Helsinki University Biomedical Dissertations No. 185

Computational methods to analyze molecular determinants behind phenotypes

Sirkku Karinen

Institute of Biomedicine,
Biochemistry and Developmental Biology &
Research Programs Unit,
Genome-Scale Biology Research Program

Faculty of Medicine

University of Helsinki

Finland

Academic dissertation

To be publicly discussed with the permission of the Faculty of Medicine of the University of Helsinki, in Biomedicum Helsinki 1, Lecture Hall 2, Haartmaninkatu 8, Helsinki, on the 31th of May 2013 at 12 noon.

Helsinki 2013

Supervisor
Sampsa Hautaniemi, DTech
Professor & Academy Research Fellow
Institute of Biomedicine and Genome-Scale Biology Research program
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Pre-examinors
Päivi Onkamo, PhD, docent
University Lecturer
Department of Biological and Environmental Sciences
University of Helsinki
Helsinki, Finland

Garry Wong, PhD
Professor
Department of Neurobiology
A. I. Virtanen Institute
University of Eastern Finland
Kuopio, Finland

Opponent
Yves Moreau, PhD
Professor Department of Electrical Engineering (ESAT-SCD)
Katholieke Universiteit Leuven, Belgium
Leuven-Heverlee, Belgium

Helsinki University Biomedical Dissertations
ISBN 978-952-10-8863-6 (paperback)
ISBN 978-952-10-8864-3 (PDF)
ISSN 1457-8433
Helsinki 2013

To my dearest: Simeon, Sofia, Laura and Pete

Contents

List of original publications iii

Abbreviations iv

Abstract v

Tiivistelmä vi

1 Introduction 1

2 Literature review 2
2.1 Genetic regulation and inheritance of phenotypic variation . . . 2
2.1.1 Phenotypes . . . 2
2.1.2 Inheritance of phenotypes . . . 2
2.1.3 Cell types . . . 4
2.1.4 Cancer cells . . . 6
2.1.5 Biomarkers . . . 7
2.2 Measurement technologies . . . 8
2.2.1 Genetics . . . 8
2.2.2 Gene expression . . . 9
2.3 Computational methods to analyze molecular basis of phenotypes . . . 11
2.3.1 Identification of putative genetic variation behind phenotype . . . 11
2.3.2 Mining of gene expression differences between phenotypes . . . 14
2.3.3 Use of existing biological knowledge . . . 15
2.3.4 Creating predictions from gene expression data . . . 17

3 Aims of the study 21

4 Materials and methods 22
4.1 Data . . . 22
4.1.1 The genome-wide SNP array data . . . 22
4.1.2 The pooled genome-wide gene expression data . . . 23
4.1.3 The RT-qPCR data . . . 24
4.2 Bioinformatics workflow execution . . . 24
4.3 Data integration method for systems biology . . . 24
4.4 Rule-based haplotype analysis from genome-wide data . . . 26
4.5 Pooling heterogeneous studies to compare cell types . . . 26
4.6 Biomarker prediction using machine learning . . . 28

5 Results and discussion 29
5.1 Using information from databases to create novel hypotheses for molecular determinants behind cancer . . . 29

5.2 Computational method to detect shared trait-associated haplotypes derived from a common ancestor 30
5.3 Pooling gene expression datasets reveals novel markers for cell types 32
5.4 Artificial neural network classification predicts phenotypes 33

6 Conclusions and future prospects **35**

7 Acknowledgements **37**

References **39**

Conflicts of interest **50**

List of original publications

PUBLICATION I	Karinen S., Heikkinen T., Nevanlinna H. and Hautaniemi S. (2011) Data integration workflow for search of disease driving genes and genetic variants. *PLoS ONE 6*, e18636
PUBLICATION II	Karinen S., Saarinen S., Lehtonen R., Rastas P., Vahteristo P., Aaltonen L. A. and Hautaniemi S. (2012) Rule-based induction method for haplotype comparison and identification of candidate disease loci. *Genome Medicine 4*, 1–18
PUBLICATION III	Keuschnigg, J., Karinen, S., Auvinen, K., Irjala, H., Mpindi, J. P., Kallioniemi, O., Hautaniemi, S., Jalkanen, S. and Salmi, M. (2013) Plasticity of blood- and lymphatic endothelial cells and marker identification. *Submitted*
PUBLICATION IV	Savilahti E. M., Karinen S., Salo H. M., Klemetti P., Saarinen K. M., Klemola T., Kuitunen M., Hautaniemi S., Savilahti E. and Vaarala O. (2010) Combined T regulatory cell and Th2 expression profile identifies children with cow's milk allergy. *Clinical Immunology 136*, 16–20

PUBLICATION IV has been used in Emma Savilahti's PhD thesis (2010) Cow's milk allergy and the development of tolerance.

Abbreviations

ANN	artificial neural network
AUC	area under curve
BEC	blood endothelial cell
cDNA	complementary DNA
cHL	classical Hodgkin lymphoma
CMA	cow's milk allergy
DNA	deoxyribonucleic acid
FN	false negative
FP	false positive
GBM	glioblastoma multiforme
GEO	gene expression omnibus
GWAS	genome-wide association study
HMEC	human microvascular endothelial cell line
IBD	identity-by-descent
IBS	identity-by-state
LD	linkage disequilibrium
LEC	lymphatic endothelial cell
MLP	multilayer perceptron neural network
MM	mismatch
NHL	non-Hodgkin lymphoma
NLPHL	nodular lymphocyte predominant Hodgkin lymphoma
mRNA	messenger RNA
PM	perfect match
PPI	protein-protein interaction
RNA	ribonucleic acid
RNA-seq	RNA sequencing
ROC	receiver operator characteristics
RT-qPCR	reverse-transcription quantitative polymerase chain reaction
SVM	support vector machine
SH	shared haplotype
SNP	single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
TCRBCL	T-cell/histiocyte rich B-cell lymphoma
TIME	telomerase immortalized microvascular endothelial cell line
TN	true negative
TP	true positive

Abstract

Phenotype is a collection of an organism's observable features that can be characterized both on individual level and on single cell level. Phenotypes are largely determined by their molecular processes which also explain their inheritance and plasticity. Some of the molecular background of phenotypes can be characterized by inherited genetic variations and alterations in gene expression.

The high-throughput measurement technologies enable the measurement of molecular determinants in cells. However, measurement technologies produce remarkable large data sets and the research questions have become increasingly complex. Thus computational methods are needed to discover molecular mechanisms behind the phenotypes. In many cases, analysis of molecular determinants that contribute to the phenotype proceeds by first identifying putative candidates by using *a priori* information and high-throughput measurements. Then further analysis can focus on most promising molecules. In many cases, the aim is to identify relevant markers or targets from a set of candidate molecules.

Often biomedical studies result in a long list of candidate genes, and to interpret these candidates, information on their context in cell molecular functions is needed. This context information can give insight to synergistic effects of molecular machinery in cells when functions of individual molecules do not explain the observed phenotype. In addition, the context information can be used to generate candidates. One of the methods in this thesis provides a computational data integration method that provides a link in between candidate genes from molecular pathways and genetic variants. It uses publicly available biological knowledge bases to systematically create functional context of candidate genes. This approach is especially important when studying cancer, that is dependent of complex molecular signaling.

Genotypes associated with inherited disease predispositions have been studied successfully in the past, however, traditional methods are not applicable in wide variety of analysis conditions. Thus, this thesis introduces a method that uses haplotype sharing to identify genetic loci inherited by multiple distantly related individuals. It is flexible and can be used in various settings, also with very limited number of samples.

Increasing the number of biological replicates in gene expression analysis increases the reliability of the results. In many cases, however, the number of samples is limited. Therefore, pooling gene expression data from multiple published studies can increase the understanding of the molecular background behind cell types. This is shown in this thesis by an analysis that identifies gene expression differences in two cell types using publicly available gene expression samples from previous studies.

Finally, when candidate molecules are available to characterize phenotypes, they can be compiled into biomarkers. In many cases, a combination of multiple molecules serves as a better biomarker than a single molecule. This thesis also includes a machine learning approach that is used to discover a classifier that predicts the phenotype.

Tiivistelmä

Fenotyyppi on joukko organismin piirteitä, jotka ovat havaittavissa joko yksilön tasolla tai yksittäisten solujen tasolla. Molekulaariset prosessit määräävät pitkälti fenotyyppien ilmentymistä, joten taustalla vaikuttavat molekulaariset prosessit myös selittävät fenotyyppien perinnöllisyyttä sekä mukautumista. Fenotyyppien molekulaarista taustaa voidaan kartoittaa tunnistamalla geneettistä variaatiota sekä muutoksia geenien aktiivisuudessa.

Määräväviä molekulaarisia tekijöitä voidaan havaita soluissa käyttämällä *high-throughput* -mittausteknologioita. Nämä mittausteknologiat tuottavat erittäin suuria data-aineistoja, ja samalla tutkimuskysymykset ovat tulleet entistä monimutkaisemmiksi. Nämä seikat ovat johtaneet siihen, että laskennallisia menetelmiä tarvitaan fenotyyppien molekulaarisen mekanismien tunnistamisessa. Usein tutkimus etenee ensin tunnistamalla lupaavia kandidaatteja käyttämällä aiempaa tietoa sekä *high-throughput* -mittauksia. Jatkoanalyysit voivat keskittyä lupaavimpiin molekyyleihin. Tällöin tavoitteena saattaa olla käyttökelpoisimpien biomarkkereiden tunnistaminen tai kohdegeenien valitseminen kandidaattien joukosta.

Usein biolääketieteen tutkimus tuottaa ison joukon kandidaattigeenejä, jolloin tulosten tulkinta vaatii tietoa kandidaattigeenien suhteesta solun muuhun molekulaariseen toimintaan. Kun tämä molekulaarinen toiminta kontekstina otetaan huomioon, on mahdollista ymmärtää geenien yhteisvaikutuksia solun toimintaan silloin kun yksittäiset geenit eivät selitä havaittua fenotyyppiä. Solun molekulaarista kontekstia voi käyttää myös kandidaattigeenien kartoittamiseen. Yksi tässä väitöskirjassa esitelty menetelmä tarjoaa laskennallisen menetelmän, jolla voidaan yhdistää kandidaatit tunnetuilta pathwaylta geneettisiin variantteihin. Tämä menetelmä käyttää julkisia tietokantoja, joista se systemaattisesti kerää molekulaarisen kontekstin kandidaattigeeneille. Tällainen lähestymistapa on erityisen hyödyllinen syöpätutkimuksessa, sillä syöpä on tyypillisesti riippuvainen monimutkaisista molekyylien signalointiverkoista.

Perittyjen genotyyppien ja sairauksien välisiä yhteyksiä on tutkittu pitkään menestyksekkäästi, mutta perinteisesti käytetyt menetelmät soveltuvat vain tiettyihin tapauksiin. Tässä väitöskirjassa esitellään menetelmä, joka käyttää haplotyyppien jakamista tunnistaakseen genomiset alueet, jotka ovat periytyneet useille kaukaisesti sukua oleville henkilöille. Tätä menetelmää voi käyttää useissa erilaisissa tutkimuskysymyksissä, ja se tuottaa luotettavia tuloksia myös hyvin vähäisellä näytemäärällä.

Geenien ekspressioanalyysin tulosten luotettavuus kasvaa samalla kun biologisten kopioiden määrä aineistossa kasvaa. Huolimatta tästä, näytemäärät ovat usein rajallisia. Tämän vuoksi geenien ekspressiomittausten yhdistäminen useista jo julkaistuista tutkimuksista voi lisätä ymmärrystä solutyypin määräävistä biologisista prosesseista. Tässä väitöskirjassa esitellään analyysi, jolla tunnistetaan geenien ekspressioeroja käyttäen geenien ekspressioainestoa, joka on yhdistetty julkaistuista tutkimuksista.

Viimein, kun fenotyyppiä selittävät kandidaattimolekyylit on tunnistettu, niistä voidaan luoda biomarkkereita. Monesti useamman molekyylin mittaus on parempi biomarkkeri kuin yksikään molekyyli yksinään. Tässä väitöskirjassa esitellään myös koneoppimisanalyysi, jolla luodaan geenien ekspressiomittauksista fenotyyppiä ennustava luokittelija.

1 Introduction

Computational methods are essential in biomedical research when studying molecular mechanisms that determine, for instance, what creates organisms' appearance, makes them susceptible to a disease, or induces them to behave as they do. The computational methods can be used to quantify differences in molecule concentrations in different types of cells. As measurement technologies evolve and our understanding of biology increases, computational methods need to be developed to meet new demands.

The topic of this work originates from traditional paradigm in medicine, or natural sciences in general, where understanding of a natural phenomenon can be solved by dissecting the question into the study of individual components [1]. This paradigm in science, *reductionism*, is the principle when studying molecular mechanisms that determine a phenotype, which is an organism's observable feature, such as eye color. Reductionistic paradigm assumes one-to-one mapping between a molecule, such as gene product, and the phenotype [2]. This approach is evident in medicine, where the research aims to identify the dysfunctional organ or molecule, and return the normal, healthy, state by repairing the problem [1]. Similarly, the estimators for disease risks are typically based on single molecule biomarkers and each disease is typically treated individually without a holistic view of a patient's physical state [1].

One limitation of reductionism is that it misses the interactions and multiplicative effects of the components in a complex biological system [1]. This has pushed biomedical research to study complete biological functions at the systems level. Systems biology assumes that the whole system has synergistic characteristics that cannot be seen from individual components [1, 2]. In this paradigm, computational approaches are in a central role as systems biology includes computational models and biological measurements [1, 3]. The carefully designed measurements verify the accuracy of the models, although molecular biology is still lacking the rich datasets that, for example, chemistry has [3].

Although reductionism has limitations due to overly simplified assumptions, it has also shown its power in biomedicine [1]. However, increasingly complex datasets and computational approaches have been produced to address the questions of biomedicine at systems-level [4, 5]. Thus, no single approach can explain the complex nature of phenotypes, which calls for approaches that target the research question on different levels.

Biomedical research requires understanding of the system level functions of the molecules, inheritance of phenotypes, the phenotypic plasticity in tissue types, and, in the end, it also requires translating these into clinical tools such as biomarkers. In Publication I we present how the molecular determinants can be estimated using *a priori* knowledge from molecular systems. Publication II and Publication III present two complementary approaches to identify candidate genes behind the phenotype. In Publication II we use genetics in a setting where the phenotype is inherited from a common ancestor. In Publication III we analyze gene expression and combine biologically and technically versatile data, and identify candidate genes where changes in expression change the phenotype of the cell types. To use these candidate genes clinically, they can be compiled into biomarkers or their synergistic effects can be studied computationally. In Publication IV we identify a gene expression signature that predicts the phenotype.

2 Literature review

The first aim of this chapter is to define the biological goals and review the molecular mechanisms known to affect phenotypes. The second aim is to address the measurement technologies that produce information from the molecular level. Finally, the relevant computational analysis approaches are reviewed.

2.1 Genetic regulation and inheritance of phenotypic variation

The concepts of phenotypes and their inheritance can be viewed both from individual-level (organism) and from sub-individual level (cell types or cell colonies) [6]. The inheritance of phenotypes on both of these levels are explained by genetic inheritance (DNA) [6] and epigenetic inheritance (regulation of DNA and chromatin) [7]. The genetic and epigenetic inheritance together contribute to a cell's molecular functions that can be observed as cell type specific gene expression [8].

2.1.1 Phenotypes

On an organismal level, a phenotype is the organism's whole collection of observable physical components; the component may be morphology, behavior, or molecular function [9, 10]. In many cases, like in this thesis, phenotype actually refers to a partial phenotype [9], that consists only of those components that are interesting or relevant in given context, *i.e.,* disease phenotype in comparison to disease-free phenotype. Thus, understanding the biology behind a partial phenotype is the guiding principle in this thesis.

Phenotype can be something as obvious as eye color. Organisms can also be classified according to the phenotype that is common to them [9], such as walking in an upright position or the ability to produce language. Species-specific phenotypes are examples of such groupings [9]. On cell type level, the phenotype can be seen as species-specific phenotype for cell types, that is the collection of physical components that is common to all cells in that cell type.

Phenotype is determined by the inherited material (which is reviewed in the next section), and by the environment [10]. When including the environmental factors into phenotype determinants, phenotypes tend to be plastic both on organismal level [11] and on single cell or cell population level [12]. In this thesis molecular regulation of a cell is considered as an inherited determinant of a phenotype and only the external signals to be environmental factors.

