of the hematopoetic lineages. There are several later stages that can generate a transient repopulation of the bone marrow.

Early work on mice determined that bone marrow cells, when transplanted into irradiated mice, colonized the spleen and showed differentiation specific to blood cells[80]. The same group indicated that there are two main branches of blood differentiation, the myeloid linage and the lymphoid linage[81]. Since then, an array of cell types and developmental paths have been described within these lineages, and, to this day, the debate on how to classify subtypes of blood cells continues. Functionally, human HSCs are still defined by xenografting into mice[18].

## Acute Myeloid Leukemia

Acute myeloid leukemia, AML, is a diverse group of blood cell neoplasms[86], characterized by abnormal, uninhibited growth of immature white blood cells, called blast cells, that accumulate in the bone marrow and suppresses normal blood cell formation, which leads to anemia, hemorrhage due to thrombocytopenia and compromised immunity[8, 49]. Each year about one million people globally are diagnosed with acute myeloid leukemia (AML) and the number seems to be increasing[85]. In Sweden, about 300 new cases are diagnosed each year, with the number of new cases per 100 000 being 3.33 for males and 3.25 for females (all ages included) in 2015[75]. Depending on acquired genetic lesions are present,

the 5-year survival rate for treated patients under 60 years of age varies from about 20% to about 75%. For elderly patients (above 60 years of age) the mean survival time for patients with favorable prognosis is about 3 years, while patients with an adverse prognosis survive less than 1 year[76]. The prevalence of survivors from acute myeloid leukemia in Sweden in 2014 was 13.7 per 100 000[33].

There are several different subgroups of AML based on what kind of genetic lesions, including translocations, DNA copy number aberrations, and point mutations, are present cite-Mardis2009. Historically, subgroups of AML were defined according to the FAB (French-American-British) system, based on cell morphology and cytochemical stains. Today the World Health Organization's (WHO) classification, first published in 2001[83], provides a classification scheme that also takes genetic markers into account. The latest major revision was done in 2008[8], with additional revisions concerning AML done in 2016[4]. The WHO classification is based on a combination of cell morphology, immunophenotyping, cytogenetics and molecular genetics[4, 8]. Leukemias tend to have a lower number of mutations per tumor compared to solid tumors. About half of all AML cases show chromosomal aberrations[37]. The most common chromosomal alterations include t(8;21)(q22;-q22.1) [*RUNX1-RUNX1T1*], t(9;11)(p21.3;q23.3)[*MLLT3-KMT2A*], t(6;9)(p23;q34.1)[*DEK-NUP214*], inv(16)(p13.1;q22)[*CBFB-MYH11*], inv(3)(q21.3;q26.2)[*GATA2, MECOM*] and t(1;22)(p13.3;q13.3)[*RBM15-MKL1*][4, 92].

A subgroup is acute promyelocytic leukemia (APL), a type of AML that has the fusion protein [*PML-RARA*], generated by a translocation (t(15;17)(q24.1;q21.2)) or its variant translocations[4, 52]. Patients with APL have a tendency for severe bleeding problems and untreated it is rapidly deadly[52]. However, this subtype is nowadays treatable with vitamin A analog (all-*trans* retinoic acid, ATRA), as part of the fusion gene (RARA, Retinoic acid receptor alpha) encodes a vitamin A receptor. ATRA is usually combined with arsenic trioxide (ATO) in practice. Arsenic is thought to interact with the PML moiety in a way that leads to degradation of the fusion protein by the proteasome[52]. Treated APL have a good prognosis with complete remission rates of 90-100% and overall survival rates of 86-97%[52].

Of AML cases, about 40-50% have a normal karyotype[55, 89], and in 25% of the cases no mutation can be found in a RT-PCR screening of previously diseased-linked genes[70]. Common genes with prognostic implications mutated in patients with normal karyotype are *NPM1, FLT3, CEBPA, MLL,* and *BAALC*[55].

Some mutations have a significant impact on disease prognosis. The prognosis for patients with mutated Wilms tumor 1 (WT1) is generally worse. For patients with normal karyotype around 5 percent have mutated WT1, while the prevalence rises to around 10 percent for patients with either APL or inv(16)(p13.1;q22); [*CBFB-MYH11*][36].

AML cases have a relatively low number of mutations in comparison to solid tumors[37, 43].

The number of mutations per patients was in one cohort of 200 patients found to be an average of 13 mutation in coding regions of the genome, out of which an average of 5 mutations occurred in genes that are recurrently mutated in AML[43]. The number of mutations per patient can rather be related to the patient's age, suggesting that the mutations have been acquired before any disease-causing mutations[89]. In experiments isolating healthy, residual hematopoietic stem cells from AML patients several pre-leukemic mutations were found. This indicates that some of the mutations found in leukemic stem cells are not in and of themselves pathogenic[30].

AML is an aggressive disease, which for patients ineligible for cytotoxic therapy has a mean survival time of 5 to 6 months[15]. Apart from the advances in the treatment of the APL subgroup, there has been no large changes in overall survival during the last three decades[65]. The basic treatment, for patients who can tolerate intensive treatment, is cytarabine (Ara-C) combined with an anthra-cycline[15, 69]. These drugs target DNA replication and thereby primarily eradicate proliferating cells[69]. For patients up to about 60 years of age 60-85% respond to chemotherapy[15]. In some cases, allogenic hematopoietic-cell transplantation is used.[15] However, relapses remain common and the reported longterm survival ranges from 40-50%[63] to less than 30% for adult patients[31].

## Leukemic stem cells

The malignant cells in AML are not a homogeneous population, but display genetic, phenotypic and functional heterogeneity. A subset of the cells, termed leukemic stem cells (LSCs) or leukemia-initiating cells (LICs), are thought to have properties that distinguish them from other cancer cells or healthy blood stem cells. They differ from the bulk of leukemic cells by being able to recapitulate leukemia on transplantation into mice. LSCs need to have a self-renewal capability, implying asymmetric division, producing one stem cell-like cell and one more differentiated cell. In order to recapitulate leukemia, LSCs need to be able initiate tumor growth and to give rise to a clonal longterm repopulation of the bone marrow[32, 61].

The idea that cancers are hierarchically organized with a stem cell-like apex first emerged during the 1970s. Because cancer treatment targets rapidly proliferating cells, the cancer stem cell hypothesis was suggested as a mechanism for treatment failure, as the stem cell-like cells are expected to have a low proliferation rate[60].

In theory, leukemic stem cells are cells have been mutated so that they can give rise to all the cell types of different maturation in the leukemic cell mass. In practice, the niche and epigenetic modifications also play a major role, since all cells in a tumor have very similar genetic material, but only some can be classified as stem cells[61]. The gold standard for

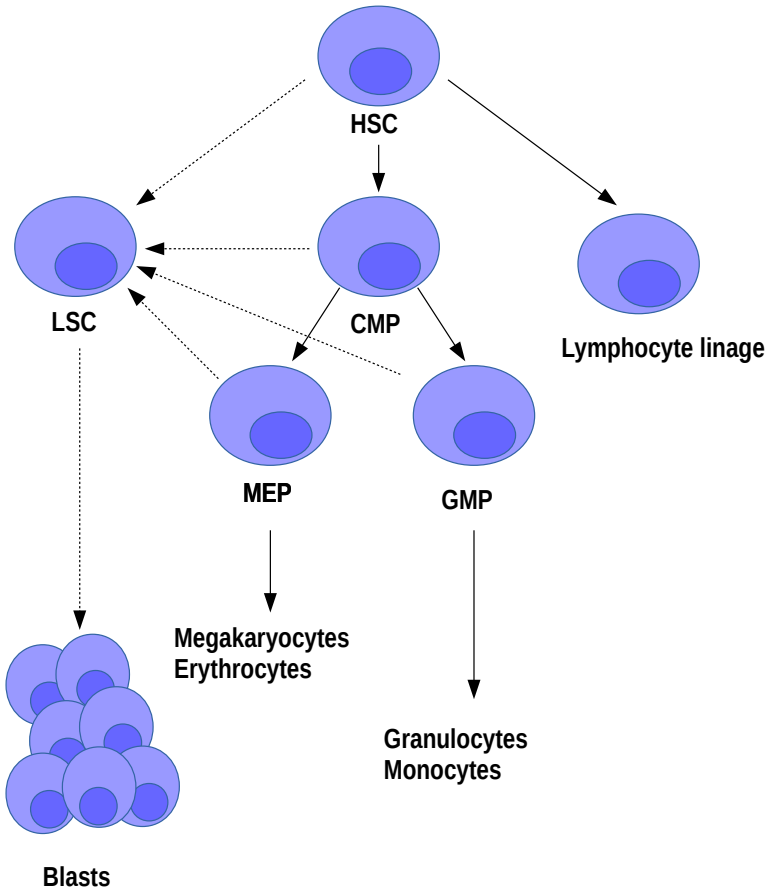


Figure 1: Simplified picture of the hierarchically organized hematopoiesis. The LSCs can originate from any of a number of early stages of blood differentiation.For AML the LSCs comes from early stages of the myeloid linage, and the bone marrow gets overpopulated with immature blast cells. The lymphoid linage is more complex than shown in the illustration. HSC: Hematopoeitic stem cell. LSC: Leukemic stem cell. CMP: Common myeloid progenitor. MEP: Megakaryocyte-erythrocyte progenitor. GMP: Granulocyte-monocyte progenitor.

establishing this is to use functional repopulation assays in mice to confirm long-term clonal maintenance. In many contexts this is not possible and *in vitro* assays are used instead[37]. The leukemic cells derive from an altered version of one of the early stem or progenitor cells. They need not be transformed from the most multipotent hematological stem cells, but can derive from somewhat later stages[53]. There are also indications that AML stem

cells are more differentiated than healthy, but has regained the limitless self-renewal[40].

