Helsinki University Biomedical Dissertations No. 170

Data Integration Methods to Interpret Genome-Scale Data from Cancers

Marko Laakso

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

Finnish Doctoral Programme in Computational Sciences

Finland

Academic dissertation
To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in Auditorium XII (3032), 3rd floor, Main Building, Unioninkatu 34,
on September 28th, at 12 o'clock noon.

Helsinki 2012

Supervised by

Sampsa Hautaniemi, DTech, Docent
Academy Research Fellow
Institute of Biomedicine and Genome-Scale Biology Research Program,
University of Helsinki
Helsinki, Finland

Reviewed by

Tero Aittokallio, Ph.D, Docent
FIMM-EMBL Group Leader
Institute for Molecular Medicine Finland (FIMM), University of Helsinki
Helsinki, Finland

Tapio Visakorpi, MD, Professor
Institute of Biomedical Technlogy, University of Tampere
Tampere, Finland

Official opponent

Hannu Toivonen, Ph.D, Professor
Department of Computer Science, University of Helsinki
Helsinki, Finland

Helsinki University Biomedical Dissertations No. 170
ISSN 1457-8433

ISBN 978-952-10-8177-4 (paperback)
ISBN 978-952-10-8178-1 (PDF)
http://urn.fi/URN:ISBN:978-952-10-8178-1
Helsinki University Print
Helsinki 2012

Contents

Abbreviations iv

Original publications and contributions v

Related publication and contributions vi

Abstract vii

Tiivistelmä viii

1 Introduction 1

2 Review of the literature 3
2.1 Genome-scale measurements . . . 3
2.1.1 Single nucleotide polymorphism microarrays . . . 4
2.1.2 Gene expression microarrays . . . 5
2.1.3 Chromatin immunoprecipitation with microarray . . . 8
2.1.4 Massively parallel sequencing . . . 8
2.2 Biological pathways . . . 10
2.3 Background of cancers studied in Publications . . . 12
2.3.1 Colorectal cancer . . . 12
2.3.2 Glioblastoma multiforme . . . 12
2.3.3 Prostate cancer . . . 13
2.4 Obtaining data from biodatabases . . . 13

3 Aims of the studies 16

4 Materials and methods 17
4.1 Detection of recessive mutations . . . 17
4.2 Data analysis framework . . . 19
4.3 Analysis of AR binding sites . . . 20
4.4 Interpreting new results with the existing information . . . 24

5 Results 29
5.1 Analysis of the CRC genotypes . . . 29
5.2 Candidate pathways . . . 30
5.3 Interplay between AR and FoxA1 . . . 32

6 Discussion 34

7 Acknowledgements 36

References 38

Conflicts of interest 47

Abbreviations

AR	androgen receptor
bp	base pair
BS	binding site
ChIP-seq	chromatin immunoprecipitation with massively parallel DNA sequencing
COSMIC	Catalogue of Somatic Mutations in Cancer (database)
CpG	a DNA methylation site where a cytosine is followed by a guanine
CRC	colorectal cancer
CRPC	castration-resistant prostate cancers
DEG	differentially expressed genes
d	a matrix of distances between genes (g) and peaks (p)
DNA	deoxyribonucleic acid
EM	expectation maximisation
FoxA1	forkhead box protein A1
g	an index for a gene
GBM	glioblastoma multiforme brain cancer
GO	Gene Ontology
GR	glucocorticoid receptor
IBD	identical by descent
IPAVS	Integrated Pathway Resources, Analysis and Visualization System
JASPAR	name of the TF BS consensus sequence database
KEGG	Kyoto Encyclopedia of Genes and Genomes (database)
LOH	loss of heterozygosity
MACS	Model-based Analysis for ChIP-Seq
MEME	Multiple EM for Motif Elicitation
MM	mismatch probe
mRNA	messenger ribonucleic acid
n	number of nucleotide positions in a DNA binding site motif
p	an index for a BS peak
PM	perfect match probe
RNA	ribonucleic acid
$S_{h/l}$	Highest/lowest possible alignment score for the given motif
siRNA	small interfering ribonucleic acid
SNP	single nucleotide polymorphism
TCGA	The Cancer Genome Atlas (data provider)
TF	transcription factor
w	a vector of read alignment overlap enrichments between the case and the control samples at peaks p

Original publications and contributions

This thesis is based on the following original publications that are referred as Publication I–III.