2.1.2 Inheritance of phenotypes

The phenotype passed on from one generation of biological entities, such as organisms or single cells, to another follows the *like-begets-like* phenomenon [6], where the offspring usually resemble the parent. Differences on individual and sub-individual level may be inherited in genetic (in DNA) or epigenetic (regulation of DNA and chromatin) manner. According to, the nowadays disputed [13, 14], *'central dogma'* of biology [15], all inherited information is stored in the DNA molecule, from which the RNA molecule is transcribed and this messenger RNA (mRNA) is later translated into proteins that perform the functions of the cells. This process is illustrated in Figure 2.1. According to this dogma, genetic inheritance is explained by transmission of DNA from parents to offspring [6], and features of the DNA molecule translate to phenotype through messenger RNA (mRNA), protein, and

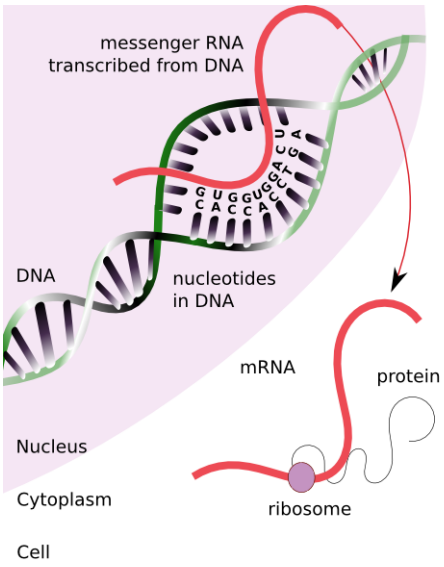

Figure 2.1: The schematic representation of the process that conveys the information from DNA to the phenotype, which is also referred as *'central dogma'* of biology. DNA double helix is opened, and the messenger RNA (mRNA) is transcribed from the DNA template. The protein is synthesized from the mRNA molecule in a cell's cytoplasm.

protein functions or through functions of non-coding RNA [16, 17]. Meanwhile, epigenetic inheritance functions without changes in the DNA sequence through epigenetic modifications and regulation of cell signaling [18, 7], but also through the maintenance and transmission of DNA [18].

Essentially, the inherited genetic or epigenetic features are manifested in the phenotype through molecular signaling: the type of produced molecules, the rate of molecule production or success of molecular interactions. At a single cell level, the inherited part of a phenotype is determined by the information content of DNA [19] and the regulation of intrinsic molecular signaling [20].

The central part of inheritance, both on individual and sub-individual level, is DNA transmission from parent to offspring [6]. The biological basis of this DNA transmission is well characterized in sexual reproduction, in meiosis, and in eukaryote cell division, in mitosis. Procaryotes and viruses also transmit DNA (or RNA) to their offspring, but this thesis is concentrated only on mechanisms in eukaryote cells.

In sexual reproduction, both parents transmit haploid chromosomes to the offspring that receives diploid chromosomes combined of the parents' genomes. The parents' haploid chromosomes are produced in meiotic cell division, which has two cell divisions but only one DNA replication [19]. Interesting parts of meiosis are the dissociation of parental chromosomes and the recombination, where homologous parental chromosomes exchange parts of the DNA chain. This, in principle, produces a random selection from paternal and maternal genomic sequences transmitted to the offspring. Mitosis, instead, produces two diploid daughter cells that have nearly identical copies of the chromosomes.

Epigenetic inheritance takes place on the single cell (sub-individual) level [7], it follows DNA transmission in cell division, although the inherited information is not in the DNA sequence [7].

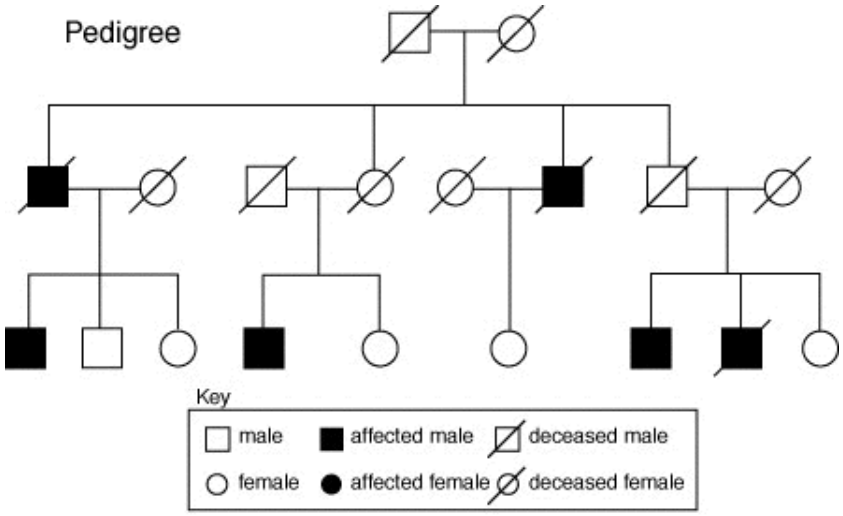

Figure 2.2: An example of a pedigree diagram. The inheritance of phenotypes can be visualized using a pedigree diagram as shown by this figure. The squares are males and the circles females, and individuals are connected to their spouses with direct horizontal lines. The children are below their parents and connected to them with vertical lines. A slash through a symbol means that this individual is deceased. [Figure downloaded 21.1.2013 from http://en.wikipedia.org/wiki/File:Pedigree-chart-example.png]

Epigenetic information is based on methylation of DNA and several different histone modifications [7, 18] that have an effect on DNA transcription, which propagates to changes in phenotype [7, 18]. Although it is important to recognize the effect of epigenetics to molecular functions and phenotypes, it is out of the scope of this thesis.

The inheritance of phenotypes is typically visualized with a pedigree chart as in Figure 2.2. Various studies have revealed a causative genetic component in the inheritance of many human phenotypes such as eye color [21], neurofibramatosis [11] or colorectal cancer [22]. However, transmission of DNA alone does not explain the inheritance of phenotypes [6, 11]. Although non-genetic part of inheritance is often explained by environmental factors[6], also epigenetic inheritance may be an important factor that can contribute to inheritance of phenotypes.

2.1.3 Cell types

All cells in an organism inherit their DNA from a common progenitor. Although some genetic changes accumulate during cell generations, cell type specific phenotypes are typically determined by regulation of gene expression [20, 23]. An example of morphological differences between cell types is shown in Figure 2.3.

During the development of an organism, the phenotype is altered by the environmental signals that trigger transcription factors that either promote or suppress the target genes' expression by binding to DNA [24, 20]. For instance, the gene *POU domain, class 5, transcription factor 1* (*Oct-3/4*) has been identified as a switch whose expression regulates differentiation from stem cells [25]. During an

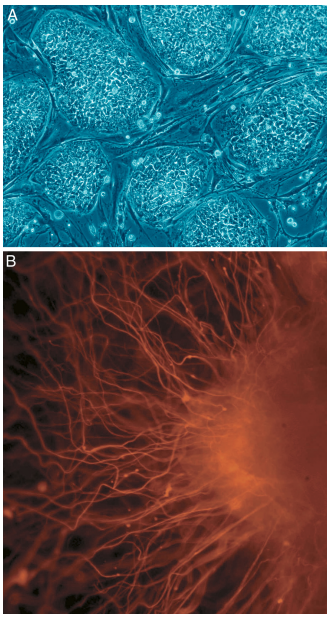

Figure 2.3: A) Undifferentiated human embryonic stem cells and B) neurons developed from the human embryonic stem cells. All the cells carry nearly identical DNA and still show dramatically different phenotypes. The cell types differentiate during development and commit to this phenotype. This image is published in PLoS Biology [26] under Commons Attribution 2.5 license (link to the file 19.11.2012 http://en.wikipedia.org/wiki/File:Human_embryonic_stem_cells.png)

organism's development, cells differentiate and cell populations maintain their differentiation [24, 23]. In biomedicine, the cell type specific molecular signatures provide attractive biomarkers and may serve as selective drug targets [8]. In addition to environmental signals for transcription factors, the sub-individual inheritance of phenotype from cell generation to another is partly due to epigenetic regulation of cell fate [24]. Thus, the gene expression regulation is an interplay between epigenetic modifications and transcription factor responses to signals.

The cell type specific phenotype is a result of complex molecular signaling, which can be described in terms of molecular signaling pathways. As gene expression produces the interacting molecules in the molecular signaling pathways, the signaling in these pathways may change when gene expression changes, as for example in the apoptosis pathway [27]. One example of transcription factors is the widely studied *tumor protein 53* (tp53) that responses to various stress signals [28, 29]. Once a cell confronts stress, such as double-stranded DNA breakage, tp53 activation in the cell increases and its ability to upregulate target genes' expression increases [28, 29]. The result may be an activation of an apoptosis pathway, whose schematic diagram from WikiPathways [30] is given in Figure 2.4. Although the molecular signaling pathways usually involve gene products instead of DNA, for sake of simplicity, we can assume one-to-one mapping of genes to their protein products and call proteins in these pathways with corresponding gene names and focus on studies of gene expression.

Several studies have identified tissue type specific gene expression signatures [31, 32, 8] that show commitment of cell types to to express their specific phenotypes. These experiments have accumulated

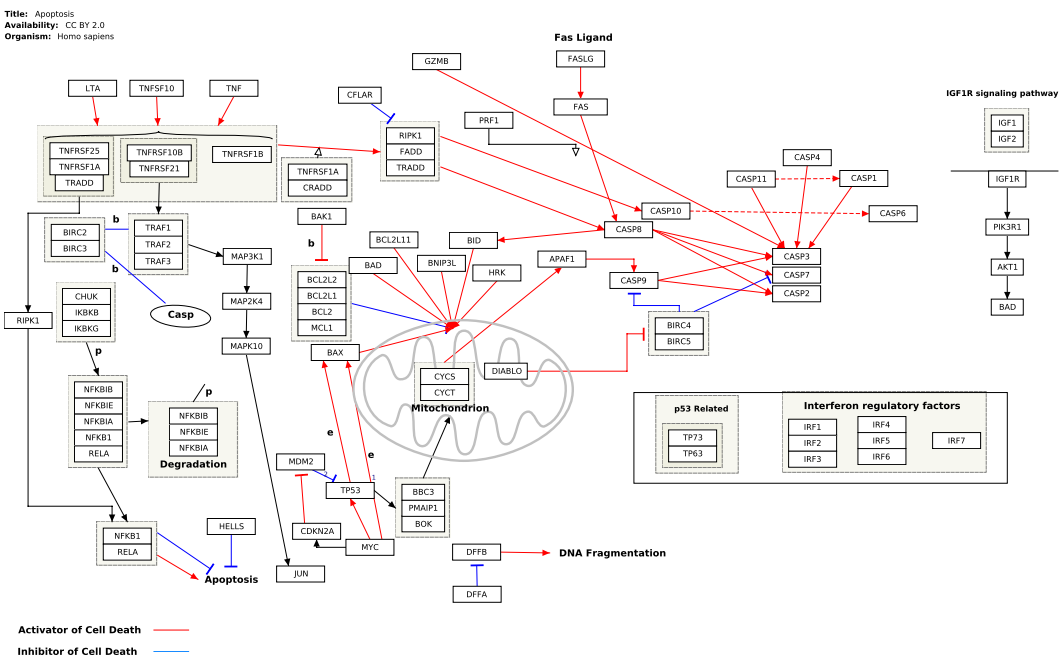

Figure 2.4: The molecule signaling diagram from the apoptosis pathway. This diagram represents a model for gene activation and repression. The squares are proteins of which we can assume a one-to-one mapping to genes, and the lines are interactions between these proteins. Proteins in the same boxes form complexes. Blocked lines illustrate repression and arrows activation. In this diagram, the red lines indicate the activation signals of apoptosis and blue lines repression of apoptosis. Image was downloaded from WikiPathways at 27.9.2012.

data to phenotype databases that define gene expression patterns in tissue or cell types [10, 33, 34].

2.1.4 Cancer cells

Cancer cells represent a subtype of cell types that have acquired a phenotype that allows uncontrollable growth of the cell population and invasion into surrounding tissues, which may create a life threatening disease by disturbing the normal functions of the body [35]. Although cancers are a group of heterogeneous malignant growths from different tissue origins, similar molecular functionalities can be seen in all cancers [36].

The established hallmarks of cancer cells are proliferation, escape from growth restriction, invasion to other tissues, unlimited number of cell cycles, ability to introduce angiogenesis and escape from apoptosis [36, 37]. In contrast to normal cell populations, cancer cells' phenotype is due to accumulated somatic mutations [37, 38], especially through genomic instability [36, 37], and additional epigenetic modifications during cell generations [37, 39]. Figure 2.5 visualizes the steps that a cancer cell population has undergone during the disease development. Cancer cells grow in a colony where the cell population is under massive evolutionary selection [37, 38], which results bias to some beneficial genetic and epigenetic changes.

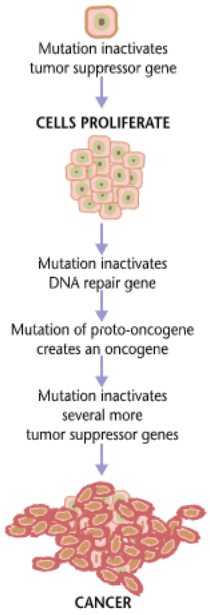

Figure 2.5: Cancer cells acquire somatic mutations which lead to an altered phenotype, which can create a disease by uncontrolled growth of cancer cells and their invasion to other tissues. This figure is published by the National Cancer Institute under public domain. (Link to file 19.11.2012, http://en.wikipedia.org/wiki/File: Cancer_requires_multiple_mutations_from_NIHen.png)

Cancer originates from normal cells that, along cell generations, undergo several changes before they develop into a life-threatening disease [37]. In some of the cases, the initial chromosomal changes are inherited in the germline, which manifests as an increased risk of cancer [22, 40]. Some of the inherited genetic risk is associated with common genetic variants, which can be observed at the population level [40]. However, one genetic or epigenetic change alone is not sufficient to create cancer [37], of this colorectal cancer is a good example. Colorectal cancers have, in many cases, concurrent *tp53* inactivation with other mutations in genes such as *liver kinase B1* (*LKB1*) or *adenomatous polyposis coli* (*APC*) [22]. Cancer cells are typically characterized by altered molecular signaling pathways. For example, classical cancer pathways are those regulated by *tp53*, thus *tp53* is deregulated either directly or indirectly in many different types of cancers [41], such as breast cancer [4].

2.1.5 Biomarkers

A biomarker is a measurable feature that can be used to detect biological processes and success of treatments [42]. Additionally, the definition of molecular biomarker has been extended to all possible biomarkers and biomarker combinations that can be detected using their molecular features [43]. The classification of the biomarkers varies depending on the context where the biomarker is intended to be used: is it targeted, for instance, to identify cell types or assess organism-level phenotypes, such as diseases [44]. The biomarkers can be based on the molecular background of the phenotype or some non-molecular feature, such as blood pressure [42]. Also, the biomarker may be the cause of the phenotype or it can be only a surrogate marker [45].

Biomarkers are widely used in disease screening for early-onset diseases [44, 46], diagnostics [46], prognosis [47, 48, 49, 46] and selecting treatments [49, 46]. Breast cancer, for example, has an established classification system that is based on biomarkers regarding tumor size, metastasis and lymph nodes [50], and additional molecular biomarkers identified from gene expression [51, 48]. The biomarkers can be used as surrogates to predict the disease outcome [42, 44], which is especially useful in drug trials [42]. In addition, molecular markers can be used to identify and monitor tissue or cell types [52, 53]. One example of a widely studied molecular biomarker is *Oct-3/4,* which has been used to identify undifferentiated stem cells [54, 53].

2.2 Measurement technologies

DNA is the building block of inheritance, and gene expression is the well characterized molecular basis of cell type specific phenotype in multicellular organisms. Therefore, the measurement technologies reviewed are those to discover the identity of DNA (genetic variation), and those to quantify the gene expression intensity (amount of mRNA).

2.2.1 Genetics

Genetic measurement technologies are based on the genetic variation profile of an individual. In each locus, each individual has two possible alleles of base pairs that constitute a genotype. Genotyping technologies aim to characterize the genotypes along the genome.