There are two major models of how the tumor cells differentiate, which posits either a hierarchical model or a stochastic model. In the hierarchical model the stem cells develop into cells that loose their self-renewal, and most cells do not have tumor-initiating capabilities. In the stochastic model, whether a cell has stemness or tumor-initiating capabilities is more evenly spread throughout the tumor and each division leads to either self-renewing or non-self-renewing "by chance". None of the models capture completely the behavior of leukemias and the truth is probably in a combination of the models[57, 61].

Most of the cancer cells are not LSCs, and LSCs, due to their low proliferation rate and small population, can survive cancer treatment and repopulate the bone marrow leading to relapse. Achieving total eradication of the leukemic cells is thought to be complicated as LSCs are thought to divide rarely, and therefore resistant to chemotherapeutics that primarily target cell division[31, 69].

The origin of LSCs are disputed, and can vary depending on patient[65, 78, 88].The LSCs can often be purified in the $CD34+38^-$ fraction[69], which is the same fraction that enriches for healthy hematopoeitic stem cells[31]. While the $CD34+38^-$ fraction patient samples most of the time contain LSCs, they can also be found in other fractions[65, 78]. This indicates that LSCs can in some cases derive from more differentiated cells, shown by different cell markers depending on patient[88]. A high proportion of $CD34+38^-$ cells indicates a poor prognosis[65].

There are a number of proposed immunophenotypic differences between LSCs and HSCs. The leukemic cells have higher expression of CD123, TIM3, CD47, CD96, CLL-1 and IL1RAP, while they are deficient in CD90 and CD117, as reviewed in [88]. These differences can be potential targets for further therapies.

The gold standard for establishing the stemness of a tentative LSC population is to use engraftment studies in mice. This functional assay was first suggested in the early 1990s[24], and the first transplant of $CD34+38^-$ cells isolated from AML patients into severe combined immuno-deficiency (*scid*) mice was done in 1994[42]. This fraction was later shown to be the only fraction that has engraftment ability in Non-Obese Diabetic(NOD)-*scid* mice when injecting mononuclear cells from patients[10]. By dilution series it was shown that between 0.2-100 engraftment events happened per million mononuclear cell depending on patient. This supports the theory that only a small fraction of the leukemic cells have a tumor initiating ability[10].

While $CD34+38^-$ cells remain the most common selection procedure for LSCs, further characterization has added $CD90+45RA^-49f+$ as further detailing. The fraction of LSCs in a sample also varies considerably between individual patients from 0.3-1% LSCs in unsorted AML cells, up to 25% in extreme cases. The amount can have some dependence on subtype

of AML[91]. Hope et al[26] suggests that CD34+38$^-$ cells have the ability to induce a transient response but not long-term repopulation. These transient responses do fade out over time and do not last more than 12 weeks[26]. Ishikawa et al[28] showed that CD34+38$^-$ cells could still initiate leukemic response after serial transplants in mice. They also showed that the leukemic cells homed to the endosteal surface in femur and to the spleen. The LSCs were resistant to Ara-C treatment and could induce relapse[28].

Mouse models which can more easily be engrafted have been developed over time. In NOD-*scid* mice there are no mature T or B cells and low levels of natural killer (NK) cells. They have a higher level of human cell engraftment, which make them useful in studies of stemness[74]. By treating them with antibodies against CD122 (interleukin-2 receptor), the NK cell population can be further disrupted[18]. There are mouse strains which have the IL2RG gene deleted or trunkated and thus have no NK cell activity; they do however lack human specific cytokines, which can influence the viability and development of human cells[74]. In studies where NOD-*scid IL2rg*$^{-/-}$ mice were transplanted with human CD34+38$^-$ cells, the bone marrow repopulated after treatment with cytarabine and it was shown that this treatment selectively eliminates cells in S-phase. By using granulocyte colony-stimulating factor (G-CSF), the LSCs were moved from resting into more active phases of the cell cycle. When treated with AraC, the mice that had been exposed to G-CSF had less relapse of leukemia than mice not pre-treated[69]. The mouse models may only give part of the answer. Some argue that engraftment assays underreport the ability of LSCs to initiate leukemic growth, since they can be dependent on human factors found in the specific environment in human bone marrow[91].

# Computational background

## Network inference

A cell is an astoundingly complex system that in many ways is a challenge to model. A common approach to modeling is to use differential equations to describe the dynamics of the system. In engineering applications, the physics of the system are often easier to model than the complex biochemistry and biology of cells. In many cases, the structure of the interactions becomes the object of study. Finding this structure in large systems, such as the gene expression of a cell, is a challenging task even when the system is described using linear models.

This section will start with a discussion of how to structure a model and discusses a number of different approaches that have been used. The section continues with some examples of uses of models and ends with a discussion of some future challenges in the area.

### Model structure

Building a model of a system is to establish which parts the system contains and in what way they interact with each other. After a (non-trivial) mathematical re-formulation, determining the structure of the model can then lead to considerations of what to measure in order to find the parameters of the model that best describe how the system reacts. When it comes to living cells, however, the limitations on measurability makes estimation of the parameters of the system a difficult task. There are many levels of interaction in a cell and, obviously, even more if the system under consideration is the whole body. One possible starting point is to look at the levels of mRNA measured for a large set of genes and to find the covariance between expression levels. This can give insight into which genes are part of the same subsystems. However, if a time series of data points is not used, the causality of the interactions can not be determined.

A gene interaction network can be seen as a net where each node is a gene, measured as the

prevalence of mRNA, and where each edge is a link between two genes. The problems of model structure boils down to determining which of the links are present. In most methods mentioned in this thesis the graphs are undirected, meaning that the sequence of the proteins is not known[11]. Directed graphs are a possibility, but have different constraints and require data from more than one time point to generate.

A network of gene interactions can be described as an inverse covariance matrix having non-zero entries where an interaction exists and zeros elsewhere. The number of genes active in a given condition is comparatively large, in the range of hundreds to thousands of genes. As the number of measured variables, $n$, increases the size of the covariance matrix grows as $n^2$ making the handling of the full-sized matrix computationally costly[14].

Sparsity has several advantages in being easier to store and handle than full matrices[17]. Biological systems have a tendency towards displaying sparsity[11]. The interactions are complex, but most gene products seem to interact only with a small number of other genes. Some genes, for example major transcription factors, interact with many other genes, but the general structure remains sparse.

## Types of network inference

The networks used to describe interactions in cells can take different forms. Four of the most common approaches are probabilistic models, information theory based models, correlation-based models and partial-correlation based models[3].

Bayesian networks are directed graphical models that that include probabilities in their definition of the edges. The network is not allowed to contain cycles, i.e. it is not possible to get back to the starting point by following the edges[90]. Bayesian inference work well for small networks, but do not tend to scale well. Disadvantages are the need to discretize the data, usually by binning[3], and the lack of ability to incorporate feedback loops in the models.

Another technique is to use an information theory based approach. ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) uses mutual information as a measure of dependency between genes and uses data processing inequality to determine whether there are any direct interactions between the genes[50]. Information theory-based methods have a tendency to generate to dense networks[56]. Mutual information have the benefit of being invariant to reparameterization[50].

Correlation-based models, such as WGCNA[41], uses the correlations calculated from experimental data. This matrix is then thresholded, using either a hard or soft threshold, to get a binary adjacency matrix which shows the presence or absence of an edge.

Partial correlation-based models uses the inverse of the correlation matrix, called the concentration matrix, to determine which edges are present. A non-zero entry in the concentration matrix is interpreted as an edge[3]. Since the number of variables (genes) is large in proportion to the number of samples, the correlation matrix is not invertible and specific computational approaches are needed[34].

## Optimization of partial correlation networks

Optimization is a process where a possible solution to a mathematical problem is found where no exact solution is possible. Limits to the ability to solve the problem can be introduced by the dimension of the problem, measurement errors making more exact solutions unreliable or mathematical formulation of the problem. The selection of solution is driven by different constraints which can be traced back to the physical system. When used on typical gene expression data sets, the direct mathematical solution process is computationally intensive to degree of not being possible to solve[54].