Publication I **Laakso M**, Tuupanen S, Karhu A, Lehtonen R, Aaltonen LA, Hautaniemi S. (2007). Computational identification of candidate loci for recessively inherited mutation using high-throughput SNP arrays. *Bioinformatics*, 23(15):1952–1961.

Publication II Ovaska K, **Laakso M**[*], Haapa-Paananen S[*], Louhimo R, Chen P, Aittomäki V, Valo E, Núñez-Fontarnau J, Rantanen V, Karinen S, Nousiainen K, Lahesmaa-Korpinen A-M, Miettinen M, Saarinen L, Kohonen P, Wu J, Westermarck J, Hautaniemi S. (2010). Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine*, 2(9):65.

Publication III **Laakso M**, Hautaniemi S. (2010). Integrative platform to translate gene sets to networks. *Bioinformatics*, 26:1802–1803.

[*] equal contribution to the work

The author's main contributions to the papers are:

Publication I The development of the algorithm, visualisation and the data analysis, execution of the data analysis, probabilistic characterisation of the algorithm, drafting of the article.

Publication II The functional design of the Anduril framework together with KO, contribution to the development of over 60 Anduril components, analysis of the gene expression and survival data together with EV, revision of the article critically for important intellectual content.

Publication III The design of the database structure and the related accession methods, application of these methods to the analysis of genes with a survival association in glioblastoma, drafting of the article.

Related publication and contributions

The following publication relates to this thesis but has not been included in the official publications. Faculty of Medicine recommends that at most half of the original publications of each thesis are shared in other theses. This unshared publication will be reserved for the first author planning to use it in his thesis.

RelPublication Sahu B, **Laakso M**, Ovaska K, Mirtti T, Lundin J, Rannikko A, Sankila A, Turunen JP, Lundin M, Konsti J, Vesterinen T, Nordling S, Kallioniemi O, Hautaniemi S, Jänne OJ. (2011). Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *The EMBO Journal*, 30(19):3962–3976.

The author's main contributions to the paper are:

RelPublication Development of the data analysis pipeline for the gene expression microarrays and the transcription factor binding sites, revision of the article regarding to the data analysis.

Abstract

The genetic alterations of cancer cells vary between individuals and during the progression of the disease. The advances in measurement techniques have enabled genome-scale profiling of mutations, transcription, and DNA methylation. These methods can be used to address the complexity of the disease but also raise an acute demand for the analysis of the high dimensional data sets produced.

An integrative and scalable computational infrastructure is advantageous in cancer research. First, a multitude of programs and analytic steps are needed when integrating various measurement types. An efficient execution and management of such projects saves time and reduces the probability of mistakes. Second, new information and methods can be utilised with a minor effort of re-executing the workflow. Third, a formal description of the program interfaces and the workflows aids collaboration, testing, and reuse of the work done. Fourth, the number of samples available is often small in comparison with the unknown variables, such as possibly affected genes, of interest. The interpretation of new measurements in the context of existing information may limit the number of false positives when sensitive methods are needed.

We have introduced new computational methods for the data integration and for the management of large and heterogeneous data sets. The suitability of the methods has been demonstrated with four cancer studies covering a wide spectrum of data from population genetics to the details of the transcriptional regulation of proteins, such as androgen receptor and forkhead box protein A1. The repeatable workflows established for these colorectal cancer, glioblastoma, and prostate cancer studies have been used to maintain up-to-date registries of results for follow-up studies.

Tiivistelmä

Syöpäsolujen geneettiset muutokset vaihtelevat potilaittain ja taudin edetessä. Mittausmenetelmien kehittyminen on mahdollistanut mutaatioiden, transkription, sekä DNA-metylaation genominlaajuisen kartoittamisen. Genomin kattavia menetelmiä voidaan käyttää monitekijäisten syöpäsairauksien tutkimuksessa, mutta niiden myötä on syntynyt tarve moniulotteisen tiedon tarkasteluun soveltuville menetelmille.