SNP microarrays Single nucleotide polymorphism (SNP) is a variant that is polymorphic in a given population, *i.e.,* it has a minor allele frequency, which is the allele frequency of rarer allele in a population, higher than some predefined threshold, such as 5% [55]. Biallelic (two possible alleles in a population) SNPs are used as a genotyping surrogates in SNP microarrays [56, 55, 57]. One individual can be distinguished from others by measuring a sufficient number of SNPs. In meiosis, parental chromosomes can be recombined to form a new chromosome for offspring that has DNA stretches from both parental chromosomes. In a population this leads to a phenomenon called linkage disequilibrium (LD), where DNA regions tend to be observed in blocks, which is usually measured by correlation between loci [55, 57]. The concept of using SNPs in genotyping is based on an assumption that a SNP tags the DNA region that is in LD with the SNP. Thus, by measuring these SNPs we actually indirectly measure the identity of the surrounding DNA region [55].

A SNP microarray contains single-stranded complementary DNA (cDNA) probes for biallelic SNPs that have been selected to the array [56, 57]. The principle of SNP microarray technology is illustrated in Figure 2.6. The probes are organized into probe sets having one probe set for allele A and one for allele B [56]. The single-stranded and fragmented sample DNA is hybridized on the probes in the array and the hybridized probes then contain label, usually biotin, to capture the signal [56, 57]. Next the array is scanned and the image is read for the genotypes [56]. The signal is measured from probe sets for alleles A and B, and the loci is determined either to be homozygous AA or BB genotype or heterozygous AB genotype, and the SNP locus is mapped to the probe location in the array [56, 57].

Sequencing A direct way to detect genotypes from any species is sequencing, and the most recent sequencing technologies are rapidly growing in popularity over the SNP microarray genotyping. The

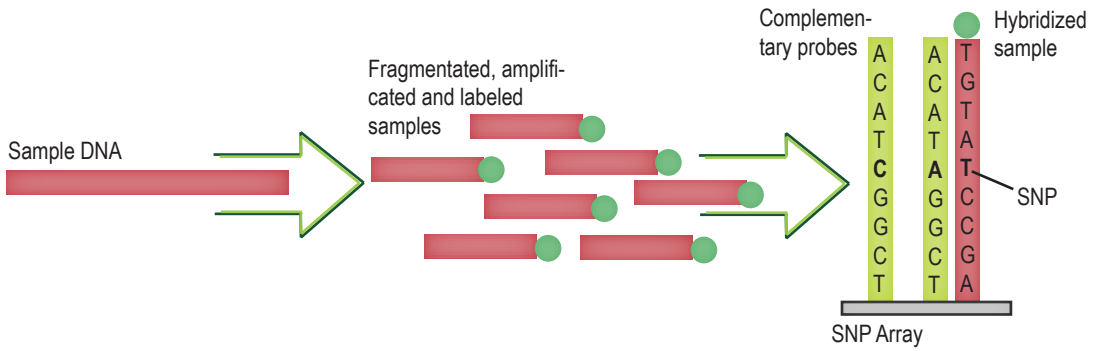

Figure 2.6: Schematic illustration of the SNP array technology. Original figure is in Karinen, S. 2008.

benchmark technique is Sanger sequencing, which is generally feasible in sequencing limited chromosomal regions [58, 59]. The Sanger sequencing uses a DNA template where DNA is generated by polymerase chain reaction, which is terminated by a dideoxyribonucleotide [59]. The sequence length is determined by each of the four sequence terminating nucleotides, which enables deduction of the order of nucleotide sequence [59]. The Sanger method is today used merely for validation purposes [58].

Currently, the newest sequencing technologies enable sequencing of the whole genome or selected regions, such as protein coding regions (exome sequencing) [58]. The sequencing technologies are developing rapidly [58], but the standard scheme in various platforms is to generate base pair calls for a read, which is a string of DNA from one hundred to more than thousand nucleotides long [60]. These reads overlap, which creates a read coverage along the genome [61]. The actual sequence is inferred from these overlapping reads.

Measurements produced by sequencing were not used in the publications in this thesis, however, recognizing the importance sequencing technologies is essential for modern biomedical research. Also, the methods presented in this thesis to analyze genetic variation can be applied to data derived from sequencing technologies, as well.

2.2.2 Gene expression

Gene expression measurement quantifies the level of mRNA transcribed from genes [62]. Gene expression can be measured either from selected genes of interest or genome-wide from all the genes. The selected technology depends on the target: expression of a small set of genes can be measured with reverse-transcription quantitative polymerase chain reaction (RT-qPCR) [62], the whole-genome expression profile can be captured with gene expression microarray technologies [63], and the most comprehensive whole-genome expression quantification can be achieved with RNA sequencing (RNA-seq) [64].

RT-qPCR Reverse-transcription quantitative polymerase chain reaction (RT-qPCR) can be used to quantify gene expression of selected genes from up to a thousand samples, and it is especially useful

in biomarker deployment [62]. RT-qPCR is highly sensitive, economical [65], and it can produce measurements from a very limited amount of RNA.

In RT-qPCR, RNA is converted into cDNA using reverse-transcription, which is then amplified by using PCR for a pre-determined number of cycles and the relative concentration is then quantified [62]. Real-time qPCR allows simultaneous amplification and detection of the RNA [62]. The detection is based on fluorescence dyes that are attached to reporter sequences, that in turn bind or hybridize on the target molecule [62]. The amount of cDNA doubles at each PCR cycle, therefore the expression values require specific transformation steps. In the analysis, RT-qPCR data can be used as absolute values or relative expression values [62]. For the relative gene expression quantification, a housekeeping gene is used as an internal control to subtract the expression value from pre-defined PCR cycle threshold (Ct) [65] from target genes' value in Ct to get the gene-specific value $\triangle Ct$ [62, 66]. An outside calibrator sample can also be used to get the normalized value $\triangle\triangle Ct$ between assays that is usually reported as $2^{-\triangle\triangle Ct}$ [66, 62].

Gene expression microarrays Gene expression microarrays follow the same principle as SNP genotyping microarrays. They have cDNA probes attached on an array where target DNA is hybridized and the fluorescence signal is scanned by special fluorescence microscopy from the microarray, and the target is then identified using probe position coordinates in the microarray [67, 68]. In the gene expression microarrays, however, the target DNA is produced using reverse-transcription from mRNA [67, 69]. The probes selected for the microarray measure the expression of the target gene. Several commercial providers offer gene expression microarrays having different technical solutions, such as one-channel or two-channel signal [70].

The one-channel Affymetrix platform has perfect match (PM) probe sets that consist of multiple perfectly matching 25 nucleotides long oligonucleotides and corresponding mismatch (MM) oligos with one nucleotide difference in the middle of the probe [68]. Labeled cDNA from a single sample is hybridized on these probes. These probes can be mapped to the most up-to-date annotations to obtain mRNA measurements from correct chromosomal loci, and the gene-specific expression can be combined from the probes that are mapped to the same gene [68]. The two-channel microarrays provided by Agilent can hybridize two samples having different labels [70]. One subtype of gene expression microarrays is an exon array, whose high density probes aim to capture expression of individual exons. This measurement technology enables to quantify expression of alternatively spliced transcripts [71].

The raw values from the gene expression microarray are rarely used, but a normalization step follows the reading of the data to account for background and inter-array differences [63]. The normalization method is selected according to the used measurement technology, and it affects the data quality [72], however, there are no definite guidelines on how to select the optimal normalization method [73].

RNA-seq RNA-seq is currently the most advanced gene expression quantification method as it allows measuring the level of all RNA in the cell as well as determining specific transcripts [62, 64]. However, RNA-seq is a relatively new technology, especially when compared to qPCR, whose development was initiated from discovery of PCR in 1983 [62]. In comparison to DNA sequencing, the RNA-seq technology has additional complexity because it aims to quantify the gene expression

instead of just giving distinct judgment of the presence of a variant [74]. The RNA-seq also starts from reverse-transcription that transforms the target RNA into cDNA, which is then sequenced to obtain data of overlapping reads [74]. The RNA level is then quantified from these overlapping reads.

Data produced by RNA-seq were not used in publications of this thesis. However, use of the RNA-seq measurement technology is increasing, and computational method in PUBLICATION IV can also be applied to RNA-seq data.

2.3 Computational methods to analyze molecular basis of phenotypes

The human genome consists of approximately 3 billion bases and 20,000 protein coding genes, where the regulation of molecular mechanisms is complex [75]. Thus, it is evident that computational methods are required to understand the DNA-originated molecular mechanisms behind phenotypes. The computational methods reviewed in this section either aim to identify molecular candidates behind the phenotype or compile these candidates into a clinically relevant predictive model. This separation of the computational methods follows an approach in biomedicine where the understanding of complete processes is initiated by identifying a set of candidate molecules that is then analyzed to prioritize the findings [76] or integrated into more complex functional models [77].

The mining of molecular determinants can be based on genotype or gene expression data. The mining is typically dependent on the existing biological knowledge, and this knowledge is used as a reference to interpret results, to rank the findings, or to create hypotheses for an analysis [76]. One example of existing biological knowledge is the human genome that is used as reference basis for research in biomedicine [75].

A clinically relevant model can be a diagnostic or prognostic predictor. Methods to discover these predictors typically use the candidate molecules produced by the mining and create an application that selects biomarkers for a model that predicts the phenotype from the data [78]. Although the aim of these methods is prediction, the study of biomarkers can also give an insight to molecular processes behind the phenotype [53].

2.3.1 Identification of putative genetic variation behind phenotype

Traditionally, medical genetics has aimed to identify chromosomal regions that are associated wiht inheritance of a phenotype in families [79]. The goal has usually been to target very rare and striking phenotypes, such as Mendelian diseases, that presumably are caused by rare, high penetrance genetic variants, *i.e.*, variants that cause a disease in a high proportion of carriers [80].

Another statistical approach is a large population study, in which an association between genetic variants and phenotype is sought for. These studies use the *common variant – common disease* model, which assumes common, low-penetrance variants to cause subtle phenotypes, such as cancer predisposition [81, 82]. Since many phenotypes do not fall directly to these categories [82], a plethora of other computational methods, which are reviewed later in this section, target phenotypes in between these inheritance models.

Statistical linkage and association analysis The linkage analysis tests whether a genetic locus and an inherited phenotype are segregating together in a pedigree [83]. Figure 2.7 shows the genetic basis of this analysis where the closest measured marker correlates the most with the causative allele

[80]. A statistical linkage analysis uses families to identify high penetrance genetic variants [84]. This analysis requires estimations of inheritance of alleles, penetrance and allele frequencies [79]. The linkage analysis can include either affected sibling pairs or families with multiple affected individuals [84, 80]. The linkage analysis requires the genotypes in the family to be informative in or near to the causative variant locus, *i.e.,* to show heterozygosity in the pedigree to distinguish the maternally and paternally inherited alleles [84]. The linkage analysis is highly powerful in locating a genetic locus carrying the causative allele. However, the use of the linkage analysis is restricted to limited research settings where information on pedigrees, number of individuals and families, inheritance model, and allele frequencies is available. In practice, all this information is only rarely known.

A genome-wide association study (GWAS) for *common variant – common disease* model tests statistical association with SNPs' allele frequencies and a phenotype, such as a disease, from thousands of samples [81, 57]. When this test is done genome-wide, it results in hundreds of thousands of tests, which requires a very large set of samples for enough statistical power [81, 57]. The phenotype may be discrete or continuous, which affects selection of the statistical test [81]. However, an association test may find a false positive variant because of a confounding effect such as population stratification, which is also emphasized when the number of samples increases [84, 81].

Identity-by-descent analysis A phenotype is not necessarily caused by a common genetic variant but by multiple rare variants, which requires analyzing shared ancestry [81]. Identical chromosomal regions between individuals may be inherited from a relatively recent common ancestor [85, 86] or they may emerge due to a random event, these regions are called identity-by-descent (IBD) or identity-by-state (IBS), respectively. Figure 2.7 illustrates how chromosomal segments are broken in a generation. The causative allele is usually surrounded by a longer stretch of DNA, which is the IBD region between individuals.

The rare causative variants can be enriched in families [82], and the IBD regions in populations can be used in genetic analysis [55, 82]. The IBD analysis assumes that individuals who share a rare, phenotype-causing, variant inherited from a common ancestor also share a longer chromosomal region surrounding the causative variant [81, 85, 86]. The analysis identifies these IBD regions between pairs of individuals taking into account the genome-wide relatedness between this pair [81], estimated haplotype sharing [85], or identical haplotype matching [86]. To find the locus harboring the causative variant behind the phenotype, the IBD sharing between cases can be compared to IBD sharing between controls [81, 85]. Homozygosity detection can also be used to identify IBD. It is especially useful when studying a phenotype that shows recessive mode of inheritance [81, 87].

Haplotype analysis A general IBD analysis can be extended to investigate haplotypes. A haplotype is a chromosomal segment that is inherited in a pedigree or in a population without recombination[19]. In population level, haplotypes are observed as haplotype blocks that have reduced haplotype diversity or high LD, depending on definition [88]. Examples of haplotypes are visualized in Figure 2.7.

As genome is inherited in haplotypes, it is organized accordingly [89, 90, 91, 55]. Recent mutations are expected to be inherited in a specific haplotype [81, 90, 92]. However, single-marker tests do not capture the effect of rare variants, especially with a small number of samples [93, 94, 55]. Thus, haplotype-based analysis has been expected to increase statistical power due to reduction in dimensionality [90, 93]. Therefore, various efforts have been implemented to account for haplotypes

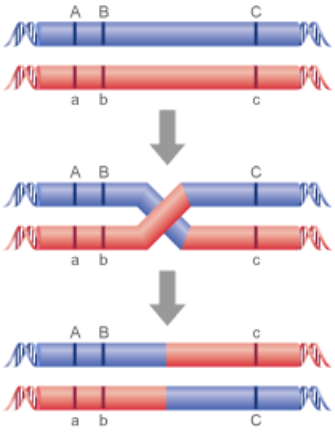

Figure 2.7: The genetic analysis is based on the genetic linkage between loci. The recombination of maternal and paternal chromosomes produces a new combination of the chromosome. The up-most image shows the original chromosomes of a parent. In the middle, the chromosomes are recombined to produce two copies that can be transmitted to the offspring. Here the uppercase letters show genetic markers passed from the father and lowercase letters those passed from the mother. The causative allele is segregating with the marker A, which is more often inherited with B than with marker C. This information is used in linkage analysis. The statistical association analysis, instead, finds association between marker A and the phenotype. The IBD analysis targets to identify the chromosomal segment carrying marker A from multiple individuals with common ancestry. Haplotype analysis identifies the paternal haplotype with marker A based on haplotype with markers A and B inherited to offspring. This figure is published by Wellcome Trust websites under Creative Commons Attribution 2.0 UK license [link to file 20.11.2012, http://genome.wellcome.ac.uk/doc_WTD020778.html].

in statistical association analyses [81, 89, 92, 91, 94, 95, 96, 97, 98]. Typically those methods identify haplotypes and test for their statistical association with the phenotype.

The haplotype identification approaches vary, as the haplotypes are identified using various sliding window approaches [99, 100, 97], based on haplotype blocks [93], by identifying haplotype clusters [98], building a cladistic model [92] and taking the haplotypes that are divided to multiple loci [91, 95]. Often haplotype information is not available in the genotypes, and the analysis requires haplotype estimation, *i.e.,* a phasing step [57]. The purpose of phasing is to resolve how alleles are inherited from a father and a mother. Haplotype detections use genotype data that is either with phasing [92, 97, 98] or without phasing [91, 93, 94, 95, 96, 99]. The haplotype analysis methods are targeted to genome-wide analysis [93, 98, 99] or to investigate limited chromosomal areas [96, 100].

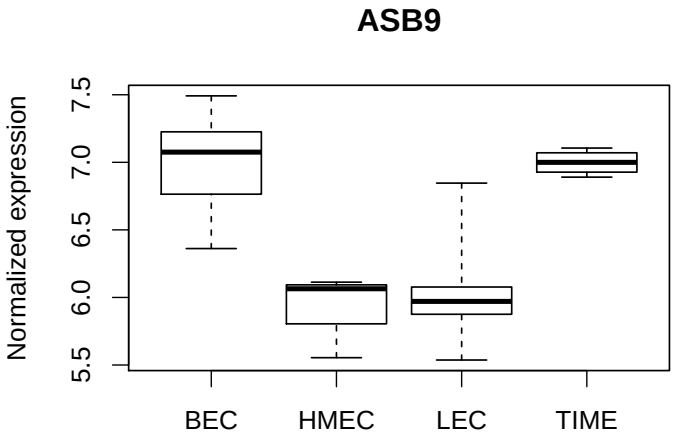

Figure 2.8: Box plot visualization of gene expression measurements of a gene *ankyrin repeat and SOCS box containing 9 gene* (*ASB9*) from samples in different cell lines: human microvascular endothelial cell line (HMEC), telomerase immortalized microvascular endothelial cell line (TIME), blood endothelial cells (BEC), and lymphatic endothelial cells (LEC) from the normalized gene expression data used in Publication III. The average gene expression of BEC and TIME samples is higher compared to average gene expression of HMEC and LEC samples. The gene expression analysis identifies genes where such differences exist.