The number of parameters in a covariance model increases proportional as $0.5(p + 1)$ where $p$ is the the number of measured factors. The variables can be modeled as belonging to the exponential family, which are log linear, meaning that logarithm of the probability density follows a linear model. The entropy, a measure of simplicity of the model, can then be maximized by selecting a empirical correlation matrix so that regularized inverse correlation matrix contains zeros where no relation exists[14].

Constraints on the optimization process are used to exclude mathematically feasible solutions that are physically or biologically non-feasible. In bioinformatic applications these are decided by a combination of the biological system and the measuring set-up. One of the most used constraints is a non-negativity constraint on, for example concentration or amounts of different proteins. Constraints that limit the number of non-zero elements are also commonly used, as the average number of interactions per protein or gene is expected to be low.

The non-negativity constraint is applicable in many different systems, as keeping all variables at zero or positive values gives an interpretation consistent with many types of physical restrictions. This type of restriction is used in non-negativity-constrained least squares (NNLS) problems.

When it comes to the computational execution of the algorithms, the dimensionality of most collections of microarray data pose a problem. The number of genes (variables) are in most cases large and the number of samples are proportionately small. Simulation studies confirm the intuitive assumption that a higher number of samples lead to higher accuracy, but the effect saturates for high sample numbers[3].

## Gaussian graphical models

Gaussian graphical models (GGMs) is an approach to inferring partial correlation networks. The approach was first used by Dempster in 1972 and assumes that the parameters measured are drawn from distributions that are normally distributed and then use their natural logarithms[14]. The original method was developed for problems that have a small number of variables and a relatively large number of samples. In most current applications to genetic data this is not true; the most common case being lots of variables, e.g. gene level expression, and a relatively small number of replicates.

The determination of the inverse covariance matrix has two different elements: finding the entries that should be zeros and calculating the value for the non-zeros. Determining the values for the non-zeros can be done by a combination of likelihood equations and Cholesky decomposition[34].

The lasso ('least absolute shrinkage and selection operator') is a method for estimating a log-likelihood function and introducing a L1-norm to promote sparsity[79]. The graphical lasso is further development of this method for solving NNLS problems developed by Friedman et al[20]. For problems with large number of variables, the running time for lasso can be considerable[35].

Sparsity of the inverse covariance matrix is a desirable feature from both a mathematical and a biological perspective. In order to promote sparsity a $L1$ constraint has been used in addition to the log-likelihood used for estimating inverse covariance[6].A comparison between different methods for building Gaussian graphical models finds that lasso based models have a tendency to create too dense networks, whereas shrinkage-based models tend to become overly sparse[35].

One approach to solving large optimization problems is to subdivide them into smaller and thus more easily computed problems. Block coordinate-descent is a method for subdividing the problem by sequentially solving each column and row individually while keeping the rest of the solution temporarily constant[6].

## Challenges

Newer methods for measuring the levels of RNA transcripts can be used in a similar way to microarray data. The challenge is to go from individual reads to a good measure of the total amount of mRNA related to a specific gene. Proteome data might also be used to construct similar types of networks[51]. There is also data that on a broad scale measures the interactions of a protein with DNA (see papers III and IV), that if a large number of experiments are combined can give more insight into how interaction network function.

One of the main challenges in bioinformatics is the handling and processing of the ever increasing amounts of data. The different types of data could be exploited in parallel to best extract the information in them. The large number of parameters measured and the usually small number of samples in comparison can lead to computationally expensive and badly conditioned problems.

One problem with many methods is the lack of theoretically grounded ways to fine-tune the parameters. As shown in Friedman et al[20], the value of the penalty parameters can give networks with different number of edges and the judgment of which level is realistic is left to the user.

The behavior of a biologic system varies depending on the conditions it is exposed to. Including data from different conditions or a time series can show interactions not caught by a single time-point[51]. Comparing networks inferred from data from different conditions (e.g. healthy/diseased) can indicate pieces of the disease process.

## Practical applications

The practical use of these types of methods is to analyze the data and give input for selection of follow-up experiments. These can be found in many areas, including chemometrics[82] and microarrays[77]. The indicated interactions can then guide further experimental work. We will focus primarily of microarray data.

One approach to network inference, that is possible when computational time is not a limiting factor, is to use a 'wisdom of crowds' approach. In this case the results from a number of different methodologies are combined and weighted according to their known strengths and weaknesses[48].

By using data measuring gene expression in yeast across a range of different conditions, mostly single gene deletions, Rung et al[68] created gene disruption networks showing which genes changed expression due to deletion of a single gene. The networks, however, do not separate direct and indirect effects[51].

Some networks are expected to be sparse, and there needs to be mathematical formulations in the cost function to take this into account. One approach is to use a L1 penalty to select which links to retain in the network. Banerjee et al[5] develops a model for handling problems using block coordinate descent. This approach optimizes over one column and row at a time, solving a series of small problems repeatedly until convergence is achieved. Another approach is to use an added cost based on the cardinality, i.e. the number of nonzero entries[13]. In both cases a parameter is used to balance the influence between the sparsity's weight and the rest of the cost function.

Bien and Tibshirani[9] developed a method that, instead of focusing on the inverse covariance matrix, finds the covariance matrix where zeros can be explained in terms of marginal independences.

# Deconvolution

Deconvolution is an *in silico* approach for analyzing gene expression in heterogeneous samples. Tissue samples generally contain more than one cell type. This means that the measurements done on a sample will be measuring a combination of signals from the constituent cell types weighted by their abundance.

There are methods for physical separation of samples into sub-populations, including fluorescence-activated cell sorting (FACS) or magnetic bead-based cell sorting (MACS), microdissection using a capillary pipette[67], or laser capture microscopy[87]. The isolated cells are then analyzed by gene expression profiling. The techniques are work-intensive and also leave the cells exposed to harsh *ex vivo* conditions for considerable amounts of time, which stresses the cells and may affect the quality of gene expression measurements.

A newer approach is single cell mRNA sequencing, where single cells are sequenced separately. The process starts with very small amounts of mRNA present in a single cell, about 1 pg per cell[87]. The resulting data has a tendency to have high noise levels and the amplification step can also introduce distortions, as well as being sensitive to contamination[62, 87]. Although these methods have transformed our ability to identify cell type-specific gene expression patterns, they are comparatively labor- and cost-intensive. In addition, these methods may also affect gene expression patterns due to prolonged handling of the cells *ex vivo*[84, 46].

As a complementary approach to methods that physically separate cells, the analysis of data generated from complex samples can be improved by taking heterogeneity into account. If results from complex samples are used to draw conclusions about one sub-population of the sample the situation assumed in the analysis is not actually present in the sample, giving unnecessarily large error margins. The most basic model uses a linear combination of the abundances of the cells together with the typical gene expression pattern. This approach assumes that we can know something about the cells in advance. However, this is not always the case and variations between samples (e.g. patients) are to be expected.

## Modeling the mixture

From a signal theory perspective, it is important to have an idea of how the different signals combine to form the output registered from the microarray assay. The signals or a transformation of the signals being added linearly is the most often used model. This gives a model that can be expressed as

$$A = WH + \epsilon$$

where A is a $m \times n$ matrix containing the measured gene expression values, W is a $m \times p$ matrix showing the (unknown) gene expression per cell type, H is a $p \times n$ matrix showing the relative amount of each cell type, and $\epsilon$ is the noise due to the measuring process. Without additional constraints on either W or H, there is an infinite number of mathematical solutions. To reduce the solution space, penalties are used to limit the possible solutions to those compatible with the biological and physical properties of the system, e.g. only accepting solutions having positive amounts for cell abundances.

All models are based on some set of assumptions about the model system, some of which can be more or less accurate. Some methods include assumptions about the data that can be seen as unrealistic, e.g. total lack of correlation between cell types[84]. Others are limited by just being able to separate the sample into two subpopulations, for example separating the samples into cancer and non-cancer cells[2].

Most approaches use extra information about the experimental system to guide the estimation. Depending on the kind of information available the output of the methods are used to fill in the gap in either proportions or cell expression profiles. Some methods estimate both using constraints.

When it comes to data from microarrays, a common processing step is to log-transform the data before handling it further. This has some advantages in handling outliers and gives gene expression values a more normal distribution. However, if using log-transformed data in a deconvolution setting, subject to the constraint that the frequency of all cell types should sum to 1 in each sample, the result will always be a lower estimate than the true value[95]. In assessing the false discovery rate through permuting the data, Shen-Orr et al[72] found that which approach is best seems to be dependent on the particular data sets used. The use of log-transformation can affect the rank ordering of significant genes, when the data is used to determine differentially expressed genes[45].

This transformation of H also makes the interpretation of the results more difficult. In order for the mixing weights to be directly interpretable in a natural sense, they need to be limited to non-negative numbers and the sum of them should ideally be 1. This complicates the mathematical solutions, since in many cases the mathematically optimal solution would include negative cell fractions. A strict sum-to-1 restriction for the cell abundances is introduced by Gong et al[23]. Allowing a slight deviation from 1 can be compatible with a modeling of noise.