Joitakin syöpätutkimukseen liittyviä haasteita voidaan ratkaista yhdistävällä ja skaalautuvalla laskennallisella infrastruktuurilla. Ensimmäiseksi, erilaisten mittausten yhdistämiseen tarvitaan useita sovelluksia ja tarkasteluvaiheita. Kokonaisuuden automatisoitu suoritus ja hallinta säästävät aikaa ja pienentävät virheiden mahdollisuutta. Toiseksi, uutta tietoa ja menetelmiä päästään hyödyntämään pienellä vaivalla uudelleen suorittamalla työnkulku. Kolmanneksi, ohjelmistorajapintojen ja työnkulkujen määrämuotoinen kuvaus helpottavat yhteistyötä, testausta ja tehdyn työn uudelleenkäyttöä. Neljänneksi, saatavilla olevien näytteiden lukumäärä on usein pieni verrattuna kiinnostuksen kohteena oleviin tuntemattomiin muuttujiin, kuten mahdollisesti vioittuneisiin geeneihin. Uusien mittausten tulkinta olemassa olevan tiedon yhteydessä saattaa vähentää väärien positiivisten määrää kun tarvitaan herkkiä menetelmiä.

Olemme esitelleet uusia laskennallisia menetelmiä tiedon yhdistelyyn, sekä laajojen ja vaihtelevan muotoisten aineistojen käsittelyyn. Menetelmien käyttökelpoisuutta olemme havainnollistaneet soveltamalla niitä neljässä syöpätutkimuksessa, jotka liittyvät paksunsuolen syöpään, glioblastoomaan ja eturauhassyöpään. Tutkimusten aihealueet kattavat kirjon populaatiogenetiikasta transkriptiotekijöiden, kuten androgeenireseptorin ja FoxA1:n toiminnan, yksityiskohtiin. Tutkimusten puitteissa toistettavaan muotoon rakennetut työnkulut ovat tuloksineen tarjonneet ajantasaisen tietolähteen pohjaksi jatkotutkimuksille.

1 Introduction

Cancer is characterised by the cells that have lost their growth controlling ability as a consequence of a set of genetic or epigenetic defects accumulated to their genome (Vogelstein and Kinzler, 2004; Martini et al., 2011). The pattern of defects varies between and within individual tumours and during their progression (Dancey et al., 2012; Gerlinger et al., 2012). The malignant growth and the invasive nature of the cancer cells may disturb the normal body function. Indeed, 7.6 million annual cancer caused deaths are reported world wide (WHO, 2011). In the following text, the focus is on colorectal cancer, glioblastoma, and prostate cancer, which have been studied in Publications I–III and in RelPublication.

Molecular measurement techniques have evolved rapidly over the last decades enabling genome-scale analysis of the DNA sequences and gene expressions (Chen et al., 2012). The advanced understanding of human genome and genetic variation between individuals has improved the resolution of the assays as more specific probes can be produced and denser maps of markers have become available (Frazer et al., 2007; Van der Ploeg, 2009; Levsky and Singer, 2003). At the same time, throughput of measurements has increased due a higher level of parallelisation and multiplexing. The invention of microarrays and the development of new stainings and sequence labels have enabled simultaneous observation of many biological variables (Hoheisel, 2006; Levsky and Singer, 2003). The computational analysis of the produced data sets, especially the integration between platforms and experiments, has become a challenge (Wilkes et al., 2007). Although new methods have been introduced for the sample collection and preparation, the reproducibility and the accuracy of the biomedical measurements can be improved still (Wilkes et al., 2007; Van der Ploeg, 2009).

The number of biological databases and biomedical articles is increasing and they provide a rich source of information that may facilitate functional and causal interpretation of the observations. The dimensions of data produced in genome-scale studies are challenging for the traditional statistics. The number of unknown variables (genes, genomic loci) is vast compared to the number of cases (samples) (Houlston and Peto, 2004; Easton et al., 2007). Existing information regarding the variables can be used to compensate the related uncertainty and some of the challenges can be solved by using biological databases during the data analysis. An automated version of such analysis enables rapid adjustments of the results as new information becomes available.

External data sources can be useful when selecting candidates for the validation, but an integrative system that covers them may become fragile and hard to maintain. Each external resource adds new complexity to the system as their interfaces and content tend to change in time. The latest information cannot be captured without revising the resources periodically, which means that one has to maintain compatibility between the systems. Another technical challenge lays in the heterogeneity of the biomedical resources. For instance, different assays have been used to measure patient genotype, somatic mutations in tumours, gene expression, and DNA methylation. Related knowledge is represented in terms

of frequent alterations in certain tumours, models describing the interactions between some bioentities, and functional annotations assigned to the genomic loci. Although new standards have been established for biomedical data (Turenne, 2011), the representation and the accession method of all this information typically varies according to the providing source.