2.3.2 Mining of gene expression differences between phenotypes

The analysis of gene expression data usually detects genes that have differential expression between samples, or studies the similarity of gene expression between genes [63]. The whole-genome gene expression analysis usually finds a small set of genes to be differently expressed between the samples while the majority of the genes do not show significant differences [72].

Usually a gene expression study includes gene expression measurements from several technical or biological replicates under one or multiple types of conditions from which the gene expression patterns are quantified. Typically, normalized gene expression values are used in logarithmic scale to ease computation and make gene expression differences independent from the level of gene expression [63].

Figure 2.8 shows distribution from the gene expression measurements of one differentially regulated gene (multiple hypothesis corrected p-value < 0.05), where the average of the gene expression is significantly different between two of the sample groups although the expression values may vary within the sample groups. The sample groups in Figure 2.8 are blood endothelial cells (BEC), human microvascular endothelial cell line (HMEC), lymphatic endothelial cells (LEC), and telomerase immortalized microvascular endothelial cell line (TIME). These data were used in Publication III. Usually, gene expression analysis aims to identify genes that show these kind of gene expression distributions.

The differential expression, which is also stated as up- or downregulation, can be tested using a statistical test of difference between expression distributions [63, 70]. The differentially expressed genes can be identified using a p-value threshold, or simply by calculating a *fold change* of gene expression between samples [63, 70]. These values can be used jointly or separately to evaluate gene expression differences. A statistical test, like Student's t-test, gives a confidence value, *i.e.*, a p-value, to reject the null hypothesis, which is interpreted as confidence for the true gene expression difference [73]. The fold change value is used only to rank the genes and quantify the difference because it does not include any information about variance of the gene expression [73]. At simplest, the gene expression analysis results a set of differentially regulated genes.

The similarity between gene expression profiles can be analyzed using supervised [63] and unsupervised (clustering) [63, 77, 73] learning approaches. Also, an enrichment of pre-defined gene classes can be tested to interpret the results [73]. These classes can be chosen for each study, but often gene ontology [101] classes are used [73].

The gene expression analysis results may contain false positives and false negatives [73]. A golden standard has been that the significant genes are validated using independent biological samples and the same experimental settings to rule out reporting false positives, but false negatives are rarely evaluated in an independent study [73].

2.3.3 Use of existing biological knowledge

Molecular measurements and the subsequent analysis alone are rarely enough to understand the molecular mechanisms behind phenotypes. Instead, the results are interpreted or the study is designed in the light of existing knowledge. Here, biological databases are key resources. An essential feature of these databases is that they can be accessed automatically, which allows systematic use of all information stored in these databases. Furthermore, this enables the use of computational approaches to modify the data and to integrate biological knowledge with study results [102, 103, 104, 105]. Often integration aims to understand the findings of the analysis either based on gene annotations, for example, characterizing affected genes in breast cancer tumors [5], molecular functions, or molecular networks [106, 5]. The existing knowledge can also generate new candidates for further analysis which can be based on the molecular networks and functions [103].

Biological experiments and the subsequent analysis may produce such a large number of candidate genes that downstream analysis and validation is not feasible for all of them [76]. Here, the protein-protein interaction (PPI) or another functional relation can be used to infer which candidates are the most relevant by using a functional link between a causal gene already identified and the candidate gene [76]. However, one data source, such as PPI, is rarely enough to create a comprehensive context for the candidate gene prioritization, thus multiple sources should be integrated [76]. Also, the candidate gene prioritization can be executed backwards from knowledge to measurements, by using existing biological knowledge to generate candidates to guide the design of biological experiments [76].

Biological knowledge bases In this thesis, all computational resources that provide information from multiple levels of molecular functions to support biomedical research are seen as biological knowledge bases. They can be either traditional curated databases or tools to generate knowledge. Table 1 shows a grouping of knowledge bases relevant for this work. The biological knowledge bases

Table 1: The grouping of knowledge bases. The biological knowledge used in biomedical research is stored into various centralized repositories. These repositories can be grouped into single mutation effect, chromosomal or protein annotations, systems biology, and systems biology meta server resources.

Type of knowledge base	**Use of the knowledge base**	**Examples of knowledge bases and computational resources**
Single mutation effect	Prioritizing of findings, creating hypotheses, selecting tagging SNPs, listing known mutations	Sift [107], SNPs3D [108], PolyPhen [109], Cosmic [110]
Gene and protein information	Interpretation of findings, reference knowledge	Ensembl [111], Entrez [112], InterProt [113], HapMap [55]
Systems biology	Interpret systems effect of the findings, understand dynamics of the system, predict new candidates	KEGG [114] , Reactome [115], PINA [116], GO [101]
Systems biology meta servers	Integrate information, computational modeling of the system	Moksiskaan [103], BioMine [102]

provide information about the effect of single mutations or variations, chromosomal or protein annotations, and biological systems. In addition, meta servers collect information from multiple sources and integrate them into a knowledge base.

The single-mutation knowledge bases give information on whether variants potentially affect gene regulation or protein [117, 118]. They can also be used to generate a set of interesting variants for instance to study segregating SNPs in a population [118]. In the case of diseases, the question is even simpler: is the variant detrimental for the protein, thus preventing normal cell functions [117]? This single-molecule information can be predicted [118, 117], for example, using sequence conservation [107] or a larger vector of molecular features [109, 119], or it can be verified by mutation-phenotype consequence, such as in cancer [110].

Biomedical research uses chromosomal annotations to give a species-specific reference to the data. Genome databases maintain the most recent annotations of the genomes of various species [112, 111]. From these databases one can find the chromosomal coordinates of the genes, known variants, transcripts, peptides, citations to original publications [112, 111], allele frequencies in populations and population linkage structure [55]. In the study of gene products, proteins, information from the research is stored into protein databases including annotations such as protein sequences, domains,

or protein structure [113].

Studying individual genes or mutations follows the reductionistic paradigm where understanding of the system should be achieved by studying its parts. However, systems biology aims to understand the system as a whole, and in this work the existing knowledge of molecular interactions, *i.e.,* molecular networks, is essential [120]. The networks can be used to interpret the causal consequence of the findings or predict new hypothesis.

The pathway, PPI, and molecular function databases can be seen as systems biology databases [120, 121]. These databases have different emphasis: pathway databases usually contain information about molecular signaling transduction or other molecular pathways where information is conceptualized to biologically defined pathways and networks [120, 121]; PPI databases list verified or predicted protein pairs that have physical interaction [120, 121]; and molecular function databases link molecular functions and cellular locations with genes, to describe relationships among these concepts [120, 101].

Meta servers in systems biology integrate information from existing molecular network databases [121, 103, 102]. These meta servers in systems biology also compile the data into a complete network representation of this information [102, 103]. They also overcome the scattered nature of the databases, as they combine information from multiple sources [121].

2.3.4 Creating predictions from gene expression data

The differentially expressed genes from gene expression analysis can be used directly as biomarkers to discriminate phenotypes, although a more complex panel of gene expression signatures may be a more effective biomarker [48]. Classification, or supervised learning, can capture complex features in the data, and thus produce a powerful predictor for a phenotype [48]. Although classification is an integral part of gene expression analysis [73], it is also a very large field in biomarker discovery covering many computational methods [48, 122, 123, 124]. The goal of classification analysis is to find the gene expression pattern and classifier that yield the best prediction for the phenotype [78]. The technical modifications, such as probe mapping and normalization, in gene expression microarray data processing may complicate creating a predictive classifier [78], however, this may not be the case in other gene expression measurement technologies.

Computational classification by supervised learning The computational classification uses training data where each gene expression signature is labeled according to the phenotype, such as disease and healthy [78]. The classifier defines a mathematical function for producing the classification based on the data, and, from the training data, it learns parameter values that best produce the phenotype labels [78]. The outcome of this process is a classifier that can be applied to new datasets to predict previously unknown phenotype labels [78, 125].

Classification is often impaired due to a large number of genes in the gene expression signatures [78]. The learning typically includes feature selection, which is gene selection in the case of gene expression data, to produce a set of genes used in the classification [48, 78, 125, 126]. The feature selection aims to select a subset of the original dataset and, unlike other, more general dimensionality reduction methods, does not transform original values [126]. When the number of genes in the dataset is high, as it usually is in gene expression microarrays, the feature selection requires specialized algorithms [48, 78, 126]. However, in a small set of genes, it is possible to go through all available gene expression combinations to find the optimal set of genes for classification.

Table 3: Hypothetical example of a classification to predict a disease. The confusion matrix shows the quantity of correctly classified true positives (TP) and true negatives (TN), and incorrectly classified false negatives (FN) and false positives (FP).

	Disease	Healthy
Predicted disease	10	2
	Correctly classified TP	Incorrectly classified FP
Predicted healthy	3	12
	Incorrectly classified FN	Correctly classified TN

Some widely used classification functions deploy different approaches to achieve the classification. One powerful method to classify instances is artificial neural networks (ANN) [127]. They are comprised of neurons and weighted connections between them, where each neuron defines an activation function to calculate an output that is fed to the next neuron, and the network output is the classification for each instance [78, 125, 127]. The ANNs have many variations; one type of ANN is multilayer perceptron (MLP), which is a feed-forward model where neurons are organized into layers, and each layer feeds the output to the next layer of neurons [127]. Each neuron implements a non-linear activation function that is a combination of all the inputs [127]. The parameter values learned from the data are the weights for the combination function for each input [127]. Finally, learning is achieved by minimizing an error function [127].

Some other classification functions that have been used in biomedical analysis are: Instance based learning, where the prediction is based on similarity with known instances, and similarity metrics are typically distance metrics, such as Euclidean distance [78, 125]. Linear discriminants define a linear function that returns a value from gene expressions, which is used to divide instances to classes based on a threshold [78]. Statistical classifiers use a probability model for classification of the instances [125]. Support vector machines (SVM) use a hyperplane that maximizes the distance to instances of classes in multi-dimensional space [78, 125]. Logic based methods [125], like decision trees, create a set of rules, that can be in tree structure, for classification [78, 125]. If the data has multiple phenotype classes, these classification methods may require some adjustments [125, 78].

Biomedical research has multiple examples of biomarker discovery using computational classification. The breast cancer prognosis predictor from 70 gene expression signatures [48] is a well-known example. It uses wrapper feature selection that selects genes using rank-based filtering while evaluating the classification along the gene selection [126]. Another study identified a 74 gene signature to predict breast cancer patients' response to chemoterapy [122]. This study used a specialized algorithm for gene selection and compared multiple classifiers to find the best performing classifier for the treatment outcome [122].

Classifier evaluation Classifiers are evaluated by using a validation dataset, and the evaluation is based on correctly and incorrectly classified instances [78, 125]. The validation dataset can be an

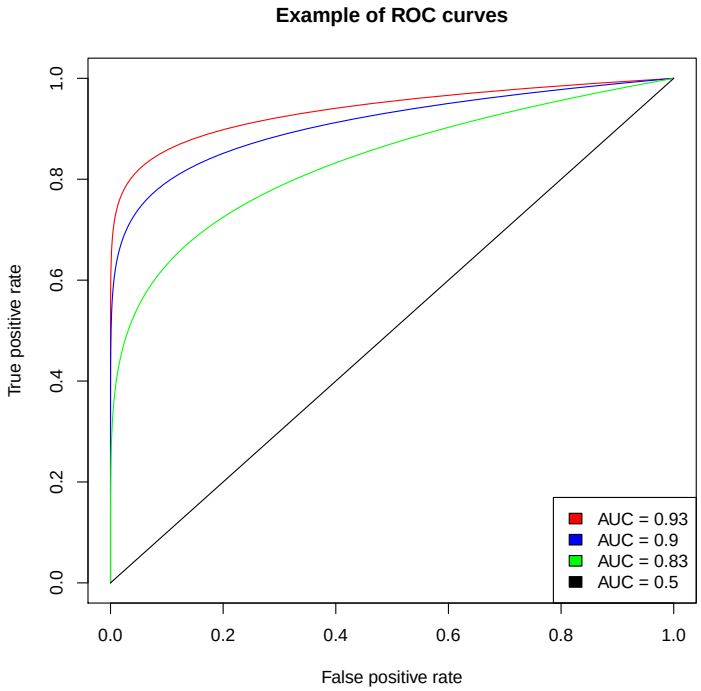

Figure 2.9: An example of four hypothetical receiver operating characteristics (ROC) curves and their corresponding area under curve (AUC) values. Black curve with AUC value 0.5 presents poor classifier where in average each true positive classification produces another false positive classification. Red curve with AUC value 0.93 is produced for most accurate classifier in this example.

independent dataset or cross-validation can be used [78, 125]. The independent training and validation datasets can be constructed by splitting the original dataset into two sub-datasets [125, 78]. In cross-validation, the dataset is divided into subsets that each are used as the validation dataset in turn, and the statistics from these validations estimate the accuracy of the classifier built from the whole dataset [78].

The validation produces a so called confusion matrix that shows how many times classification was correct. In practice, the confusion matrix is a table where the columns are real classes and the rows are predicted classes, and it can be used for visualization and as a source for performance statistics. Table 3 shows an example of a hypothetical confusion matrix.

A typical metric for classification accuracy is the receiver operator characteristic (ROC) curve, which is a graphical presentation of classification sensitivity and specificity over different classification parameter stringencies where, for each class, true-positive rate is plotted against false-positive rate [128]. Area under curve (AUC) is the value that describes the size of the area that the curve occupies from the graph [128], and it is used to give a score for a ROC curve. Figure 2.9 shows an example of four hypothetical ROC curves and their corresponding AUC values. When evaluating a ROC curve, random guess produces 0.5 AUC value and perfect classification produces the value 1.

One somewhat rarely used metrics is the κ-value, which describes how much the agreement on

classification results differs from random guessing between different raters [129]. This concept can be extended to evaluate the agreement between true and predicted classes, as is done in the data mining software, Weka [130]. There, κ-statistics are calculated from the values in a confusion matrix as $\frac{(P_o - P_c)}{(1 - P_c)}$, where P_o is the observed agreement and P_c is the agreement that would be achieved by chance.

Classification has also a risk of overfitting, which means that, although generated classifier produces correct predictions, it cannot be generalized beyond the training data [73, 78]. The risk becomes even more prevalent when a complex function with multiple parameters is trained [78].

3 Aims of the study

The research in this thesis aimed to create computational methods for finding molecular features contributing to phenotypes on multiple levels of biomedical research. These levels include genes in the same functional context, inherited phenotypes, plastic cell type specific phenotypes and predictive gene expression signatures.

Publication I Cancer evolves from aberrant molecular signaling, therefore, systems level view of molecules behind the phenotype is needed. Publication I presents a meta server (CANGES) that integrates information from multiple biological knowledge bases [114, 115, 116, 111, 55, 107, 108, 131, 109] and creates hypotheses for understanding the molecular basis of the phenotype. CANGES also suggests new SNP candidates that can be tested experimentally by either directly genotyping or using linkage with the tagging SNPs. Publication I demonstrates the use of this approach in *tp53* network and in candidate gene generation for survival analysis in glioblastoma multiforme (GBM).

Publication II The analysis of phenotypes in *rare variant – moderate penetrance* assumption usually cannot be done using traditional statistical methods. Publication II presents an analysis method (Haplous) that uses shared haplotypes to identify IBD regions from multiple individuals having a recent common ancestor. Also, in Publication II performance of Haplous is tested in various settings with simulated data and with data that has haplotype phasing errors. It is also used to analyze shared haplotypes in family members affected by lymphoma in multiple Finnish families.

Publication III The cell type specific phenotype is determined by altered gene expression. The identification of the differentially expressed genes between cell types is often hampered by a limited number of samples. Publication III pools multiple samples from multiple gene expression platforms to analyze molecular determinants behind blood endothelial cells (BECs) and lymphatic endothelial cells (LECs). The approach identifies novel markers that are validated. This approach can be extended to any other tissue types.