Generally, it is a good idea to include known information about the system in an estimation of some other property of the system. It is however important that this data is well-characterized, even if most methods allow some deviations from the guide. A schematic of the different inputs and outputs that has been included in deconvolution of microarray data can be seen in Fig. 2. The data available before the deconvolution includes the data from the mixed samples. Furthermore, information regarding cell types in the sample such

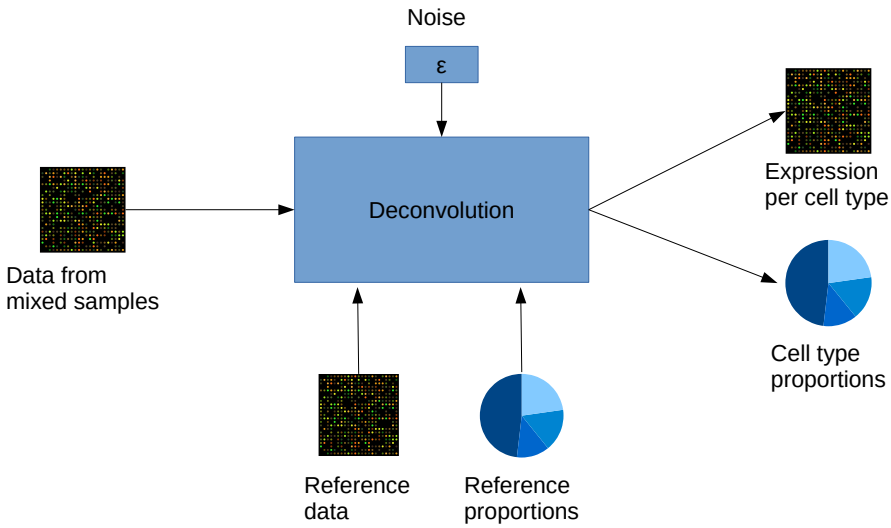


**Figure 2:** Schematic picture of the work flow during deconvolution. The input (on the left) are measured data from complex samples. Also needed as input are some kind of guiding data (from the bottom). The guides can be expected proportions, number of cell types or gene expressions per expected cell type. The process can also introduce or amplify some noise (top), that might also be present within any of the inputs. The deconvolution process is represented as a box and can be any of the deconvolution methods mentioned in the text. Outputs (on the right) are proportions of each cell type in each sample and/or expression profiles for each cell type. Note that not all methods discussed need all of the possible inputs nor do they necessarily generate all of the mentioned outputs.

as number of cell types, expression from pure samples or estimations of the proportions of cell types can be included in the deconvolution scheme.

## Previous work on deconvolution

Venet et al[84] were the first to show that mathematical deconvolution of complex samples is possible given circumstances where an assumption of no correlation between cells can be used. They found that they could recreate the amount of muscle tissue present in samples from colon cancer. The study also gives predictions of genes associated with each cell type.

Abbas et al[1] were the first group to try to deconvolve blood samples. They used a method based on Gene Set Enrichment Analysis (GSEA) to weight the gene probe set to take more highly differentially expressed genes more into account in the cells of interest. They solve the problem as a linear least squares problem. After showing that the method could deconvolve a known mixture of cell lines, they continued to look at leukocytes and found them to correlate well with complete blood counts. Finally they compared systemic lupus erythematosus (SLE) with healthy samples and found that the SLE patients had higher numbers

of activated NK cells, which correlated well with the data from complete blood cell counts.

A different approach called DSection uses a Bayesian model to estimate a prior containing the behavior of the constituent cell type. This prior is then used to estimate the composition of a given sample. The use of a prior allows the expression in each sample to vary slightly, which is realistic in considering the noisy signal. They show that this works better than simple linear regression in cases involving a small number of sufficiently different cell types[19].

In addition to methods guided by expected expression profiles or expected cell type abundances, there has also been unguided approaches. Principal component analysis (PCA) has been tried as a tool to group samples, but the outcome has not been sufficiently better than clustering based on gene expression without PCA[93]. Depending on the differences between the cell types that one wishes to separate and the information available on the contents of a given sample a method independent of previous information can be attractive. PCA can also be used as a part of the method, as in Lähdesmäki et al[44], where it is used to reduce the number of parameters by grouping genes in their principal components.

## Challenges

Deconvolution offers new ways of looking at complex samples. However, there are limitations that have not been sufficiently explored and need to be addressed if the methods are to be used in more demanding situations than exploratory investigations.

The limits of detection when it comes to proportionately small cell subpopulations are currently unexplored[71]. Depending partially on the nosiness on of the signal, small populations can be hard or impossible to detect. Finding the detection limits is an important issue for further consideration.

Most of the methods discussed above, e.g. [84, 94], only apply their algorithms to samples assumed to contain only a small number of cell types. Some methods instead try to characterize the tissue types in the sample, e.g. [2, 19, 23, 96], where each tissue is known to contain several cell types. Blood, being one of the more common sample types, contains about 20 cell types, depending on how you classify them, and some of them in very low frequencies. There is also a large variation between patients and conditions[66].

In practice, the number of cell types possible to detect is related to the detection limit for small cell type populations. A sample containing numerous cell types in small proportions present a different challenge than a sample containing mostly one cell type but a small number of a different cell type. Finding the detection limits is an important challenge for further use of these methods.

## Practical applications

Differences in expression for a specific cell type for different disease conditions can be hard to detect in mixed samples. Shen-Orr et al[73] used known mixtures of liver-brain-lung tissue to validate that a linear combination of microarray values works as a model for signal combination. They also identified differences in gene expression in the leukocyte fraction of whole blood from kidney transplant patients with and without transplant rejection. These differences were not detectable without deconvolution.

One method including an extra step for identifying the number of cell-types has shown some success when the cell types that the sample contain are sufficiently different[96], something that however is not the case in differential hierarchy of the bone marrow. The groups, however, are hard to interpret and tend to cluster around functions rather than cell types[64]. An algorithm based on Bayesian linear unmixing found better performance measure in mean square error of the deconvolved answer, but the factors found did not correlate to cell types, but rather to cell functions[7]. This makes unguided approaches unsuitable for determining the expression of specific cells, while they can have some utility in looking at the roles of a tissue as a whole.

Machine learning is an area that has had influence over the development of bioinformatics. One example is a deconvolution method using linear support vector regression, a machine learning derived-technique, and also does a select of the most informative subset of genes to use for estimating blood cell fractions and expressions from microarray data[58].

While the methods mentioned above use microarray data for their investigations, other types of high-dimensionality data can potentially be used in a similar manner. Several methods use data from RNA-seq, a high throughput sequencing method, and assumptions on the proportions of cell types in the sample to estimate the abundance of each transcript per cell type[22, 94]. With adding preprocessing of the sequencing data, it is possible that more of the methods presented above can be adapted to handle RNA-seq data or similar data types. There has also been efforts to make the methods more user friendly[21].

# Main results of the research papers

This section starts by describing the aims of the studies. The second part gives a brief outline of the methods used. It is followed by a discussion of the results from our investigations separatedly. The section ends with a general discussion and future research prospects.

## Aim

The overall aim of the thesis is to develop new methods to extract information from high-dimensionality data, and apply these to data from AML. The methods used are *in silico* processes, using experimental data as a foundation for discovery of potential interactions that in a later step can be verified experimentally.

Paper I aims to develop a computationally efficient method for finding correlations between expression levels of different genes using optimization to arrive at a sparse inverse correlation matrix. We look at the interactions in a collection of data sets from AML patients to validate the method. We also evaluate our method against several other methods in the field.

Paper II focuses on finding expression and proportions of cells simultaneous from complex samples containing different cell types. We develop an *in silico* approach to find new insights from available microarray data from AML.

Paper III aims to find the DNA-binding pattern of the oncoprotein DEK and to find a function for DEK in myeloid linage cells.

Paper IV aims to determine the functions of *WT1*, a gene encoding a transcription factor previously shown to be important for a number of different organ systems. We focus on the transcriptional functions of two isoforms of the WT1 protein.

# Laboratory methods

## Gene expression profiling

Transcribed genes are represented in the mRNA population of that cell. One way of capturing RNA expression is to use DNA microarray chips. The RNA is purified and reverse-transcribed to generate a more thermodynamically stable population of cDNA fragments. Fluorescent dyes are added to the sample and the sample fragments are hybridized to probe sequences on a microarray chip. Probe sequences are arranged in a spot pattern on the chip and the results are read out by measuring the fluorescence of each spot[27].

The presence of RNA detected in a microarray is used a proxy for gene activity in the cell. The data is normalized to a log2-normal distribution to minimize the effects of outliers.

In our experiments, we used the cell lines U937 (for paper III) and K562 (for paper IV). The cell line U937 is derived from histiocytic lymphoma and shows myeloid characteristics. The cell line K562 is derived from a chronic myelogenous leukemia patient in blast crisis and carries the Philadelphia chromosome[47].

With the advances in sequencing technologies RNA-seq has taken over as the method of choice. After isolating and reverse-transcribing the RNA to cDNA, the fragments are sequenced and then mapped to the genome in a way that allow for gaps in the sequence to take introns into account. RNA-seq has the advantage of not requiering any previous knowledge of a sequence to be able to detect it.