A computational infrastructure can help in repeating and maintaining analyses that are comprised of steps implemented in various computer programs (Almeida, 2010; Evans, 2011; Podpečan et al., 2011). *Workflow engines* are software frameworks specialised in the management and execution of computational processes consisting of tasks and the dependencies between them. A workflow engine based computational infrastructure can be used for the simultaneous analysis of multiple data sets of various kinds. We demonstrated the advantages of such an approach by analysing survival associations in The Cancer Genome Atlas (TCGA) glioblastoma multiforme (GBM) data set. TCGA is established by the National Cancer Institute and the National Human Genome Research Institute (at the United States of America) in order to improve the molecular understanding of cancer by providing a shared repository of data. GBM data set consisted of samples of 338 patients and it was among the widest genome-wide cancer data sets available in 2009 (McLendon et al., 2008).

An efficient analysis of the data sets is a prelude to the functional and causal interpretation of the genome-scale data. The existing pipelines of case-control studies produce distribution profiles and literature annotations of individual entities over the sample sets. The results are typically summarised in terms of pathway impacts (Tarca et al., 2009) and enrichments of certain annotations among them (Subramanian et al., 2005; Ovaska et al., 2008). New methods are needed for the automated pathway integration that could address the crosstalk between the pathways (Bauer-Mehren et al., 2009). We directed our efforts to compare pathways and result sets by establishing a database called Moksiskaan that combines information about relationships between entities such as genes, proteins, diseases, drugs, pathways, cellular components, and biological functions. This database enabled a construction of connectivity graphs between the entities of interest, which aid the interpretation (Liikanen et al., 2011; Heinonen et al., 2011; Louhimo et al., 2012).

The extraction of the relevant set of relationships is one of the key challenges when a large pool of heterogeneous interaction data is used in a specific biological context. Functional dependencies between the biological entities can be learned by measuring biological cascades at various levels (Pe'er and Hacohen, 2011). For instance, genome-wide profiles of the transcription factor (TF) binding sites (BS) can be integrated with the cellular responses in order to identify the functional targets. The pre-existing information about the DNA binding motifs of other TFs (Portales-Casamar et al., 2010) is also useful in this context, as we demonstrated in the study of androgen receptor (AR) in prostate cancer (RelPublication).

2 Review of the literature

This section provides a summary and the key references of the published knowledge regarding the production and analysis of genome-wide measurements of glioblastoma (Publication II; Publication III), colorectal (Publication I), and prostate cancers (RelPublication).

2.1 Genome-scale measurements

The release of the human reference genome enabled studies that relied on the genomic loci of the sequence fragments (Lander et al., 2001). The reference is also being used to estimate how many times certain fragments occur at the genome and to identify unique fragments (Gräf et al., 2007). The reference itself does not represent a real genome but it has been constructed by combining samples of different individuals (Lander et al., 2001). More importantly, a fixed reference is not able to capture the variations seen between individuals, not to mention the defects seen in cancer cells.

Still, once fixed the reference provided the means by which these variations can be described and, indeed, massive projects have been established for this purpose. HapMap (Frazer et al., 2007) collects data about the variants and their segregation in humans. Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes et al., 2008) is a cancer project that focuses on somatic variants present in tumours. The joint information about the reference and the common variants has been used to discover short but still unique sequence fragments of the genome. A variety of high-throughput measurement techniques have emerged from the opportunity to first measure these sequences from the sample material and then to map these signals to their positions in the genome.

Genome-scale studies of DNA and its transcriptional regulation are based on the measurements aiming for a complete or at least unrestricted set of targets. An important aspect of these studies is that the number of null hypotheses is typically high. The same hypothesis (such as 'the expression of this gene does not differ between the case and control samples') may be applied to tens of thousands of genes. Large sample sets may be needed in order to produce reliable results while keeping the number of false positives low. Multiple hypothesis corrections, such as Bonferroni correction and false discovery rate (FDR) (Benjamini and Hochberg, 1995), can be used to estimate overall reliability of the results.

The total length of a human genome is over $3.1 \cdot 10^9$ base pairs (bp) (Flicek et al., 2011), which enables an enormous potential for variation. HapMap describes over $3.1 \cdot 10^6$ single nucleotide polymorphisms (SNP) that are estimated to cover 25%–35% of the SNP variations among the studied four populations of African, Asian, and European ancestry. The genome-scale studies focusing on phenotype-genotype relationships are typically faced with challenges of low numbers of samples versus the dimensions and the variability of the genome. Not all loci are equally susceptible for alterations (Martini et al., 2011).