Publication IV The studies that identify molecular determinants behind phenotype rarely result in one marker but a list of molecules that may have combinatorial effects on the phenotype. Publication IV shows how a predictive gene expression signature of multiple genes can be identified from the expression data. In this publication we apply computational classification to acquire the molecular signature that best predicts recovery from cow's milk allergy (CMA).

4 Materials and methods

The overall goal of this work is to develop computational methods, whose utility can be demonstrated using SNP and gene expression datasets. The phenotypes selected to the analysis manifest on different organismal levels: cancer, whose cellular regulation has changed from the normal, inherited cancer predisposition, cell type differences, and transient cow's milk allergy. The computational methods developed aim to discover various molecular features behind these phenotypes and also detect new biomarkers. This section gives an overall description of the materials and methods. The details for technologies and parameters can be found from the original publications.

4.1 Data

The material used in this study consists of measurements of genetic variation and gene expression. Table 4 shows a summary of types of data used in the publications. The measurements of genetic variation are from high-throughput SNP technologies, and the gene expression measurements use both whole genome approach and more targeted expression data from selected genes. This section gives an overview of the data. The details can be found in the original publications.

4.1.1 The genome-wide SNP array data

Publication I and Publication II used genome-wide SNP microarray data. The data were used to estimate the effect of inherited genotypes to the phenotype.

To show the utility of integrating information from biological knowledge bases to enrich the genetic analysis, we used genome-wide SNP array data from cancer patients in Publication I. The Cancer Genome Atlas (TCGA) database provides clinical information and high-throughput measurements from cancer patients and their tumors from heterogeneous measurement technologies. The first launch of the TCGA database consisted of glioblastoma multiforme (GBM) patients and a few unaffected seizure patients as controls [132]. GBM is a type of grade IV tumor that originates from the glial tissue in the brain. The patients show very poor survival with median time of 12 months [132]. For Publication I, we used genome-wide SNP data derived from 209 GBM patients' blood samples from the TCGA database. The SNPs were genotyped with 550K SNP microarray.

In Publication II, we used genotypes from three Finnish families to test our approach to identify shared haplotypes from individuals having a recent common ancestor. These families may have a lymphoma predisposition, since multiple individuals in these families are affected with several different lymphoma subtypes. The lymphoma subtypes in this analysis were: nodular lymphocyte predominant Hodgkin lymphoma (NLPHL), T-cell/histiocyte rich B-cell lymphoma (TCRBCL), non-Hodgkin lymphoma (NHL), and classical Hodgkin lymphoma (cHL).

We had the DNA from blood of nine affected family members and, when possible, their parents. Also, DNA from the children and siblings of deceased lymphoma-affected family members was used. This resulted in 29 samples. All the samples from the lymphoma families were genotyped with 370K SNP microarray. Haplotype phase for these samples was predicted based on the pedigree using MERLIN [83].

To control for typical haplotypes in the Finnish population, we included genotypes from 250 unaffected Finnish individuals from the Nordic Center of Excellence in disease genetics (NCoEDG)

Table 4: List of data types used in publications. PUBLICATION I and PUBLICATION II used data from SNP microarray measurement technology. The data set in PUBLICATION I used the genotypes of glioblastoma multiforme (GBM) patients measured from germline. PUBLICATION II used germline genotypes from Finnish lymphoma families, unaffected controls, and family trios from a HapMap database. PUBLICATION III integrated three different Affymetrix gene expression microarray platforms from a Gene Expression Omnibus (GEO) database. PUBLICATION IV used gene expression data from 12 genes measured by RT-qPCR.

Publication	Measurement technology	Description
PUBLICATION I	SNP microarray	SNP genotypes of GBM patients measured from germline
PUBLICATION II	SNP microarray	SNP genotypes from germline of lymphoma families and unaffected controls
	SNP microarray	SNP genotypes of family trios from HapMap database
PUBLICATION III	Gene expression microarray	Three different Affymetrix platforms from GEO database
PUBLICATION IV	RT-qPCR	Gene expression of 12 genes

control database [http://www.ncoedg.org/]. The haplotype phase of these controls was predicted using HaploRec [133].

The HapMap database has genotypes from parent offspring trios whose haplotype phase can be determined based on the pedigree [55]. As the haplotype phasing may produce uncertainty to the data, we used phase-known SNP microarray data from HapMap phase 3 database using European population (CEU) and chromosome 12 [55]. The 29 samples from these trios were used as a reference in PUBLICATION II to study the effects of haplotype phasing in the results. The haplotype phase was estimated using HaploRec [133].

4.1.2 The pooled genome-wide gene expression data

The genome-wide gene expression data was used in PUBLICATION III where we identified new gene expression markers that determine the phenotype in cell types. We downloaded the raw gene expression data of 33 BEC samples and 14 LEC samples from public Genome Expression Omnibus (GEO) database [34]. These samples are from three different Affymetrix platforms: Human Genome U133 Plus 2.0 Array (14 BECs, 10 LECs), Human Exon 1.0 ST Array (7 BECs, 4 LECs), and Human Genome U133A 2.0 Array (12 BECs).

Prior to the study in PUBLICATION III, 189 candidate samples from relevant studies were collected for the analysis from 18 different gene expression platforms. 112 samples were BECs, 40 LECs, and 37 samples were unsorted microvascular cells. The unsorted microvascular cells were discarded at the beginning since they could not have been included to comparison of BEC and LEC cell types. After discarding unsorted microvascular cells, the data had 17 different gene expression platforms. Other

candidate samples, except the 47 samples from Affymetrix gene expression platforms, were discarded because of various technical problems in integration of the samples.

4.1.3 The RT-qPCR data

The gene expression data from 12 genes used in PUBLICATION IV were measured from mononuclear cells of peripheral blood from 44 children after β–lactoglobulin stimulation *in vitro*. 13 of these children were non-atopic healthy controls; 16 children had had CMA, but they had achieved the tolerance by the age of three; 15 children had persistent CMA. The gene expression was quantified with the RT-qPCR method. The genes selected for study were T-cell markers *T-bet*, *GATA-3*, *IFN-*γ, *CTLA4, IL-10, IL-16, TGF-*β*, FoxP3, Nfat-C2, TIM3, TIM4*, and *STIM-1*.

4.2 Bioinformatics workflow execution

All methods presented in this thesis are implemented as pipelines that can be executed using an Anduril bioinformatics workflow engine [105]. Anduril is component-based: each analysis step is implemented as a component that can be connected to each other using their inputs and outputs. These inputs and outputs are usually in files on a computer disc, and Anduril takes care of passing the data from one component to another. Also, intermediate results from each component are stored. Anduril transforms the pipeline into a network presentation and evaluates the dependencies between each component. Information about the component dependencies are used to parallelize the execution as the independent components can be executed concurrently and, in the case of re-execution, only components that have modifications are executed and results from others can be used without re-execution. As the dependencies are clear, modifications and maintenance of analysis pipelines are easy. The pipeline itself is constructed using special language, AndurilScript, which can be used as description of the analysis.

4.3 Data integration method for systems biology

PUBLICATION I introduces a method for creating hypotheses and for interpreting results using existing knowledge from systems biology databases, gene databases, and single-molecule resources. This method, CANGES, is essentially a meta server that integrates information from multiple biological resources. The schematic representation of the CANGES workflow and technologies is shown in Figure 4.1. CANGES identifies *focal genes*, *i.e.,* candidate genes for the causative molecules behind the phenotype. For the focal genes, CANGES collects *central SNPs,* which are SNPs in the focal gene's chromosomal region. For these SNPs, CANGES collects many relevant annotations into one repository for later use, and creates a list of markers that are in LD with the central SNPs. Finally, CANGES fetches predictions on the coding SNP's impact.

CANGES assumes that the putative impact of focal genes to a given phenotype is defined by the same function that the already known genes have. These genes are used in CANGES as *query genes* to seed the focal gene search. In practice, the focal gene is in the same pathway or having a PPI with the query gene's protein product. The focal genes are collected from pathway databases KEGG [114] and Reactome [115], and from the PPI database PINA [116], using the query gene as a search seed.

The focal genes are annotated for their central SNP from the Ensembl database [111], and the markers that are in LD with those central SNPs are collected from the HapMap database [55] using

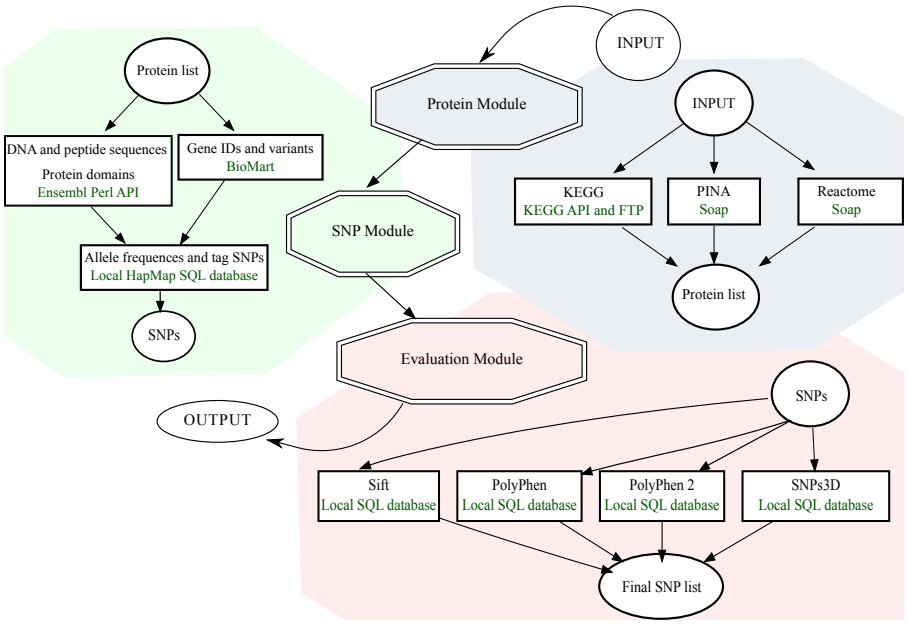

Figure 4.1: The CANGES data integration workflow consists of three modules that each access different databases to collect the data. The octagons are the modules, rectangles are data accessions, circles are data files, and the arrows represent the data flow in the system. This diagram shows the knowledge bases that are used, and the technology for accession is denoted with green font. SNP Module requires the collection of information from the Ensembl database using different queries and connection types.

an appropriate population. Ensembl also provides annotations of the central SNP's function and their chromosomal region.

Finally, the SNPs falling into coding regions are evaluated for their function. For this, CANGES uses the SNP function prediction resources SIFT [107], SNPs3D [108], PolyPhen [131], and PolyPhen2 [109]. These SNP function predictors give an estimate for damaging effects of the coding SNPs to the protein. SIFT relies on peptide sequence conservation [107] and SNPs3D, PolyPhen, and PolyPhen2 classify amino acid changes based on a set of molecular features [108, 131, 109].

CANGES is implemented on the Anduril bioinformatics workflow engine [105], and, when possible, it uses the public programming interfaces provided by the databases, including BioMart query tool [134].

We applied CANGES on two research settings: we characterized the focal genes from the *tp53* molecular signaling network and created a list of the central SNPs; and we generated a set of central SNPs in GBM, which were analyzed using high throughput genotype data. The central SNPs in GBM were identified using six putative GBM predisposition genes [135] as query genes: *telomerase reverse transcriptase* (*TERT*), *cyclin-dependent kinase inhibitor 2B* (*CDKN2B*), *cyclin-dependent kinase inhibitor 2A* (*CDKN2A*), *regulator of telomere elongation helicase 1* (*RTEL1*), *coiled-coil domain containing 26* (*CCDC26*) and *pleckstrin homology-like domain, family B, member 1* (*PHLDB1*) [135].

4.4 Rule-based haplotype analysis from genome-wide data

We developed a rule-based computational method, Haplous, to identify *shared haplotypes* (SH) from distantly related individuals (Publication II). Haplous uses rules to identify and filter SHs that are seen in cases showing the phenotype, but which are almost absent in unaffected controls. The rules used in Haplous are expert defined and implement a natural deduction of what makes a SH interesting.

Haplous uses phased genotype data that can be bi-allelic SNPs or multi-allelic markers. It identifies the SH by comparing each sample pair using a fixed-size sliding window that allows mismatches in the SHs. The window size is determined by markers in a window, although Haplous has support for using base pairs as the metric. Then the SHs between sample pairs are collapsed into one data structure, which allows identifying with whom each sample shares the haplotype. The detailed algorithms are provided in Publication II, but Figure 4.2 visualizes the process of SH identification and the output it creates. From these SH data, Haplous filters interesting SHs using rules. These rules take into account homozygosity and heterozygosity and allow defining the frequency in which the SH is seen in cases and in controls. It also filters SHs based on informativeness that is calculated from allele frequencies. Finally, SHs are given a score according to their length and sharing in cases, and the score is penalized if it is shared by controls. Haplous also doubles the score if SH is shared in a homozygous region. Haplous identifies loci where multiple individuals share the same haplotype, which is more interesting than finding multiple individuals to share different haplotypes.

We first compare the Haplous results in a setting where there are no switch errors from the phasing step in the data and then when the haplotype phase estimation has produced switch errors. For this we use samples in the HapMap [55] database. Then we apply Haplous to Finnish lymphoma families and identify chromosomal regions shared by multiple individuals. Finally, we simulate a dataset that corresponds to our study with lymphoma families and compare it with an existing method [85]. We also test the effects of false assumptions to the results by using inappropriate parameter values.

4.5 Pooling heterogeneous studies to compare cell types

In Publication III we use gene expression analysis to identify differentially expressed genes between phenotypes. To acquire a large set of samples for analysis, we increase the number of available samples to 47 by pooling samples from three genome-wide gene expression platforms from a public gene expression database GEO [34].

For pooling the samples, we created an Anduril pipeline for data processing. First the pipeline takes a list of GEO sample identifiers as an input and downloads raw gene expression data, and when the raw values are not available, downloads preprocessed data. An output from the download is a database that consists of gene expression signals of each sample in separate files, annotations of data location in GEO server, information about biological source of specimens, and some technical information. This database also contains information about gene expression measurement platform types. After download, the processing pipeline allows manual intervention to exclude gene expression platforms from further processing by setting a flag to the database. For preprocessing the pipeline transforms raw gene expression signals into gene expression values, normalizes samples within each platform separately with user-defined normalization methods, and annotates each gene with their Ensembl gene identifiers [111]. As a final step of preprocessing, this pipeline produces a quality

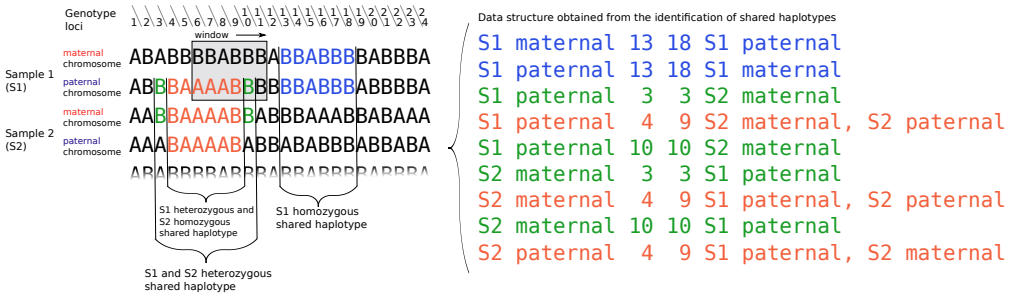

Figure 4.2: Visualization of the identification of shared haplotypes and output from this process in Haplous. Two samples in this visualization are denoted by symbols S1 and S2. For both samples, Haplous takes the phased chromosomes as input, and slides a fixed-size window over all pairs of available chromosomes. Alleles that are the same within this window constitute a shared haplotype (SH). SHs are visualized with colores (green, red, blue). As SHs between chromosomes are overlapping (for example the green and red SHs in loci from three to ten) they are collapsed into a specialized data structure visualized on the right, using corresponding colors. This structure shows which of the original chromosomes share the SH and allows easy lookup of homozygous and heterozygous SHs.

control report with visualizations of data including box plot, cumulative distribution, and hierarchical clustering for each used normalization method. This report can be used to select the preprocessed datasets to final analysis. After these steps all samples are integrated into one expression matrix using the gene identifiers to map gene expression values between samples. Then follows two filtering steps in which all samples that have missing values in more than user-defined percentage of genes are removed, and all genes that have missing values in more than the user-defined percentage of samples are removed. Finally the pipeline normalizes all samples with user-defined normalization methods to adjust all samples to the same scale. The actual gene expression analysis pipeline is only semi-automated since statistical tests and fold change calculations require several definitions of sample groups used in comparisons.