## ChIP-seq

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a method to determine where a protein binds to the genome. It is used to look at DNA associated proteins, which includes proteins that are directly DNA binding or a part of a protein cluster that binds to DNA. By cross-linking the proteins before sonic fragmentation of the DNA and then using immunoprecipitation to isolate the protein-DNA complexes containing the protein of interest, the sequences close to the putative binding site are isolated. The proteins are then removed and the short DNA fragments are sequenced[39].

The resulting sequences are mapped to a standard genome. The loci in the genome where the sequences differ from the background are identified as "peaks". We measure the distance from the peaks to identified transcription start sites.

## Methods

In paper I, we used microarray data compiled from AML cases[25]. We normalized the data to a lognormal distribution and gene networks were inferred by calculating the inverse covariance matrix was calculated using our developed method, Ultranet.

In paper II, we developed a *in silico* method to discern the expression of subgroups of cells in complex samples. We used data from sorted blood cells[59] to guide the analysis. We also used a compilation of blood and bone marrow samples containing several cell types from different leukemias[25].

In paper III, we used the cell line U937 and did ChIP-seq experiments using antibodies against DEK and H3K4me and compared the results with pre-existing ChIP-seq and DNA hypersensitivity data from the ENCODE project and also CAGE experiments for the same cell line.

In paper IV, we used the cell line K562 and created clones expressing biotinylated WT1 -KTS and WT1 +KTS, respectively. We did ChIP using streptavidin capture followed by sequencing. The data was processed in a similar way to paper III.

# Results

**Paper I: Ultranet: efficient solver for the sparse inverse covariance selection problem in gene network modeling**

Genes are regulated in a complex interplay of different factors. Regulatory relationships can be identified through network modeling techniques. By treating different gene expression levels as variables it is possible to view this as a high-dimensional statistical inference problem.

Network modeling can be used to identify potential regulatory relationships between genes. Identifying gene regulatory networks from gene expression data at a global level can be viewed as a high-dimensional statistical inference problem. In Paper I, we developed a network reconstruction algorithm based on Graphical Gaussian Models (GGMs). To establish such models we solve for the inverse $\Theta$ of the correlation matrix S, and then compute the partial correlation matrix $P^*_{ij} = \Theta^*_{ij} / \sqrt{\Theta^*_{ii}\Theta^*_{jj}}$. Because of the high dimensionality of genomics data, computationally efficient methods are needed to calculate solutions to SICS in practice. The SICS problem can be defined as,

$$\text{maximize } \log\det\Theta - \text{tr}(S\Theta) - \lambda\|\Theta\|_1$$
$$\text{s.t. } \Theta^T = \Theta, \Theta \succ 0$$

where $\Theta$ is a regularized inverse correlation matrix, S is the empirical correlation matrix for the input data and $\lambda \in [0, 1]$, through the $L1$ norm, controls the sparsity of the solution.

We developed a novel algorithm to solve SICS problems efficiently, and implemented it as a software tool, Ultranet. Our method combines several mathematical and programming techniques that exploit the structure of the SICS problem and enables computation of genome-scale GGMs without compromising analytic accuracy.

Conclusions:

- We have established a method for estimating $\Theta$, the inverse of the sample correlation matrix S. Our method, Ultranet is implemented in C++.
- Ultranet solves the SICS problem significantly faster than previous methods.
- When looking at a known hematopoeitic transcription factor, GATA1, Ultranet finds an environment enriched in known hematopoietic genes. Most neighboring genes were also present in ChIP-seq experiments for GATA1, illustrating how GGM-based inference can identify functionally relevant interactions.

**Paper II: Deconvolution of gene expression in cancer cell hierarchies**

Leukemias and some solid tumors are organized as cell hierarchies, sustained by a population of cancer stem cells (CSCs), or leukemic stem cells (LSCs) at the apex[18]. These cells give rise to more differentiated cells in the tumor, and have the ability to transfer disease in xenotransplantation assays. Additionally, CSCs and LSCs are thought to be resistant to commonly employed cancer therapeutics, and can thus cause disease recurrences[29].

Traditional approaches to characterize CSCs and LSCs rely on physical cell sorting, a laborious procedure. In Paper II, we developed a new method to study gene expression in cancer cell hierarchies by using large-scale mathematical deconvolution of gene expression profiles of unsorted tumor cells. The method is based on pre-determined expression profiles for cells similar to the expected populations, and applies a $L1$-constrained optimization to find the expression of the disturbed cell hierachies. Following initial simulation experiments, we tested our method against a gold standard assay based on *in vivo* xenografting of sorted cell fractions and independently recover markers for AML stem cells.

Conclusions:

- We developed a method to infer, from mixed samples, the expression patterns of the constituent cells. The method, tdeconv, uses typical profiles for the cell types expected in the sample to ensure that the inferred gene expression patterns can be interpreted as perturbed versions of the gene expression patterns of a corresponding normal cell type.
- In a proof-of-principle study, we found that the AML cell gene expression patterns identified *in silico* as perturbed versions of the gene expression patterns of normal HSCs overlapped with gene expression signature extracting in an *in vivo* xenografting study of $CD34+38^-$ AML LSCs[18].

**Paper III: The DEK oncoprotein binds to highly and ubiquitously expressed genes with a dual role in their transcriptional regulation**

This study looked at the genomic binding pattern of the DEK oncoprotein, that is highly expressed in a number of cancer types and is a known translocation partner in AML with t(6;9)(p23;q34.1); [DEK-NUP214]. Previously, it has been unclear whether DEK is a promoter-binding transcription factor or DNA-binding factor that is more generally distributed across the chromatin. It is known to be a part of human chromatin, but its function and binding properties have not been fully characterized. Our analyses showed that DEK preferentially binds transcription start sites, supporting that it is indeed a transcription factor.

Here we mapped the binding sites of DEK at a genome-wide level by ChIP-seq in the myeloid cell line U937 and combined it with epigenetic and gene expression analysis. ChIP-seq is an experimental method for selecting gene fragments bound by a protein that can be immunoprecipitated. The resulting fragments are then sequenced and aligned to the reference genome assembly hg19[38]. The goal is to identify genomic regions that the DEK binds to.

The bioinformatic part of the project included mapping of sequence reads to the reference genome assembly, peak calling to identify chromosomal regions showing enriched binding, comparison of binding sites between different experiments, identifying enriched binding patterns, and analyzing distances from the nearest transcription start site at both single read and peak level.

Using data from the Encyclopedia of DNA Elements (ENCODE)[12], we also compared the binding pattern of DEK with a large number of DNA-binding proteins and chromatin features.

Conclusions:

- DEK preferentially binds to open chromatin and to the transcription start sites of highly transcribed genes, suggesting that DEK plays a role in transcriptional regulation rather than in making up chromatin structure. The comparative study shows that the binding pattern of DEK shows similarities with the binding patterns of RNA polymerase II in several different cell lines.
- Functional studies showed that DEK knockdown by shRNA resulted in both significant up- and down-regulation of individual DEK bound genes, suggesting a dual role in the regulation of gene transcription.

**Paper IV: Distinct global binding patterns of the Wilms tumor gene 1 (WT1) -KTS and +KTS isoforms in leukemic cells**

This study looked at how an insertion of three amino acids changes the binding of Wilms tumor gene 1 (WT1) to DNA. We used K562 cell line and transfected them to induce expression of biotinylated isoforms of WT1 and *E. coli* biotin protein ligase (BirA).

Wilms' tumor gene 1 (WT1) is a zinc-finger transcription factor that acts as an oncogene in AML. An alternative splice-variant of WT1 removes the three amino acids KTS between zinc-finger three and four. A change in the balance between these isoforms has been linked to worse prognosis in leukemia. In Paper IV, we examined the genome-wide binding patterns for biotinylated WT1 -KTS and +KTS in leukemic K562 cells using ChIP-seq by streptavidin capture.

Using bioinformatic approaches similar to those in Paper III, we found that the WT1 -KTS isoform binds close to transcription start sites and enhancers, similar to other transcription factors, while WT1 +KTS binds within gene bodies. Motif searches revealed differences as to how the two isoforms previously reported WT1 motifs, and that some motifs are bound by both isoforms. We also show a large overlap of the genes bound by WT1 -KTS and +KTS respectively, but they bind to different parts of the same gene. Using CAGE data and H3K4me3, histone mark for active transcription start site, we find that both isoforms bind to actively transcribed genes.

Conclusions:

- The removal of three amino acids (KTS) changes the binding pattern of WT1. Both forms bind to actively transcribed genes but with different specificity.
- The WT1 -KTS isoform binds close to transcription start sites and enhancer. The WT1 +KTS isoform binds within gene bodies. Both isoforms bind to actively transcribed genes, sometimes the same genes but different positions, suggesting different functions.

# Discussion

Bioinformatics methods are required to interpret genomics data. Paper I and II have presented methodological improvements in the area of gene expression data analysis, whereas paper III and IV represent collaborative projects were previously developed methods were used to address specific questions in experimental hematology.