2.1.1 Single nucleotide polymorphism microarrays

The contemporary SNP-microarrays enable simultaneous genotyping of hundreds of thousands of SNPs (Kathiresan et al., 2009; LaFramboise, 2009). The exact SNPs and the expected alleles are defined in advance during the manufacturing of the microarray, which is in contrast to the less biased sequencing techniques discussed in Section 2.1.4. SNP-microarrays can be used to measure genotypes of hundreds of thousands up to a million biallelic SNPs (LaFramboise, 2009). The SNPs are distributed non-uniformly but relatively densely across the genome (Madsen et al., 2007).

SNP-microarrays, like other complementary DNA hybridisation microarrays, are based on short single stranded DNA fragments called *probes* that are fixed on the solid surface of the array. Probes have been designed so that they represent the complement strand for the sequence around the target sequence and are unique to that. The selection of unique probes has biased commercial arrays towards SNPs on non-coding sequences, which limits their capability to detect gene variants (Nicolae et al., 2006). The information content of the probes depends on how independently their alleles segregate in respect to the neighbouring alleles. Nicolae et al. (2006) demonstrated that about 40% of the SNPs selected for the commercial Affymetrix array were highly dependent on their neighbour, thus the practical resolution of the array is not necessarily as high as assumed based on the number of SNPs covered. The chosen probes are clustered to distinguishable spots consisting of clones of an identical sequence and each of these spots represents an allele of a SNP. Typical commercial arrays support biallelic SNPs, which means that they provide at least two probe sequences for each SNP, both having a different nucleotide at the site of the SNP.

In Affymetrix GeneChip® Human Mapping 100K Set, used in Publication I, there are 40 different 25bp long probe sequences for each SNPs. Half of the probes are so called mismatch probes (MM) having a middle nucleotide which does not match either of the two alleles in the reference genome. The purpose of these probes is to reflect the level of non-specific cross hybridisation, but their necessity has been of debate, especially in the context of gene expression microarrays as will be discussed in Section 2.1.2 (LaFramboise, 2009). The other half of the probes, the perfect match probes (PM), represent exact alignments for the allele specific sequences but vary in their positioning in respect to the SNP. Figure 1 illustrates a hypothetical quartet of probes and a perfect alignment of one allele.

Genotypes of the sample DNA are predicted based on its hybridisation with the microarray probes. Before the hybridisation, sample DNA is enzymatically fragmented to match the array probes. The digestion enzymes are sequence specific and thus the SNPs have to be selected so that the enzymes can produce suitable fragments for the probes (Mao et al., 2007). Next, the fragments are labelled with a fluorescent marker so that they can be detected on the microarray once hybridised. The labelled fragments are placed on the microarray where they can bind with the probes, and the excess material is washed away. The results of the hybridisation are read by scanning the microarray with a laser which detects