The 152 BEC and LEC candidate samples originally intended for the analysis had to be reduced into 47 samples because of the following challenges: 51 samples did not have raw gene expression signals available and they have been preprocessed using unknown protocols. Six samples were copy number variation measurements and could not be used. Automatic reading of the expression signals was not possible from four samples produced with non-commercial platforms. Furthermore, technical production of non-commercial platforms uses unknown and heterogeneous protocols, thus we decided not to create separate readers for them. One Affymetrix platform was represented with only one sample, thus that sample was also excluded. Finally, the average correlation between platforms was poor, except for the selected Affymetrix platforms where it was ≥ 0.68.

Unlike the typical gene expression studies, this approach requires two normalization steps to acquire comparable samples. The first step uses the platform dependent analysis method, which is necessary to combine the probe set expressions to gene expressions. Then all the samples are normalized again to enable inter-platform comparison. This pooled expression dataset can then be tested for differentially expressed genes. In this study, we used Student's two-tailed t-test combined

with fold change to find differentially expressed genes. We also visualized the findings with box plots. In addition, we performed the same analysis using a subset of samples, 14 BECs and 10 LECs measured with Affymetrix Human Genome U133 Plus 2.0 Array gene expression microarray technology.

4.6 Biomarker prediction using machine learning

In PUBLICATION IV we present an application for identifying biomarkers from gene expression data. For this biomarker discovery, we use machine learning technique, multilayer perceptron neural network (MLP), which is a type of ANN. We use an implementation in a Weka data mining software [130]. In MLP the neurons are organized in layers that process the information and feed it to the neurons on the next layer. The first layer is an input layer, the last is an output layer, and they are connected with at least one hidden layer. Here we allowed Weka to automatically organize a MLP. This organization has one input neuron for each gene used in the classification, one output neuron for each three classes and $\frac{\text{number of genes+number of classes}}{2}$ neurons in one hidden layer. We choose the optimal set of markers by testing all 4,095 marker combinations from the 12 genes. For each combination, we trained and evaluated a classifier for three phenotype groups. The optimal classifier was chosen using the confusion matrix, AUC value and κ-statistics.

5 Results and discussion

The result of this work is a set of computational approaches for analyzing the molecular background of phenotypes. Sometimes molecular profile in phenotypes can not be described with only one approach; often a multi-level analysis from genetics to gene expression is needed. Ultimately, this analysis can lead to more comprehensive understanding of molecular entities behind phenotypes, and, in medicine, improved screening, diagnosis and treatments. The work done for this thesis provides a database integration method to identify candidate genes and SNPs, presents a rule-based haplotype analysis to study rare familial phenotypes, shows how pooling of data from gene expression studies can create new insight on cell types, and shows that candidate genes can be compiled into classifiers that predict phenotypes.

5.1 Using information from databases to create novel hypotheses for molecular determinants behind cancer

In comparison to normal tissues, cancer is characterized by altered molecular signaling that allows cancer cells to multiply uncontrollably [37]. The widely studied *tp53* signaling network is a fundamental example of altered molecular signaling driving cancer [37, 41]. It is not only that *tp53* itself acquires mutations, but other genes contributing to the same signaling network produce the same cancer phenotype when their function is altered [41]. One recent example of a systems level effect is an activated *epidermal growth factor receptor* (*EGFR*) signaling in a colorectal cancer cell line, that makes the treatment targeted to mutated *v-raf murine sarcoma viral oncogene homolog B1* (*BRAF*) (V600E) ineffective [136]. The same treatment in melanoma is highly effective [137], probably because melanoma is mostly missing the *EGFR* signaling that acts parallel with *BRAF* [136]. Understanding such molecular signaling networks allows pinpointing molecules that are putative drivers of cancer, thus being lucrative targets for treatments or diagnostic biomarkers.

Publication I presents a metaserver, CANGES, that integrates information from multiple biological knowledge bases to generate hypotheses from the molecular functions behind a phenotype. CANGES can be used in cases where systems level molecular functions are the main interest, such as in cancer. Thus in Publication I, we show the benefits of CANGES in two case studies. In the first case study, we used a known driver gene behind cancer, *tp53*, as a query gene to produce a list of focal genes in breast cancer. This query produced 1,914 focal genes from the KEGG [114], Reactome [115], and PINA [116] databases. For these focal genes, we identified 47,163 central SNPs from the Ensembl database [111] and 3,465 tag-SNPs from HapMap European population (CEU) [55]. Then we predicted the impact of coding SNPs to the protein using four prediction tools: SIFT [107], SNPs3D [108], PolyPhen [131], and PolyPhen2 [109]. These four methods gave the same prediction of detrimental effects for 158 SNPs.

In the second case study, we identified focal genes for GBM genes *TERT*, *CDKN2B* and *CDKN2A* identified by a previous study [135]. The initial list of query genes included also *RTEL1*, *CCDC26* and *PHLDB1*, but they were not in the pathway or PPI databases. For GBM, CANGES returned 1,346 focal genes, which had 33,428 central SNPs that were tagged by 2,657 SNPs in HapMap European population (CEU). From these 35,622 SNPs, 1,888 were in the TCGA [132] genome-wide SNP microarray. For those SNPs we evaluated their survival effect. The most significant survival associated SNP was a tagging SNP for *5-methyltetrahydrofolate-homocysteine methyltransferase* (*MTR*) gene,

which has a PPI in PINA with *CDKN2A* that is in KEGG annotated to the *tp53* pathway. Thus, *MTR* is a novel candidate as a GBM driving gene. The largest effect was in a tagging SNP for a gene *cyclin B1* (*CCNB1*) that is located downstream from *tp53* in the *tp53* signaling pathway in KEGG annotations.

The implementation on Anduril bioinformatics workflow engine [105] allows easy maintenance of the database integration and possibility for re-execution of the analysis of CANGES. Therefore, the same queries were re-executed on 3.12.2012 to investigate whether the databases have accumulated more information since the original publication.

Indeed, CANGES identified 3,688 focal genes for the *tp53* network, which is a significant increase from the original 1,914 genes. The GBM query genes yielded a list of 3,710 focal genes, but these genes were found only for query genes *TERT*, *CDKN2B* and *CDKN2A*. Thus, candidate genes *RTEL1*, *CCDC26* or *PHLDB1* are still absent from these databases. Our first analysis for Publication I confirmed that the biological knowledge bases were not complete. Although their information has increased, as our newest analysis shows, a bias towards a subset of molecules still exists.

After the first release, Ensembl transcript identifiers have been added to CANGES results, and BioMart [134] has stopped supporting peptide changes, thus, these features have been removed. An unfortunate update has been the removal of KEGG from CANGES because of licensing policies; in year 2013 the KEGG programming interface changed, and efficient usage of KEGG database is no longer feasible. This change will require new implementations for CANGES. Possible substitutes for KEGG are WikiPathways [30] or generating new queries using Moksiskaan [103]. Also, PolyPhen2 [109] has completely replaced PolyPhen [131].

Our results show that using systems level biological knowledge from multiple sources provides valuable information for analysis. For instance, the gene *MTR* was not annotated to the *tp53* pathway, although it may have a role in the signaling through PPI in GBM. Also, the comprehensive *p53* network with its central SNPs provides a landscape to investigate major molecular determinants behind cancer cells. Although our case studies are from cancer, a similar approach can be used to other interesting phenotypes whether they are diseases or cell type specific phenotypes. This approach can be applied to generate putative genes or genetic variants to identify relevant biomarkers from experimental measurements.

5.2 Computational method to detect shared trait-associated haplotypes derived from a common ancestor

Publication II presents Haplous, an analysis method to identify haplotypes shared by multiple individuals. Haplous identifies those individuals who have shared haplotype in heterozygous and homozygous conformations and corresponding chromosomal regions. Haplous uses flexible rules that can be adjusted based on the research question at hand. Many methods for identifying the genetic variant behind a phenotype, whether they are based on single-variants or IBD regions, are based on statistical significance [85, 89, 81]. However, in many cases the phenotype itself or the samples that can be collected are not suitable for statistical testing because of low statistical power [82]. Nevertheless, these cases should also be analyzed carefully, as genetic variants have considerable contribution to disease risk [82]. Therefore, Haplous is targeted to *rare variant – moderate penetrance* paradigm, where the phenotype has a molecular determinant that is inherited from a recent ancestor and it causes an incompletely penetrant phenotype. Haplous is a rule-based method to detect haplotype

sharing in multiple individuals. Haplous does not require an exact hypothesis of an inheritance model.

To investigate the effect of incorrect haplotype phasing to the SH identification, we evaluated SHs from phase-predicted samples against SHs from phase-known samples. From this evaluation we calculated the true positive (TP) and false positive (FP) rates. This was done by reducing the window size from 2,000 SNPs to 1 SNP, and allowing one mismatch in the window. The TP rate was the number of SNPs that were included to SHs both in phase-known and phase-predicted datasets compared to all TP SNPs and the false negative (FN) SNPs that were discovered only from the phase-known dataset. The FP rate was quantified by the number of SNPs falsely included to SH compared to the number of SNPs that are different between samples. The window sizes from 30 to 100 produced a high ratio between TP and FP rates. The phasing accuracy depends on the population at hand [138]. For instance, a population of African origins has more haplotype diversity, thus a larger number of samples is needed to estimate haplotype phase correctly [138]. However, the phasing accuracy and its effect on the results can be estimated using available trios similarly as we have estimated in Publication II.

We applied Haplous to find SHs from three Finnish lymphoma families. Lymphoma is especially well suited for haplotype analysis as some familial predisposition has been confirmed in Hodgkin lymphoma [139]. We identified SHs using a 100 SNP long window and assumed that the mutated SH is shared at least by five individuals. This number requires the haplotype to be at least in two smaller families or in one big family. To exclude IBS haplotypes, we excluded those SHs that were in three control samples. This resulted in 1,288 chromosomal regions having SHs that were larger than 30 SNPs, which included 273 protein coding genes. We estimated that this region is significantly shared between individuals. We also used a text mining tool [119] to find genes associated with lymphoma terms. This resulted in seven genes. As molecular background of lymphoma is still largely unknown, the risk variant may lay in other, yet unidentified, genes. However, we discovered that one of the families does not extensively share haplotypes with two other families, thus, the members of this family may carry a variant from different origins.

We also compared the length of SHs from lymphoma families to the length of SHs shared by multiple controls, and found that the longest SHs were found in cases. Based on an analytical estimation, we also estimated that the interesting SHs that are shared by five or more family members are significant.

To compare Haplous with an existing method, Beagle [85], we simulated a chromosome for 100 families that are similar to our lymphoma analysis. Using this data, Haplous performed better in terms of locating the mutated region. However, we recognize that Beagle is targeted to a very different purpose than Haplous, as it identifies pairs of individuals and their IBD regions. This comparison shows that an approach implemented in Haplous is needed in the field of IBD analysis approaches.

The simulated data were also used to assess the performance of Haplous when the assumptions in the study are wrong. We used three different wrong assumptions: 1) Parameter values are too stringent or too loose; 2) the number of controls is too low; and 3) some of the cases are in fact unrelated controls. The stringent parameter values resulted in an increased number of FN with increased specificity, which means that Haplous returned less results, but the mutation was discovered more frequently and the location was more accurate. When the number of controls was reduced, Haplous identified the mutation robustly, but the chromosomal location of the mutation was mapped less precisely. In the worst cases, where cases and controls were mixed, Haplous did not return anything, but when the assumptions nearly corresponded the data, the performance of Haplous

increased rapidly to normal level.

When using Haplous, the parameter value selection can be done iteratively first assuming the most stringent conditions and then relaxing these criteria in next iterations. This way the first results finish very fast and they show the most striking features in haplotype sharing. The latter iterations can be used to expand the information of haplotype sharing into regions that are less confidently IBD. Our results show that when the assumptions are far from reality, as might be the case in the most stringent iteration, Haplous returns an empty result set instead of false positives. These results increase the confidence on our suggestion for iterative approach.

Currently, Haplous is applied to new research questions that are very different than those presented in the original publication. We have analyzed a large cohort of unrelated, or very distantly related, people. We have also applied this approach to compare SHs between families by first compiling a pseudo-haplotype for each of the families. These recent applications show that Haplous is very flexible and can be applied to various research questions. Also, the current implementation of Haplous has a new feature to filter SHs regardless to their homozygosity or heterozygosity.

5.3 Pooling gene expression datasets reveals novel markers for cell types

To get robust results for differential expression, a large enough number of samples is needed [140]. This number increases with the biological variability of the phenotypes, and with the amount of gene expression change that has an impact on the phenotype [140]. Some recent studies are limited only to a small number of samples, which was seen in our data pool, as one source study had only two replicates for each treatment [141]. In Publication III we pooled gene expression microarray measurements from multiple studies from a public database [34] to investigate the gene expression differences between BECs and LECs, and to find robust biomarkers to separate these cell types. Although, BECs and LECs are separate, differentiated cell types, they originate from the same tissue structure of endothelial cells [142]. During development, LECs separate from early vascular structures and commit to their distinct gene expression profiles that create a LEC specific phenotype [142].

The work that enabled using the pooled set of gene expression samples required implementation of a complete pipeline that produces information of the candidate BEC and LEC samples automatically and allows monitoring the quality of data and interfering with the analysis. The integration of several gene expression measurement platforms had obvious challenges to map genes with each other and to acquire comparable expression values. Additional problems were caused by non-commercial measurement platforms, unknown processing of the data, and meta data that is either missing or very difficult to access systematically. These challenges emphasize that integrating gene expression measurement from several sources requires careful consideration and technical implementation that enables managing gene expression data.

In the analysis of the pooled expression data, we identified 28 differentially expressed genes (fold change ≥ 2 and multiple hypothesis corrected p-value ≤ 0.05), from which we validated two genes (*MCAM* and *COLEC12*) using immunofluorescence staining. *Melanoma cell adhesion molecule* (*MCAM*) was selectively expressed in BECs and not in LECs, and *Collectin placenta 12* (*COLEC12*) was expressed in LECs and not in BECs. Although pooling the samples required many processing steps, this validation shows that the gene expression analysis was able to identify two novel markers for the cell types.

We compared the differentially expressed genes with genes found by three other analysis: two

previously published studies [143, 144] and an analysis of only one gene expression microarray platform from our dataset. Only few differentially expressed genes were found by our analysis and the previous two studies. This shows that many assumptions used in sample processing, measurements and analysis steps bring different bias to the results as expected according to a previous study [140]. Furthermore, differential expression is not a dichotomous state, but rather a continuous phenomenon, and gene expression studies define the threshold that states where a gene becomes differentially regulated. The differentially expressed genes from the analysis of our dataset as compared to only one gene expression microarray platform also showed partial overlap.

Our analysis shows that pooling samples from multiple gene expression platforms can produce reproducible gene expression biomarkers for a phenotype. This is especially useful since it allows using published data sets, which can lead to increased sample sizes. This may also give a hint of biological variability in the cell types as different conditions produce somewhat alternated sets of differentially expressed genes. Thus, these analyses are essential to discover the full molecular portrait behind the phenotypes.

5.4 Artificial neural network classification predicts phenotypes

In Publication IV we show how supervised machine learning can be used both to discover biomarkers and to understand complex relationships of molecular actors behind the phenotype. Methods from other publications of this thesis produce lists of candidate markers that alone are not necessarily causal or useful in practice [48]. Thus, biomarker discovery is an important step when identifying the molecular background of a phenotype. This step produces very focused information of the biology and may result relevant tools that can be brought to use in labs and clinics.