The generalizability of the methods developed in paper I and II is high. Ultranet can be used to handle microarray data from other biological systems, and with minor pre-processing of data can handle data from other high-dimensional measuring systems, for example single-cell sequencing.

A limitation of network inference in paper I is that it does not inform about the causality of the identified interactions, exemplifying the classical dilemma of causation and correlation. How the interaction takes place is not revealed through the network modeling, and there can be several different ways mRNAs or proteins can affect the presence of a specific mRNA type, ranging from direct causal relationships (e.g., the mRNA level of a transcription factor likely correlates with the mRNA levels of its targets) to pure false discoveries (e.g., links that just represent noise but were detected as a consequence of multiple testing).

The design of the microarrays can lead to unknown systematic errors. For example, while there are several probes per known gene the strength of the binding between the probe and the sample cDNA varies depending on the nucleotide content of the sequence and suboptimal probe design may affect the network. Microarrays can also introduce errors due to the non-linearity of the technique.

In paper II we have looked at the ability of tdeconv to find differential expression of genes in a small subpopulation of the sample. In simulations sampling from a given distribution with simple known changes in the expression of a few randomly chosen genes, tdeconv could find the changed genes could be identified. In a real data set, the underlying distribution varies from patient to patient and the noise is not necessarily evenly distributed.

One critical factor in applying the method developed in paper II is how small a cell type population can be and still be detectable. The noisy nature of microarray data gives a lower threshold of the effect possible to detect. Another important consideration is how different from each other the cell types are. Cell types that have similar expression tends in practice to be lumped into one cell type, as the other cell type proportions go to zero. The detection threshold is also dependent on the number of samples available and how homogeneous they are, as the expression profile of a given cell type is shared between all the samples. If the individual samples differ from the expected expression pattern in individual ways, this is more likely to be interpreted as noise.

The data used for paper I and II are collected from a large number of patients, considering

the number of patients affected. However, when looked at from a model identification perspective, the number of samples should ideally be high in proportion to the number of parameters (inverse correlation coefficients) to be identified. The proportionally very large number of measured variables makes the probability of false positives a concern. The networks will include false connections and the best use might be to find the strongest links and then experimentally further explore these linkages.

In paper III and IV cell lines are used as a proxy for leukemic cells. As a mass of cells were used, the cells are potentially in different stages of the cell cycle. As it is known that protein expression changes over the cell cycle, this can lead to extra noise in the data. Single-cell sequencing can be a way of getting less noise from variation in a population, but currently has a high monetary cost. There is also the question of how well immortalized cell lines represent the actual micro-environment and cell types present and interacting in the bone marrow of leukemia patients.

The cell line results have been compared with results from a wide variety of tissue types. While this can give a general sense of which kind of proteins have similar binding patterns, it is doubtful if a certain protein is active in the same way in all cell types it is present in. For WT1, different variants are transcribed in different contexts and the regulatory effect can be traced to the balance between isoforms.

# References


[1] Alexander R. Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F. Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE*, 4(7):1 – 16, 2009.

[2] Jaeil Ahn, Ying Yuan, Giovanni Parmigiani, Milind B. Suraokar, Lixia Diao, Ignacio I. Wistuba, and Wenyi Wang. Demix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865, 2013.

[3] Jeffrey D. Allen, Yang Xie, Min Chen, Luc Girard, and Guanghua Xiao. Comparing statistical methods for constructing large scale gene networks. *PLoS ONE*, 7(1):1–9, 01 2012.

[4] Daniel A. Arber, Attilio Orazi, Robert Hasserjian, Jürgen Thiele, Michael J. Borowitz, Michelle M. Le Beau, Clara D. Bloomfield, Mario Cazzola, and James W. Vardiman. The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405, 2016.

[5] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[6] Onureena Banerjee, Laurent El Ghaoui, Alexandre d'Aspremont, and Georges Natsoulis. Convex optimization techniques for fitting sparse gaussian graphical models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 89–96, New York, NY, USA, 2006. ACM.

[7] Cécile Bazot, Nicolas Dobigeon, Jean-Yves Tourneret, Aimee K. Zaas, Geoffrey S. Ginsburg, and Alfred O. Hero. Unsupervised bayesian linear unmixing of gene expression microarrays. *BMC Bioinformatics*, 14(1):1 – 20, 2013.

[8] Bryan L. Betz and Jay L. Hess. Acute myeloid leukemia diagnosis in the 21st century. *Archives of Pathology & Laboratory Medicine*, 134(10):1427–1433, 2010.

[9] Jacob Bien and Robert T. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[10] Dominique Bonnet and John E. Dick. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Medicine*, 3(7):730–737, Jul 1997.

[11] Hyonho Chun, Xianghua Zhang, and Hongyu Zhao. Gene regulation network inference with joint sparse gaussian graphical models. *Journal of Computational and Graphical Statistics*, 24(4):954–974, 2015.

[12] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[13] Alexandre D'Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

[14] A.P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[15] Hartmut Döhner, Daniel J. Weisdorf, and Clara D. Bloomfield. Acute myeloid leukemia. *New England Journal of Medicine*, 373(12):1136–1152, 2015.

[16] Sergei Doulatov, Faiyaz Notta, Elisa Laurenti, and John E. Dick. Hematopoiesis: A human perspective. *Cell Stem Cell*, 10(2):120 – 136, 2012.

[17] John Duchi, Stephen Gould, and Daphne Koller. Projected subgradient methods for learning sparse gaussians. In *Conference on Uncertainty in Artificial Intelligence, UAI*, 2008.

[18] Kolja Eppert, Katsuto Takenaka, Eric R. Lechman, Levi Waldron, Bjorn Nilsson, Peter van Galen, Klaus H. Metzeler, Armando Poeppl, Vicki Ling, Joseph Beyene, Angelo J. Canty, Jayne S. Danska, Stefan K. Bohlander, Christian Buske, Mark D. Minden, Todd R. Golub, Igor Jurisica, Benjamin L. Ebert, and John E. Dick. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nature Medicine*, 17(9):1086–1093, Sep 2011.

[19] Timo Erkkilä, Saara Lehmusvaara, Pekka Ruusuvuori, Tapio Visakorpi, Ilya Shmulevich, and Harri Lähdesmäki. Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571 – 2577, 2010.

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[21] R. Gaujoux and C. Seoighe. Cellmix: A comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, 29(17):2211–2212, 2013.

[22] T. Gong and J.D. Szustakowski. Deconrnaseq: A statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–1085, 2013.

[23] Ting Gong, Nicole Hartmann, Isaac S. Kohane, Volker Brinkmann, Frank Staedtler, Martin Letzkus, Sandrine Bongiovanni, and Joseph D. Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE*, 6(11):1 – 11, 2011.

[24] M. F. Greaves. Stem cell origins of leukaemia and curability. *British Journal of Cancer*, 67(3):413–423, Mar 1993.

[25] Torsten Haferlach, Alexander Kohlmann, Lothar Wieczorek, Giuseppe Basso, Geertruy Te Kronnie, Marie-Christine Béné, John De Vos, Jesus M. Hernández, Wolf-Karsten Hofmann, Ken I. Mills, Amanda Gilkes, Sabina Chiaretti, Sheila A. Shurtleff, Thomas J. Kipps, Laura Z. Rassenti, Allen E. Yeoh, Peter R. Papenhausen, Wei min Liu, P. Mickey Williams, and Robin Foà. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the international microarray innovations in leukemia study group. *Journal of Clinical Oncology*, 28(15):2529–2537, 2010.

[26] Kristin J. Hope, Liqing Fin, and John E. Dick. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nature Immunology*, 5(7):738–743, Jul 2004.

[27] Kenichi Inaoka, Yoshikuni Inokawa, and Shuji Nomoto. Genomic-wide analysis with microarrays in human oncology. *Microarrays*, 4(4):454 – 473, 2015.

[28] Fumihiko Ishikawa, Shuro Yoshida, Yoriko Saito, Atsushi Hijikata, Hiroshi Kitamura, Satoshi Tanaka, Ryu Nakamura, Toru Tanaka, Hiroko Tomiyama, Noriyuki Saito, Mitsuhiro Fukata, Toshihiro Miyamoto, Bonnie Lyons, Koichi Ohshima, Naoyuki Uchida, Shuichi Taniguchi, Osamu Ohara, Koichi Akashi, Mine Harada, and Leonard D. Shultz. Chemotherapy-resistant human aml stem cells home to and engraft within the bone-marrow endosteal region. *Nature Biotechnology*, 25(11):1315–1321, Nov 2007.

[29] Farhadul Islam, Bin Qiao, Robert A. Smith, Vinod Gopalan, and Alfred K.-Y. Lam. Cancer stem cell: Fundamental experimental pathological concepts and updates. *Experimental and Molecular Pathology*, 98(2):184 – 191, 2015.