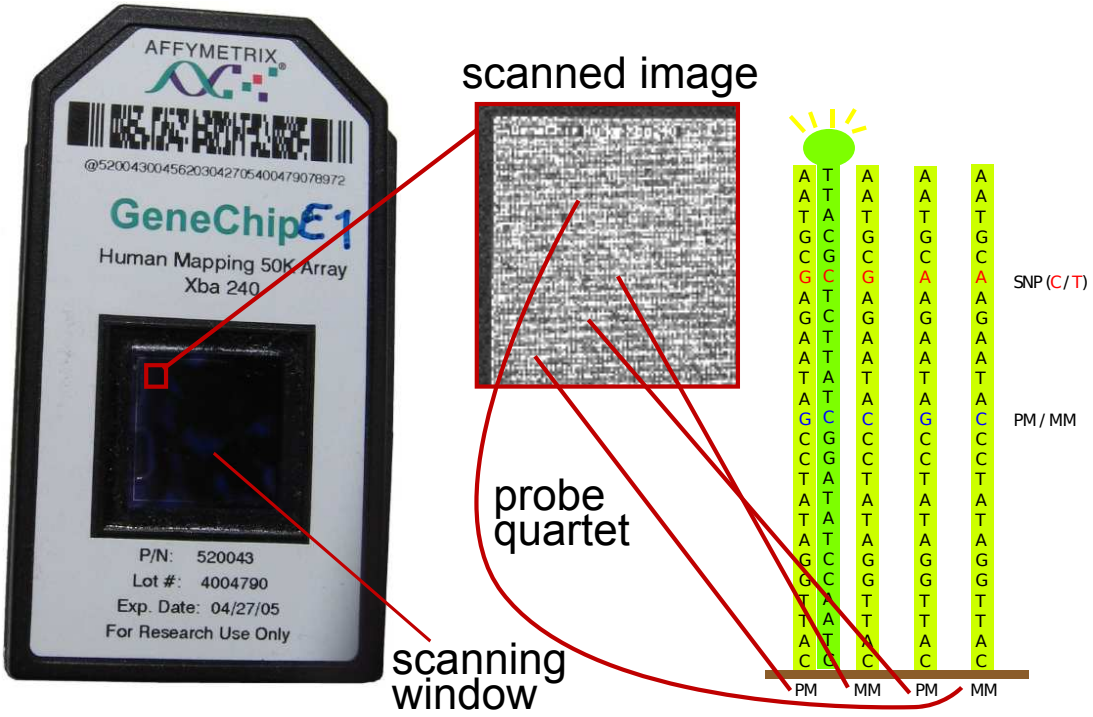

Figure 1: A fluorescent labelled sample oligonucleotide has hybridised with the PM of the C allele, which belongs to a quartet designed for a C/T SNP. Each probe belongs to a spot consisting of its clones. The intensities of these spots are determined from the scanned image of the array. The image has been edited from Laakso (2007).

the fluorescent marker. The scanning produces a figure of the array surface, which is then analysed computationally. Spots of the probe clusters are detected and the signal intensity of each cluster is estimated for the included pixels.

Probe cluster intensities are compared between the clusters representing different alleles of the same SNP, and the relative intensities are interpreted as genotypes. A heterozygous sample containing one copy of alleles A and a produces a balanced signal for the corresponding spots. A homozygous sample (AA) produces a higher signal for A, whereas the signal of a remains at the level of the background produced by the cross hybridisation between the partially matching probe and sample DNA. The signal intensities, especially when analysed together with the neighbouring SNPs, can be used to estimate local amplifications of the genome (corresponding genotypes such as AAa and aaaa) (LaFramboise, 2009).

2.1.2 Gene expression microarrays

Gene expression microarrays are used to measure messenger ribonucleic acid (mRNA) levels of the cells. The technology of these microarrays resembles that of the SNP-microarrays, except that the probe sequences have been selected from the mRNA sequences of the genes. Several short probe sequences are typically used for each targeted gene. The sample preparation is also different since the

original RNA is first converted to DNA using a reverse transcriptase before it is labelled and hybridised.

Quantitative analysis of the gene expression is more sensitive for the preprocessing of the image data than the discrete classification of the genotype calls. The intensity scale of an individual microarray is a result of the sample concentration, volume and the hybridisation conditions providing relative information about the probes within the array (Quackenbush et al., 2001). Normalisations are used to adjust these scales for a better comparability between the arrays by reducing the effects of non-biological origin (Bolstad et al., 2003).

In RelPublication, we have used quantile normalisation that assumes a common overall distribution of probe intensities for each array, although the signals of individual genes may vary (Bolstad et al., 2003; LaFramboise, 2009). The values of each sample are sorted independently, and then the value at each sorted position is replaced with the mean of the values at the same position on all arrays. Figure 2 illustrates the effects of the normalisation in RelPublication. Each normalised sample provides the same set of values and thus the distributions are equal although the values may belong to different genes. The improved comparability of the normalised values can be seen in the average distance based hierarchical clustering that makes a clear distinction between the corresponding replicates and the other samples. The gene expression values are formed by combining the values of the probe spots related to the gene. In RelPublication, a median was used to combine these values.

Some probe and sample fragments are hybridised together although their sequences are not fully complement to each other. The MM of the Affymetrix GeneChip® microarrays are intended for the background correction of this non-specific hybridisation that contributes to the signals of PM. Alternatively, the background signal can be estimated from the PM assuming the observed signals are sums of exponentially distributed signals of the fully compatible sequences and a normally distributed (with a positive truncation) background signal. Robust multi-array average is a microarray normalisation method that combines the PM based background correction with the quantile normalisation (Irizarry et al., 2003). This method has been used for the pre-normalised glioblastoma data we obtained from TCGA (Publication II; Publication III).