From the 4,095 marker combinations that were used for classification, gene expression signature of *FoxP3*, *Nfat-C2*, *IL-16*, and *GATA-3* produced the best predictive classifier for cow's milk allergy (CMA) persistency. However, in our study, selecting the optimal classifier was challenging, since we classified three different outcomes (non-atopic, tolerant CMA and persistent CMA). The classification result was a trade-off between the classes. When one class is predicted with high accuracy, two other classes may produce very poor classification. In our study misclassification between non-atopic and persistent CMA was considered a more serious error than mixing tolerant CMA with other groups. Here, we chose to favor the correct prediction of persistent CMA patients and minimize false positive prediction of CMA persistency outcome. In practice, we chose the best classification outcome using $\kappa-$value that favors overall correct separation of classes. The quality of classification was inspected from a confusion matrix, which resulted 12 persistent CMA patients classified correctly and one patient predicted to a healthy and one to tolerant group. Also, one tolerant and one non-atopic patient were falsely predicted to belong to persistent CMA group. Second best classification had already three non-atopic patients classified to persistent CMA group (five FP in total) although 14 persistent CMA patients were correctly classified. We concluded that this classifier identifies processes related to persistent CMA poorly. This data was suitable for supervised learning approach, since the phenotype, *i.e.* disease outcome, of the patients was known. Therefore we were able to use this information to identify complex biomarker combinations that best predict this phenotype.

Here, we were able to enumerate all possible gene combinations to find the optimal predictive gene expression signature. In a concurrent study [145], we created a classifier to predict the CMA recovery based on antibody binding on protein epitopes . This study had measurements from 289

epitopes, thus enumerating all epitope combinations was not possible. This required the use of a feature selection algorithm to select the most relevant epitopes for classification [145]. The feature selection is a typical step when using supervised machine learning.

6 Conclusions and future prospects

The focus of this thesis is phenotype. Although the molecular mechanisms in cells can be studied from purely intellectual aspects, usually the interest rises from very practical questions regarding the phenotype. Diseases are in particular focus, since targeted treatments would increase the life span or life quality of many people. In these cases, the molecular mechanisms affecting the phenotype are sought because repairing the disease causing molecules would be the needed treatment. Some encouraging examples for targeted treatments exist already in cancers such as melanoma [137] or breast cancer [146], where the specific molecular signature leads to treatment with a specific drug. Also, the diagnostic markers can improve the life quality when a treatment can be targeted to those that benefit from it and others will be spared from unnecessary side effects [49].

This thesis presents a panel of computational methods to discover the molecular background of phenotypes. These methods generate candidate genes based on information from genetics, gene expression and systems biology. These candidate genes can then be subjected to more detailed analyses in order to discover clinically relevant tools. Hence, this work provides new access to existing knowledge that will integrate systems biology into data analysis and interpretation of the results. This has been considered as a beneficial approach [77, 76], and its importance will increase as the quality and coverage of biological knowledge bases increases at the same time as studies will scale up and include increasingly heterogeneous measurements.

This work complements current methods to map inherited genetic variants to phenotypes in a research setting where traditional statistical methods can not be applied. These statistical methods have been very successful when studying high-penetrance Mendelian diseases or common variants in large populations, however, striking Mendelian disorders are mainly exhausted as a research material and common variants explain only a small fraction of inheritance [82]. These realizations accompanied by sequencing technologies have pushed the research towards finding rare variants in small families or studying isolated somatic events. The work in this thesis provides a method for the first types of studies, where IBD is used as a way to identify the inherited causal variation. Haplous can be applied to sequencing data, especially because it accepts multiallelic variants and its memory and time demand increases only linearly with the number of markers. This ability to accept multiallelic markers has already been exploited in an unpublished study of comparing pseudo-haplotypes between families.

The identification of inherited causal variants is hardly enough for understanding the phenotype, thus, cell type specific information of gene functions is needed. This was demonstrated in a recent study, where the same mutation in melanoma or colorectal cancer requires different treatments because of tissue specific gene activation [136]. The work in this thesis also presents a gene expression analysis of cell types by pooling multiple datasets from previously published studies. This approach utilizes scattered published gene expression studies that individually may have quite a small number of samples. When these are pooled together, the result may capture more of the biological variation and bring more power to statistical tests.

Finally, while understanding the biological background in phenotypes is in itself interesting, another important goal in biomedical research is to find biomarkers. In this step, a vast number of candidate molecules is turned into practical information with for example machine learning. In some cases this will also give new insights on the molecular mechanisms.

The essential characteristics of science is reproducibility and accumulation of knowledge that

create a cycle where hypotheses are tested and revised in the light of new information. PUBLICATION III is an example of such a case, revising the current paradigms, as we could show that the widely used practice to study immortalized cell lines may lead into false results.

Computational methods should support this iterative process. All the methods in this thesis are implemented on the Anduril workflow engine [105], which enables the application of these methods effortlessly on constantly evolving datasets. An example of re-execution was given in this thesis when the list of focal genes for PUBLICATION I was recreated using the most recent information. Here we can see that incomplete understanding of the molecular pathways has been complemented with new information, but still much remains to be discovered. All of the methods are freely available, and they can be integrated to other analyses pipelines in Anduril.

Efforts are currently focused on analyzing individuals with clearly genetically determined diseases to shed light on the molecular background of these diseases, but also they are expected to bring a novel treatment for affected people [58, 82, 147]. This has also pushed medicine to find personalized treatments [147, 148, 149], which can be extended to provide personalized genetics also to healthy people [150]. In addition, personalized genetics can be accompanied with gene expression measurements, which are essential especially in cancer genetics [148]. The terabyte-level of data produced by these measurement technologies have increased the importance of computational methods in understanding the molecular functions and isolating the driving molecules or clinically relevant biomarkers. The importance of using existing biological knowledge to pinpoint candidates from such analyses has been recognized to be a key component in ongoing biomedical projects [147, 149]. A diagnosis pipeline intended for clinical use has already integrated existing biological knowledge with a complete whole-genome sequence analysis [149]. In the near future, heterogeneous measurements, such as protein binding, protein expression or spatiotemporal information from biological processes, can be integrated with genome information and gene expression datasets. Extracting knowledge from these data will require extensive models that take into consideration the dynamic nature of biology.

7 Acknowledgements

This thesis was done at Department of Biochemistry and the Developmental Biology at Institute of Biomedicine and at the Genome-Scale Biology Program in the Research Programs Unit. The Institute for Molecular Medicine Finland has provided computational facilities to carry out demanding computations and data management during my work. I would like to thank FICS graduate school and especially its coordinator, Ella Bingham, for providing me traveling and start-up funding and possibility to network with other young scientists.

I sincerely thank my supervisor, professor Sampsa Hautaniemi, who has given me an opportunity to do my research in his research group. Additionally I have had an opportunity to learn from his excellent instructions and scientific advices. Sampsa Hautaniemi provides such a combination of great leadership, strong team and creative freedom that is a key for successful work.

I am thankful for Tiia Pelkonen for helping me in various parts in my thesis. She has spent numerous days for proof-reading my manuscripts and organized practical issues during the process. Also she has been a friendly and supportive colleague.

Sampsa's lab has provided me extraordinary support from which I am thankful for. We have helped each other in technical details, but also provided creative ideas to push the research onwards. Doing research in this group has been a privilege. I would like thank for those lab members that have worked with me: Ping Chen, Chengyu Liu, Alejandra Cervera, Rony Lindell, Ville Rantanen, Erkka Valo, Riku Louhimo, Kristian Ovaska, Miko Valori, Viljami Aittomäki, Marko Laakso, Javier Núñez-Fontarnau, Lauri Lyly, and Elmo Saarentaus. Furthermore, I would like to thank those lab members that have already relocated to other positions or having a break from working life: Mikko Kivelä, Vladimir Rogojin, Anna-Maria Lahesmaa-Korpinen, Elena Czeizler, and Lilli Saarinen. Also I would like to thank the newer lab members that I had only short time to work with: Julia Casado, Chiara Facciotto, and Amjad Alkodsi.

The Heli Nevanlinna's research group was my first position as a bioinformatician, and I would like to thank Heli Nevanlinna to giving me an opportunity to work with her excellent team. Especially I would like to thank Tuomas Heikkinen, who gave his insight to Publication I. Tuomas's contribution has been extremely important for the success of our project.

During my years in Faculty of Medicine, I have had close collaboration with professor Lauri Altonen's research group. I am grateful for the whole team that has provided high quality data and extremely interesting research questions. Lauri Aaltonen has truly created a talented group that is a privilege to work with. Rainer Lehtonen was my supervisor already during my master's thesis, and we have continued collaboration since. I am thankful for Rainer's patience in pinpointing me the details in genetics and tedious work he has done for acquiring me data in suitable formats. Also, Rainer has always had new ideas to improve my analysis. Pia Vahteristo continued the Rainer's demanding work with data analysis, and I would like to thank her for that. Silva Saarinen has ensured that the research questions in my analysis have been biologically relevant, and she has gone through the results from my analysis. This has emerged into long discussions where we have tried to understand haplotype sharing and inheritance of lymphoma predisposition. I would like to thank Silva for working with me in our projects. In addition, Pasi Rastas brought the final contribution to Publication II, which in part convinced us and reviewers of our work from which I am grateful to Pasi.

I would like to thank Johannes Keuschnigg and Sirpa Jalkanen for the excellent collaboration. Johannes and I had intensive and lengthy email discussions and phone conferences about our data,

results and biological relevance. Our collaboration was an example of fruitful integration of biological insight combined with computational methods. Also, whole team behind the project enabled our results. Therefore, I would like to thank all those people that gave their contribution.

I would like to thank Emma Savilahti, who gave me an opportunity to work in her research projects. Emma was truly enthusiastic of our project, and she has been a role model of a successful young researcher for me. I would also like to thank Erkki Savilahti and whole team that enabled my and Emma's work.

I warmly thank Jean-Babtiste Cazier and Ian Tomlinson for providing me an opportunity to visit their research groups in Wellcome Trust Center in University of Oxford. I was able to develop Haplous to scale up for large data sets and analyze deep sequencing data from tumors.

I would like to thank both my thesis pre-examiners, Garry Wong and Päivi Onkamo, for putting their effort on my thesis. Both pre-examiners gave important comments with fair and constructive critique. I have learned from this process, and I was able to improve my work by using the comments from both pre-examiners.

I would like to thank my old and new friends. In particular, I would like to thank Eeva Rissanen for sharing my ups and downs in my working days. Although my work might be incomprehensible for normal people, you know how I feel about it and where I have been successful and where not. Eeva Saksa I would also like to thank for listening me during my graduate studies and well before that and also for welcoming me and my family to relax in her family ranch. Maria Markkanen has also supported me and lengthly discussed with me about my work and plans during the time I have sat on her hair salon or just on her sofa. For this I am thankful for her. I would like to thank Anni Hanen and Outi Nieminen for being good friends and sharing my thesis process. I am thankful for Marko (Takku) Mylläri for giving a computer geek insight on various issues, we have always asked from each other an advice in tricky problems. Furthermore, the time spent together has been delightful. My crazy talkative friend, Nina Lagrtröm, has also my gratitude for good time and lot of discussions of all the topics in the world. Finally, I would also like to thank all my dear friends around Kallio. Although, you come from different backgrounds, are interested on different things, hardly no one practices bioinformatics. Thus, your company has truly been a time off from graduate studies. Also, I thank Pasi Haatanen for running our family enterprise for his part but also some funny moments outside of working life.

I would like to thank my family, mother Helena Heinänen, father Sakari Karinen, grand mother Helka Karinen, brother Aapo Karinen and his spouse Lilian Helkkola for all the support. My gratitude also is for my supportive aunts and uncles.

I thank my children Simeon and Sofia for being there. Also Laura, my bonus daughter I thank her for enriching my family. Finally, thank you Pete for going to grocery store, cleaning our home, taking the dogs out, making money for bills, and giving your love and support. All your effort has enabled this work.

Sirkku Karinen
Helsinki, May 2013

References

[1] Ahn, A. C., Tewari, M., Poon, C. S., and Phillips, R. S. (2006) The limits of reductionism in medicine: could systems biology offer an alternative? *PLoS Medicine 3*, e208.

[2] van der Greef, J., Martin, S., Juhasz, P., Adourian, A., Plasterer, T., Verheij, E. R., and McBurney, R. N. (2007) The art and practice of systems biology in medicine: mapping patterns of relationships. *Journal of Proteome Research 6*, 1540–1559.

[3] Sorger, P. K. (2005) A reductionist's systems biology. *Current Opinion in Cell Biology 17*, 9–11.

[4] Nik-Zainal, S. et al. (2012) The life history of 21 breast cancers. *Cell 149*, 994–1007.

[5] Koboldt, D. C. et al. (2012) Comprehensive molecular portraits of human breast tumours. *Nature 490*, 61–70.

[6] Mameli, M. (2005) The inheritance of features. *Biology & Philosophy 20*, 365–399.

[7] Bernstein, B. E., Meissner, A., and Lander, E. S. (2007) The mammalian epigenome. *Cell 128*, 669–681.

[8] Dezső, Z., Nikolsky, Y., Sviridov, E., Shi, W., Serebriyskaya, T., Dosymbekov, D., Bugrim, A., Rakhmatulin, E., Brennan, R., Guryanov, A., Li, K., Blake, J., Samaha, R., and Nikolskaya, T. (2008) A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biology 6*, 49.

[9] Mahner, M., and Kary, M. (1997) What exactly are genomes, genotypes and phenotypes? And what about phenomes? *Journal of Theoretical Biology 186*, 55–63.

[10] Freimer, N., and Sabatti, C. (2003) The human phenome project. *Nature Genetics 34*, 15–21.

[11] Pasmant, E., Vidaud, M., Vidaud, D., and Wolkenstein, P. (2012) Neurofibromatosis type 1: from genotype to phenotype. *Journal of Medical Genetics 49*, 483–489.

[12] Galli, S. J., Borregaard, N., and Wynn, T. A. (2011) Phenotypic and functional plasticity of cells of innate immunity: macrophages, mast cells and neutrophils. *Nature Immunology 12*, 1035–1044.

[13] Park, E., Williams, B., Wold, B. J., and Mortazavi, A. (2012) RNA editing in the human ENCODE RNA-seq data. *Genome Research 22*, 1626–1633.

[14] Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., and Cheung, V. G. (2011) Widespread RNA and DNA sequence differences in the human transcriptome. *Science 333*, 53–58.

[15] Crick, F. (1970) Central dogma of molecular biology. *Nature 227*, 561–563.

[16] Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (2008) The impact of microRNAs on protein output. *Nature 455*, 64–71.

[17] Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F., and Croce, C. M. (2002) Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America 99*, 15524–15529.

[18] Bannister, A. J., and Kouzarides, T. (2011) Regulation of chromatin by histone modifications. *Cell Research 21*, 381–395.

[19] Strachan, T. T., and Read, A. P. *Human Molecular Genetics*, 2nd ed.; Wiley-Liss: New York, 1999.

[20] Li, G.-W., and Xie, X. S. (2011) Central dogma at the single-molecule level in living cells. *Nature 475*, 308–315.

[21] Sturm, R. A., and Frudakis, T. N. (2004) Eye colour: portals into pigmentation genes and ancestry. *Trends in Genetics 20*, 327–332.

[22] Rustgi, A. K. (2007) The genetics of hereditary colon cancer. *Genes & Development 21*, 2525–2538.

[23] Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., and Robson, P. (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell 18*, 675–685.

[24] Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes & Development 16*, 6–21.

[25] Niwa, H., Miyazaki, J., and Smith, A. G. (2000) Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature Genetics 24*, 372–376.

[26] Russo, E. (2005) Follow the money – the politics of embryonic stem cell research. *PLoS Biology 3*, e234.

[27] Kucharczak, J., Simmons, M. J., Fan, Y., and Gélinas, C. (2003) To be, or not to be: NF-κB is the answer - role of Rel/NF-κB in the regulation of apoptosis. *Oncogene 22*, 8961–8982.

[28] Harris, S. L., and Levine, A. J. (2005) The p53 pathway: positive and negative feedback loops. *Oncogene 24*, 2899–2908.

[29] Feng, Z., Hu, W., de Stanchina, E., Teresky, A. K., Jin, S., Lowe, S., and Levine, A. J. (2007) The Regulation of AMPK 1, TSC2, and PTEN expression by p53: stress, cell and tissue specificity, and the role of these gene products in modulating the IGF-1-AKT-mTOR pathways. *Cancer Research 67*, 3043–3053.

[30] Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2011) WikiPathways: building research communities on biological pathways. *Nucleic Acids Research 40*, D1301–D1307.

[31] Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Wolfgang, H., Ukkonen, E., and Brazma, A. (2010) A global map of human gene expression. *Nature Biotechnology 28*, 322–324.

[32] Sood, P., Krek, A., Zavolan, M., Macino, G., and Rajewsky, N. (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proceedings of the National Academy of Sciences of the United States of America 103*, 2746–2751.