[30] Max Jan, Thomas M. Snyder, M. Ryan Corces-Zimmerman, Paresh Vyas, Irving L. Weissman, Stephen R. Quake, and Ravindra Majeti. Clonal evolution of pre-leukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Science Translational Medicine*, 4(149):149ra118–149ra118, Aug 2012.

[31] Liqing Jin, Kristin J. Hope, Qiongli Zhai, Florence Smadja-Joffe, and John E. Dick. Targeting of cd44 eradicates human acute myeloid leukemic stem cells. *Nature Medicine*, 12(10):1167–1174, Oct 2006.

[32] Craig T. Jordan. The leukemic stem cell. *Best Practice & Research Clinical Haematology*, 20(1):13–18, Mar 2007.

[33] G. Juliusson, J. Abrahamsson, V. Lazarevic, P. Antunovic, A. Derolf, H. Garelius, S. Lehmann, K. Myhr-Eriksson, L. Mollgard, B. Uggla, A. Wahlin, L. Wennstrom, M. Hoglund, for the Swedish AML Group Group, and the Swedish Childhood Leukemia. Prevalence and characteristics of survivors from acute myeloid leukemia in sweden. *Leukemia*, 31(3):728–731, Mar 2017.

[34] Harri Kiiveri. Graphical models and multivariate analysis of microarray data. In Christine Sinoquet and Raphaël Mourad, editors, *Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics*, chapter 3, pages 85–104. Oxford University Press, 2014.

[35] Nicole Krämer, Juliane Schäfer, and Anne-Laure Boulesteix. Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10(1):384, 2009.

[36] M.-T. Krauth, T. Alpermann, U. Bacher, C. Eder, F. Dicker, M. Ulke, S. Kuznia, N. Nadarajah, W. Kern, C. Haferlach, T. Haferlach, and S. Schnittger. Wt1 mutations are secondary events in aml, show varying frequencies and impact on prognosis between genetic subgroups. *Leukemia*, 29(3):660–667, 2015.

[37] Antonija Kreso and John E Dick. Evolution of the cancer stem cell model. *Cell Stem Cell*, 14(3):275–291, Mar 2014.

[38] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, and Kevin McKernan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

[39] Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, Peter Bickel, James B. Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I. Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J. Hartemink, Michael M. Hoffman, Vishwanath R. Iyer, Youngsook L. Jung, Subhradip Karmakar, Manolis Kellis, Peter V. Kharchenko, Qunhua Li, Tao Liu, X. Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M. Myers, Peter J. Park, Michael J. Pazin, Marc D. Perry, Debasish Raha, Timothy E. Reddy, Joel Rozowsky, Noam Shoresh,

Arend Sidow, Matthew Slattery, John A. Stamatoyannopoulos, Michael Y. Tolstorukov, Kevin P. White, Simon Xi, Peggy J. Farnham, Jason D. Lieb, Barbara J. Wold, and Michael Snyder. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Research*, 22(9):1813–1831, 2012.

[40] SW Lane, DT Scadden, and DG Gilliland. The leukemic stem cell niche: current concepts and therapeutic opportunities. *Blood*, 114(6):1150 – 1157, 2009.

[41] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, Dec 2008.

[42] Tsvee Lapidot, Christian Sirard, Josef Vormoor, Barbara Murdoch, Trang Hoang, Julio Caceres-Cortes, Mark Minden, Bruce Paterson, Michael A. Caligiuri, and John E. Dick. A cell initiating human acute myeloid leukaemia after transplantation into scid mice. *Nature*, 367(6464):645–648, Feb 1994.

[43] TJ Ley, C Miller, L Ding, BJ Raphael, AJ Mungall, AG Robertson, K Hoadley, TJ Triche, PW Laird, JD Baty, LL Fulton, R Fulton, SE Heath, J Kalicki-Veizer, C Kandoth, JM Klco, DC Koboldt, KL Kanchi, S Kulkarni, TL Lamprecht, DE Larson, L Lin, C Lu, MD McLellan, JF McMichael, J Payton, H Schmidt, DH Spencer, MH Tomasson, JW Wallis, LD Wartman, MA Watson, J Welch, MC Wendl, A Ally, M Balasundaram, I Birol, Y Butterfield, R Chiu, A Chu, E Chuah, HJ Chun, R Corbett, N Dhalla, R Guin, A He, C Hirst, M Hirst, RA Holt, S Jones, A Karsan, D Lee, HI Li, MA Marra, M Mayo, RA Moore, K Mungall, J Parker, E Pleasance, P Plettner, J Schein, D Stoll, L Swanson, A Tam, N Thiessen, R Varhol, N Wye, YJ Zhao, S Gabriel, G Getz, and So. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, 368(22):2059 – 2074, 2013.

[44] Harri Lähdesmäki, llya Shmulevich, Valerie Dunmire, Olli Yli-Harja, and Zhang Wei. In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, 6:1 – 15, 2005.

[45] W. Li, Y. J. Suh, and J. Zhang. Does logarithm transformation of microarray data affect ranking order of differentially expressed genes? *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, Supplement:6593–6596, Aug 2006.

[46] David A Liebner, Kun Huang, and Jeffrey D Parvin. Mmad: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics*, 30(5):682 – 689, 2014.

[47] CB Lozzio and BB Lozzio. Human chronic myelogenous leukemia cell-line with positive philadelphia chromosome. *Blood*, 45(3):321–334, 1975.

[48] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, James J Collins, Francesca Cordero, Martin Crane, Frank Dondelinger, Mathias Drton, Roberto Esposito, Rina Foygel, Alberto de la Fuente, Jan Gertheiss, and Pierre Geurts. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796 – 804, 2012.

[49] Elaine R. Mardis, Li Ding, David J. Dooling, David E. Larson, Michael D. McLellan, Ken Chen, Daniel C. Koboldt, Robert S. Fulton, Kim D. Delehaunty, Sean D. McGrath, Lucinda A. Fulton, Devin P. Locke, Vincent J. Magrini, Rachel M. Abbott, Tammi L. Vickery, Jerry S. Reed, Jody S. Robinson, Todd Wylie, Scott M. Smith, Lynn Carmichael, James M. Eldred, Christopher C. Harris, Jason Walker, Joshua B. Peck, Feiyu Du, Adam F. Dukes, Gabriel E. Sanderson, Anthony M. Brummett, Eric Clark, Joshua F. McMichael, Rick J. Meyer, Jonathan K. Schindler, Craig S. Pohl, John W. Wallis, Xiaoqi Shi, Ling Lin, Heather Schmidt, Yuzhu Tang, Carrie Haipek, Madeline E. Wiechert, Jolynda V. Ivy, Joelle Kalicki, Glendoria Elliott, Rhonda E. Ries, Jacqueline E. Payton, Peter Westervelt, Michael H. Tomasson, Mark A. Watson, Jack Baty, Sharon Heath, William D. Shannon, Rakesh Nagarajan, Daniel C. Link, Matthew J. Walter, Timothy A. Graubert, John F. DiPersio, Richard K. Wilson, and Timothy J. Ley. Recurring mutations found by sequencing an acute myeloid leukemia genome. *New England Journal of Medicine*, 361(11):1058–1066, 2009.

[50] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(1):S7, Mar 2006.

[51] Florian Markowetz and Rainer Spang. Inferring cellular networks – a review. *BMC Bioinformatics*, 8(6):S5, 2007.

[52] Derek McCulloch, Christina Brown, and Harry Iland. Retinoic acid and arsenic trioxide in the treatment of acute promyelocytic leukemia: current perspectives. *OncoTargets and Therapy*, 10:1585 – 1600, 2017.

[53] Corbin E. Meacham and Sean J. Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, Sep 2013.

[54] Nicolai Meinshausen and Peter Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[55] Krzysztof Mrózek, Guido Marcucci, Peter Paschka, Susan P Whitman, and Clara D Bloomfield. Clinical relevance of mutations and gene-expression changes in adult acute myeloid leukemia with normal cytogenetics: are we ready for a prognostically prioritized molecular classification?. *Blood*, 109(2):431 – 448, 2007.

[56] S. Nam. Databases and tools for constructing signal transduction networks in cancer. *BMB Reports*, 50(1):12–19, 2017.

[57] Dany Nassar and Cédric Blanpain. Cancer stem cells: Basic concepts and therapeutic implications. *Annual Review of Pathology: Mechanisms of Disease*, 11(1):47–76, 2016.

[58] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453 – 457, 2015.

[59] Noa Novershtern, Aravind Subramanian, Lee N Lawton, Raymond H Mak, W Nicholas Haining, Marie E McConkey, Naomi Habib, Nir Yosef, Cindy Y Chang, Tal Shay, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, 144(2):296–309, 2011.

[60] Michael L. OConnor, Dongxi Xiang, Sarah Shigdar, Joanna Macdonald, Yong Li, Tao Wang, Chunwen Pu, Zhidong Wang, Liang Qiao, and Wei Duan. Cancer stem cells: A contentious hypothesis now moving forward. *Cancer Letters*, 344(2):180–187, Mar 2017.

[61] Vicki Plaks, Niwen Kong, and Zena Werb. The cancer stem cell niche: How essential is the niche in regulating stemness of tumor cells? *Cell Stem Cell*, 16(3):225 – 238, 2015.