[33] Butte, A. J., and Kohane, I. S. (2006) Creation and implications of a phenome-genome network. *Nature Biotechnology 24*, 55–62.

[34] Edgar, R., Domrachev, M., and Lash, A. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research 30*, 207–210.

[35] *What You Need To Know About^{TM} Cancer - National Cancer Institute*, http://www.cancer.gov/cancertopics/wyntk/cancer/page2. `http://www.cancer.gov/cancertopics/wyntk/cancer/page2`.

[36] Hanahan, D., and Weinberg, R. A. (2000) The hallmarks of cancer. *Cell 100*, 57–70.

[37] Hanahan, D., and Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell 144*, 646–674.

[38] Badve, S., and Nakshatri, H. (2012) Breast-cancer stem cells–beyond semantics. *The Lancet Oncology 13*, e43–e48.

[39] Esteller, M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics 8*, 286–298.

[40] Zhang, B., Beeghly-Fadiel, A., Long, J., and Zheng, W. (2011) Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *The Lancet Oncology* 477–488.

[41] Vogelstein, B., Lane, D., and Levine, A. J. (2000) Surfing the p53 network. *Nature 408*, 307.

[42] Atkinson, A. J., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., Spilker, B. A., Woodcock, J., and Zeger, S. L. (2001) Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics 69*, 89–95.

[43] Laterza, O. F., Hendrickson, R. C., and Wagner, J. A. (2007) Molecular biomarkers. *Drug Information Journal 41*, 573–585.

[44] Vasan, R. S. (2006) Biomarkers of cardiovascular disease molecular basis and practical considerations. *Circulation 113*, 2335–2362.

[45] Cohn, J. N. (2004) Introduction to surrogate markers. *Circulation 109*, IV–20.

[46] Duffy, M. J. (2004) Evidence for the clinical use of tumour markers. *Annals of Clinical Biochemistry 41*, 370–377.

[47] Sotiriou, C., and Pusztai, L. (2009) Gene-expression signatures in breast cancer. *New England Journal of Medicine 360*, 790–800.

[48] van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., and Witteveen, A. T. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*, 530–536.

[49] Bueno-de Mesquita, J. M. et al. (2007) Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). *Lancet Oncology 8*, 1079–1087.

[50] Tavassoli, F., and Devilee, P. *Pathology and genetics of tumours of the breast and female genital organs*; World Health Organization, 2003; Vol. 4.

[51] Perou, C. M. et al. (2000) Molecular portraits of human breast tumours. *Nature 406*, 747–752.

[52] Prestegarden, L., Svendsen, A., Wang, J., Sleire, L., Skaftnesmo, K. O., Bjerkvig, R., Yan, T., Askland, L., Persson, A., Sakariassen, P. O., and Enger, P. O. (2010) Glioma cell populations grouped by different cell type markers drive brain tumor growth. *Cancer Research 70*, 4274–4279.

[53] Nagano, K., Yoshida, Y., and Isobe, T. (2008) Cell surface biomarkers of embryonic stem cells. *Proteomics 8*, 4025–4035.

[54] Wang, X., and Dai, J. (2010) Concise review: isoforms of OCT4 contribute to the confusing diversity in stem cell biology. *Stem Cells 28*, 885–893.

[55] Frazer, K. A. et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature 449*, 851–861.

[56] Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G., and Chee, M. S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics 37*, 549–554.

[57] LaFramboise, T. (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research 37*, 4181–4193.

[58] Baker, M. (2011) Sorting out sequencing data. *Nature Methods 8*, 799–803.

[59] Sanger, F., Nicklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America 74*, 5463–5467.

[60] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics 13*, 341.

[61] Bentley, D. R. et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature 456*, 53–59.

[62] VanGuilder, H., Vrana, K., and Freeman, W. (2008) Twenty-five years of quantitative PCR for gene expression analysis. *BioTechniques 44*, 619–626.

[63] Quackenbush, J. (2001) Computational analysis of microarray data. *Nature Reviews Genetics 2*, 418–427.

[64] AC't Hoen, P., Ariyurek, Y., Thygesen, H. H., Vreugdenhil, E., Vossen, R. H. A. M., de Menezes, R. X., Boer, J. M., van Ommen, G. J. B., and den Dunnen, J. T. (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research 36*, e141–e141.

[65] Udvardi, M. K., Czechowski, T., and Scheible, W.-R. (2008) Eleven golden rules of quantitative RT-PCR. *The Plant Cell 20*, 1736–1737.

[66] Honkanen, J., Skarsvik, S., Knip, M., and Vaarala, O. (2008) Poor in vitro induction of FOXP3 and ICOS in type 1 cytokine environment activated T-cells from children with type 1 diabetes. *Diabetes Metabolism Research and Reviews 24*, 635–641.

[67] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America 93*, 10614–10619.

[68] Dai, M. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research 33*, e175–e175.

[69] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science 270*, 467–470.

[70] Patterson, T. A. et al. (2006) Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. *Nature Biotechnology 24*, 1140–1150.

[71] Chen, P., Lepikhova, T., Hu, Y., Monni, O., and Hautaniemi, S. (2011) Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants. *Nucleic Acids Research 39*, e123–e123.

[72] Irizarry, R. A. et al. (2005) Multiple-laboratory comparison of microarray platforms. *Nature Methods 2*, 345–350.

[73] Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics 7*, 55–65.

[74] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research 18*, 1509–1517.

[75] International Human Genome Sequencing Consortium, (2004) Finishing the euchromatic sequence of the human genome. *Nature 431*, 931–945, PMID: 15496913.

[76] Moreau, Y., and Tranchevent, L.-C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics 13*, 523–536.

[77] Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D. (2005) From signatures to models: understanding cancer using microarrays. *Nature Genetics 37*, S38–S45.

[78] Simon, R. (2003) Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer 89*, 1599–1604.

[79] Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996) Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach. *The American Journal of Human Genetics 58*, 1347–1363.

[80] Risch, N. J. (2000) Searching for genetic determinants in the new millennium. *Nature 405*, 847–856.

[81] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007) PLINK: a tool set for whole-denome association and population-based linkage analyses. *The American Journal of Human Genetics 81*, 559–575.

[82] Manolio, T. A. et al. (2009) Finding the missing heritability of complex diseases. *Nature 461*, 747–753.

[83] Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2001) Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics 30*, 97–101.

[84] Houlston, R. S. (2000) Detecting low penetrance genes in cancer: the way ahead. *Journal of Medical Genetics 37*, 161–167.

[85] Browning, B., and Browning, S. (2011) A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics 88*, 173–182.

[86] Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe'er, I. (2008) Whole population, genome-wide mapping of hidden relatedness. *Genome Research 19*, 318–326.

[87] Laakso, M., Tuupanen, S., Karhu, A., Lehtonen, R., Aaltonen, L. A., and Hautaniemi, S. (2007) Computational identification of candidate loci for recessively inherited mutation using high-throughput SNP arrays. *Bioinformatics 23*, 1952–1961.

[88] Wall, J. D., and Pritchard, J. K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics 4*, 587–597.

[89] Hauser, E., Cremer, N., Hein, R., and Deshmukh, H. (2009) Haplotype-based analysis: a summary of GAW16 Group 4 analysis. *Genetic Epidemiology 33*, S24–S28.

[90] Clark, A. G. (2004) The role of haplotypes in candidate gene studies. *Genetic Epidemiology 27*, 321–333.

[91] Laramie, J. M., Wilk, J. B., DeStefano, A. L., and Myers, R. H. (2007) HaploBuild: an algorithm to construct non-contiguous associated haplotypes in family based genetic studies. *Bioinformatics 23*, 2190.

[92] Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P., and Morris, A. P. (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *The American Journal of Human Genetics 75*, 35–43.

[93] Shim, H., Chun, H., Engelman, C., and Payseur, B. Genome-wide association studies using single-nucleotide polymorphisms versus haplotypes: an empirical comparison with data from the North American Rheumatoid Arthritis Consortium. *BMC Proceedings*, 2009; p S35.

[94] Guo, W., and Lin, S. (2009) Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genetic Epidemiology 33*, 308–316.

[95] Abo, R., Knight, S., Wong, J., Cox, A., and Camp, N. J. (2008) hapConstructor: automatic construction and testing of haplotypes in a Monte Carlo framework. *Bioinformatics 24*, 2105.

[96] Purcell, S., Daly, M. J., and Sham, P. C. (2007) WHAP: haplotype-based association analysis. *Bioinformatics 23*, 255.

[97] Li, Y., Sung, W., and Liu, J. (2007) Association mapping via regularized regression analysis of single nucleotide polymorphism haplotypes in variable-sized sliding windows. *The American Journal of Human Genetics 80*, 705–715.

[98] Browning, B. L., and Browning, S. R. (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology 31*, 365–375.

[99] Allen, A., and Satten, G. (2009) Genome-wide association analysis of rheumatoid arthritis data via haplotype sharing. *BMC Proceedings 3*, s30.

[100] Guo, W., Liang, C., and Lin, S. Haplotype association analysis of North American Rheumatoid Arthritis Consortium data using a generalized linear model with regularization. *BMC Proceedings*, 2009; p S32.

[101] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., and Eppig, J. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics 25*, 25.

[102] Eronen, L. M., and Toivonen, H. T. (2012) Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics 13*, 119.

[103] Laakso, M., and Hautaniemi, S. (2010) Integrative platform to translate gene sets to networks. *Bioinformatics 26*, 1802–1803.

[104] Zhong, H., Yang, X., Kaplan, L. M., Molony, C., and Schadt, E. E. (2010) Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *The American Journal of Human Genetics 86*, 581–591.

[105] Ovaska, K. et al. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine 2*, 65–65.

[106] Thomas, D. C. (2005) The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiology Biomarkers & Prevention 14*, 557–559.

[107] Kumar, P., Henikoff, S., and Ng, P. C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols 4*, 1073–1081.

[108] Yue, P., Melamud, E., and Moult, J. (2006) SNPs 3 D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics 7*, 166.

[109] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. *Nature Methods 7*, 248–249.

[110] Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R., and Futreal, P. A. (2010) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research 39*, D945–D950.

[111] Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S. et al.Flicek, P. and Aken, B. L and Ballester, B. and Beal, K. and Bragin, E. and Brent, S. and Chen, Y. and Clapham, P. and Coates, G. and Fairley, S. and others (2010) Ensembl's 10th year. *Nucleic Acids Research 38*, D557–D562.

[112] Maglott, D. (2004) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research 33*, D54–D58.

[113] Apweiler, R. et al. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research 40*, D71–75.

[114] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research 38*, D355–360.

[115] Matthews, L. et al. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research 37*, D619–D622.

[116] Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. (2008) Integrated network analysis platform for protein-protein interactions. *Nature Methods 6*, 75–77.

[117] Cline, M. S., and Karchin, R. (2010) Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics 27*, 441–448.

[118] Karchin, R. (2008) Next generation tools for the annotation of human SNPs. *Briefings in Bioinformatics 10*, 35–52.

[119] Yue, P., Melamud, E., and Moult, J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics 7*, 166.

[120] Goh, W. W., Lee, Y. H., Chung, M., and Wong, L. (2012) How advancement in biological network analysis methods empowers proteomics. *Proteomics 12*, 550–563.

[121] Klingstrom, T., and Plewczynski, D. (2010) Protein-protein interaction and pathway databases, a graphical review. *Briefings in Bioinformatics 12*, 702–713.

[122] Ayers, M. et al. (2004) Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *Journal of Clinical Oncology 22*, 2284–2293.

[123] Basil, C. F., Zhao, Y., Zavaglia, K., Jin, P., Panelli, M. C., Voiculescu, S., Mandruzzato, S., Lee, H. M., Seliger, B., and Freedman, R. S. (2006) Common cancer biomarkers. *Cancer Research 66*, 2953–2961.

[124] Liu, J. J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L., and Ling, X. B. (2005) Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics 21*, 2691–2697.

[125] Kotsiantis, S. B. (2007) Supervised machine learning: A review of classification techniques. *Informatica 31*, 249–268.

[126] Saeys, Y., Inza, I., and Larranaga, P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics 23*, 2507–2517.

[127] Bishop, C. *Pattern recognition and machine learning*; springer New York, 2006; Vol. 4.

[128] Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters 27*, 861–874.

[129] Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*, 378–382.

[130] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009) The WEKA data mining software: an update. *SIGKDD Explorations 11*, 10–18.

[131] Ramensky, V., Bork, P., and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Research 30*, 3894–3900.

[132] McLendon, R. et al. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature 455*, 1061–1068.

[133] Eronen, L., Geerts, F., and Toivonen, H. (2006) HaploRec: efficient and accurate large-scale reconstruction of haplotypes. *BMC Bioinformatics 7*, 542.

[134] Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. (2009) BioMart Central Portal–unified access to biological data. *Nucleic Acids Research 37*, W23–W27.

[135] Shete, S. et al. (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nature Genetics 41*, 899–904.

[136] Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R. L., Bardelli, A., and Bernards, R. (2012) Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature 483*, 100–103.

[137] Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., and Maio, M. (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine 364*, 2507–2516.

[138] Browning, S. R., and Browning, B. L. (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics 12*, 703–714.

[139] Goldin, L. R., Bjorkholm, M., Kristinsson, S. Y., Turesson, I., and Landgren, O. (2009) Highly increased familial risks for specific lymphoma subtypes. *British journal of haematology 146*, 91–94.

[140] Wei, C., Li, J., and Bumgarner, R. (2004) Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics 5*, 87.

[141] Dame, T. M., Orenzoff, B. L., Palmer, L. E., and Furie, M. B. (2007) IFN-gamma alters the response of Borrelia burgdorferi-activated endothelium to favor chronic inflammation. *The Journal of Immunology 178*, 1172–1179.

[142] Adams, R. H., and Alitalo, K. (2007) Molecular regulation of angiogenesis and lymphangiogenesis. *Nature Reviews Molecular Cell Biology 8*, 464–478.

[143] Hirakawa, S., Hong, Y. K., Harvey, N., Schacht, V., Matsuda, K., Libermann, T., and Detmar, M. (2003) Identification of vascular lineage-specific genes by transcriptional profiling of isolated blood vascular and lymphatic endothelial cells. *The American Journal of Pathology 162*, 575–586.

[144] Petrova, T. V., Mäkinen, T., Mäkelä, T. P., Saarela, J., Virtanen, I., Ferrell, R. E., Finegold, D. N., Kerjaschki, D., Ylä-Herttuala, S., and Alitalo, K. (2002) Lymphatic endothelial reprogramming of vascular endothelial cells by the Prox-1 homeobox transcription factor. *EMBO Journal 21*, 4593–4599.

[145] Savilahti, E. M., Rantanen, V., Lin, J. S., Karinen, S., Saarinen, K. M., Goldis, M., Mäkelä, M. J., Hautaniemi, S., Savilahti, E., and Sampson, H. A. (2010) Early recovery from cow's milk allergy is associated with decreasing IgE and increasing IgG4 binding to cow's milk epitopes. *Journal of Allergy and Clinical Immunology 125*, 1315–1321.e9.

[146] Hudis, C. A. (2007) Trastuzumab–mechanism of action and use in clinical practice. *New England Journal of Medicine 357*, 39–51.

[147] Bainbridge, M. N., Wiszniewski, W., Murdock, D. R., Friedman, J., Gonzaga-Jauregui, C., Newsham, I., Reid, J. G., Fink, J. K., Morgan, M. B., Gingras, M.-C., Muzny, D. M., Hoang, L. D.,

Yousaf, S., Lupski, J. R., and Gibbs, R. A. (2011) Whole-genome sequencing for optimized patient management. *Science Translational Medicine 3*, 87re3–87re3.

[148] Roychowdhury, S. et al. (2011) Personalized oncology through integrative high-throughput sequencing: A pilot study. *Science Translational Medicine 3*, 111ra121–111ra121.

[149] Saunders, C. J. et al. (2012) Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Science Translational Medicine 4*, 154ra135–154ra135.

[150] Brand, A. (2009) Integrative genomics, personal-genome tests and personalized healthcare: the future is being built today. *European Journal of Human Genetics 17*, 977–978.

Conflicts of interest

Sirkku Karinen is a chairman of the board, shareholder and one of the founders of a company Significo Research Ltd. (http://www.significo.fi). Significo Research Ltd. provides data analytics and data mining consulting services for life sciences and for business intelligence. Some services provided by Significo Research Ltd. are based on the methods used in this thesis.