[62] Olivier B. Poirion, Xun Zhu, Travers Ching, and Lana Garmire. Single-cell transcriptomics bioinformatics and computational challenges. *Frontiers in Genetics*, 7:163, 2016.

[63] Daniel A. Pollyea, Jonathan A. Gutman, Lia Gore, Clayton A. Smith, and Craig T. Jordan. Targeting acute myeloid leukemia stem cells: a review and principles for the development of clinical trials. *Haematologica*, 99(8):1277–1284, Aug 2014.

[64] Marcela Preininger, Dalia Arafat, Jinhee Kim, Artika P. Nath, Youssef Idaghdour, Kenneth L. Brigham, and Greg Gibson. Blood-informative transcripts define nine common axes of peripheral blood gene expression. *PLoS Genetics*, 9(3):1 – 13, 2013.

[65] Andreas Reinisch, Steven M. Chan, Daniel Thomas, and Ravindra Majeti. Biology and clinical relevance of acute myeloid leukemia stem cells. *Seminars in Hematology*, 52(3):150 – 164, 2015.

[66] Dirk Repsilber, Sabine Kern, Anna Telaar, Gerhard Walzl, Gillian F. Black, Joachim Selbig, Shreemanta K. Parida, Stefan HE Kaufmann, and Marc Jacobsen. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, 11(1):27, 2010.

[67] Raghd Rostom, Valentine Svensson, Sarah A. Teichmann, and Gozde Kar. Computational approaches for interpreting scrna-seq data. *FEBS Letters*, 591(15):2213–2225, 2017.

[68] J. Rung, T. Schlitt, A. Brazma, K. Freivalds, and J. Vilo. Building and analysing genome-wide gene disruption networks. *Bioinformatics*, 18(suppl2):S202–S210, 2002.

[69] Yoriko Saito, Naoyuki Uchida, Satoshi Tanaka, Nahoko Suzuki, Mariko Tomizawa-Murasawa, Akiko Sone, Yuho Najima, Shinsuke Takagi, Yuki Aoki, Atsushi Wake, Shuichi Taniguchi, Leonard D. Shultz, and Fumihiko Ishikawa. Induction of cell cycle entry eliminates human leukemia stem cells in a mouse model of aml. *Nat Biotechnol*, 28(3):10.1038/nbt.1607, Mar 2010.

[70] Yang Shen, Yong-Mei Zhu, Xing Fan, Jing-Yi Shi, Qin-Rong Wang, Xiao-Jing Yan, Zhao-Hui Gu, Yan-Yan Wang, Bing Chen, Chun-Lei Jiang, Han Yan, Fei-Fei Chen, Hai-Min Chen, Zhu Chen, Jie Jin, and Sai-Juan Chen. Gene mutation patterns and their prognostic impact in a cohort of 1185 patients with acute myeloid leukemia. *Blood*, 118(20):5593 – 5603, 2011.

[71] Shai S Shen-Orr and Renaud Gaujoux. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, 25(5):571 – 578, 2013.

[72] Shai S Shen-Orr, Robert Tibshirani, and Atul J Butte. Gene expression deconvolution in linear space. *Nature Methods*, 9(1):9, 2012.

[73] Shai S. Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L. Bodian, Frank Staedtler, Nicholas M. Perry, Trevor Hastie, Minnie M. Sarwal, Mark M. Davis, and Atul J. Butte. Cell type-specific gene expression differences in complex tissues. *Nat Meth*, 7(4):287–289, Apr 2010.

[74] Leonard D. Shultz, Fumihiko Ishikawa, and Dale L. Greiner. Humanized mice in translational biomedical research. *Nature Reviews Immunology*, 7(2):118 – 130, 2007.

[75] Socialstyrelsen. Statistikområden, cancer. `http://www.socialstyrelsen.se/statistik/statistikdatabas/cancer`, 2015. Accessed: 2017-18-13.

[76] Gevorg Tamamyan, Tapan Kadia, Farhad Ravandi, Gautam Borthakur, Jorge Cortes, Elias Jabbour, Naval Daver, Maro Ohanian, Hagop Kantarjian, and Marina Konopleva. Frontline treatment of acute myeloid leukemia in adults. *Critical Reviews in Oncology/Hematology*, 110(Supplement C):20 – 34, 2017.

[77] Leo Taslaman and Björn Nilsson. A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PLoS ONE*, 7(11):1–7, 11 2012.

[78] David C. Taussig, Jacques Vargaftig, Farideh Miraki-Moud, Emmanuel Griessinger, Kirsty Sharrock, Tina Luke, Debra Lillington, Heather Oakervee, Jamie Cavenagh, Samir G. Agrawal, T. Andrew Lister, John G. Gribben, and Dominique Bonnet. Leukemia-initiating cells from some acute myeloid leukemia patients with mutated nucleophosmin reside in the cd34- fraction. *Blood*, 115(10):1976–1984, 2010.

[79] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[80] J. E. Till and E. A. McCulloch. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation Research*, 14(2):213–222, 1961.

[81] James E Till and Ernest A McCulloch. Hemopoietic stem cell differentiation. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 605(4):431–459, 1980.

[82] Mark H. Van Benthem and Michael R. Keenan. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *Journal of Chemometrics*, 18(10):441–450, 2004.

[83] James W. Vardiman, Nancy Lee Harris, and Richard D. Brunning. The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood*, 100(7):2292–2302, 2002.

[84] D. Venet, F. Pecasse, C. Maenhaut, and H. Bersini. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17(SUPPL. 1):S279–S287, 2001.

[85] Theo Vos, Christine Allen, Megha Arora, Ryan M Barber, Zulfiqar A Bhutta, Alexandria Brown, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Z Chen, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the global burden of disease study 2015. *The Lancet*, 388(10053):1545–1602, 2016.

[86] Michael L. Wang and Nathanael G. Bailey. Acute myeloid leukemia genetics. *Archives of Pathology & Laboratory Medicine*, 139(10):1215 – 1223, 2015.

[87] Qichao Wang, Xianmin Zhu, Yun Feng, Zhigang Xue, and Guoping Fan. Single-cell genomics: An overview. *Frontiers in Biology*, 8(6):569–576, Dec 2013.

[88] Xuefei Wang, Shile Huang, and Ji-Long Chen. Understanding of leukemic stem cells and their clinical implications. *Molecular Cancer*, 16(1):2, 2017.

[89] John S Welch, Timothy J Ley, Daniel C Link, Christopher?A Miller, David E Larson, Daniel C Koboldt, Lukas D Wartman, Tamara L Lamprecht, Fulu Liu, Jun Xia, Cyriac Kandoth, Robert S Fulton, Michael D McLellan, David J Dooling, John W

Wallis, Ken Chen, Christopher C Harris, Heather K Schmidt, Joelle M Kalicki-Veizer, Charles Lu, Qunyuan Zhang, Ling Lin, Michelle D O'Laughlin, Joshua F McMichael, Kim D Delehaunty, Lucinda A Fulton, Vincent J Magrini, Sean D McGrath, Ryan T Demeter, Tammi L Vickery, Jasreet Hundal, Lisa L Cook, Gary W Swift, Jerry P Reed, Patricia A Alldredge, Todd N Wylie, Jason R Walker, Mark A Watson, Sharon E Heath, William D Shannon, Nobish Varghese, Rakesh Nagarajan, Jacqueline E Payton, Jack D Baty, Shashikant Kulkarni, Jeffery M Klco, Michael H Tomasson, Peter Westervelt, Matthew J Walter, Timothy A Graubert, John F DiPersio, Li Ding, Elaine R Mardis, and Richard K Wilson. The origin and evolution of mutations in acute myeloid leukemia. *Cell*, 150(2):264–278, Jul 2012.

[90] Adriano V. Werhli, Marco Grzegorczyk, and Dirk Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531, 2006.

[91] Daniel Howard Wiseman, Brigit F Greystoke, and Tim CP Somervaille. The variety of leukemic stem cells in myeloid malignancy. *Oncogene*, 33(24):3091–3098, 2014.

[92] Hiromi Yamazaki, Mikiko Suzuki, Akihito Otsuki, Ritsuko Shimizu, Emery H. Bresnick, James Douglas Engel, and Masayuki Yamamoto. A remote gata2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating evi1 expression. *Cancer Cell*, 25(4):415 – 427, 2014.

[93] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763, 2001.

[94] Li Yi and Xie Xiaohui. A mixture model for expression deconvolution from rna-seq in heterogeneous tissues. *BMC Bioinformatics*, 14(Suppl 5):1 – 11, 2013.

[95] Yi Zhong and Zhandong Liu. Gene expression deconvolution in linear space. *Nature Methods*, 9(1):8–9, Jan 2012.

[96] N.S. Zuckerman, P.P. Lee, Y. Noam, and A.J. Goldsmith. A self-directed method for cell-type identification and separation of gene expression microarrays. *PLoS Computational Biology*, 9(8), 2013.