The general steps of hybridisation, image processing, quality control, normalisation, and the expression comparisons between the samples are common in gene expression microarray studies, but their actual forms vary between the projects. As Quackenbush et al. (2001) stated, a common approach may never fit for all applications and aspects of the research. For this reason, flexible computational infrastructures are important as they aid the quick construction of suitable combinations of the methods and the evaluation of different approaches.

Exon arrays are high density gene expression microarrays, which provide probes for each exon of the gene. The exon specific signals can be used to infer the relative abundances of the splice variants of the same gene (Gardina et al., 2006; Chen et al., 2011).

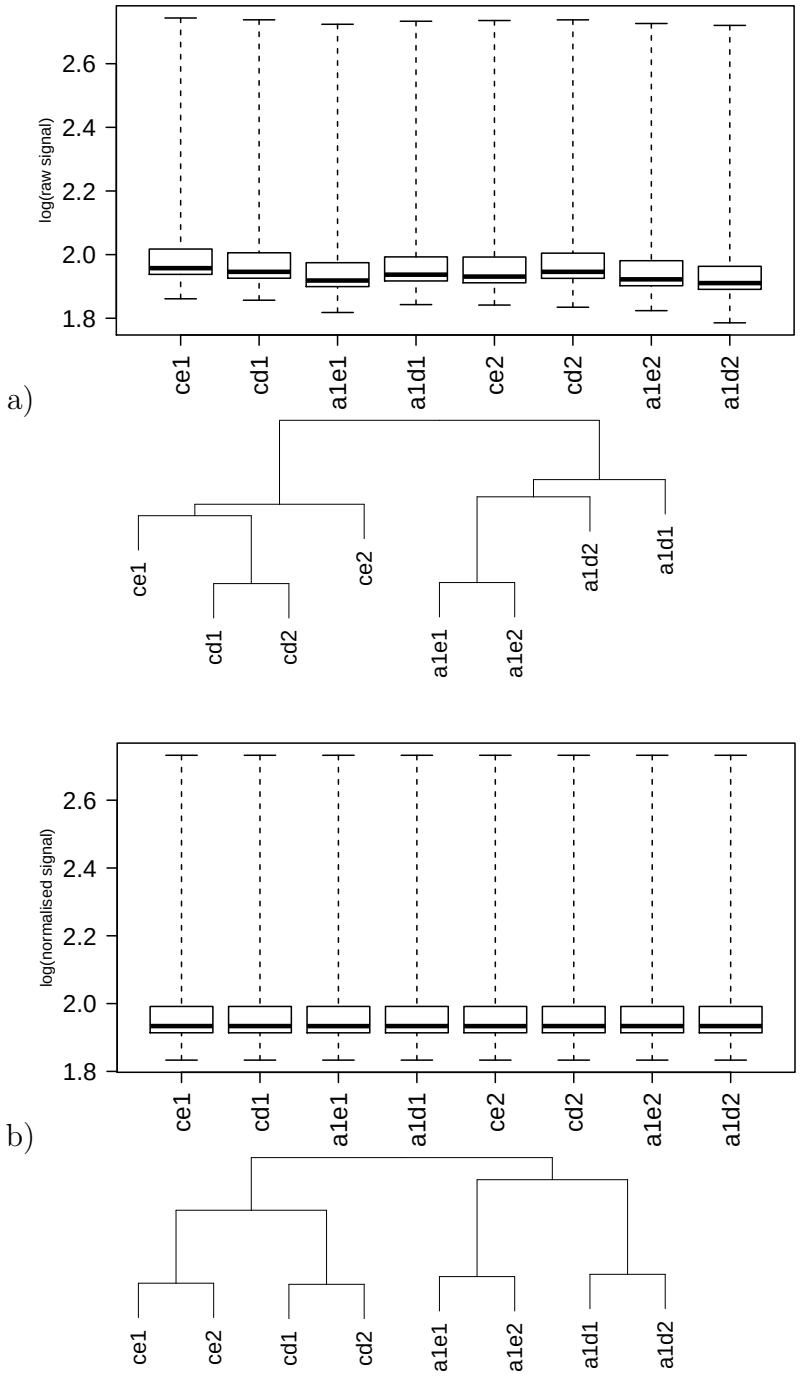

Figure 2: Signal distributions and the hierarchical clusterings of the gene expression arrays before (a) and after (b) the quantile normalisation. These Illumina HumanHT-12 v3 Expression BeadChip Kit microarrays were used to measure 5α-Dihydrotestosterone responses in parental (c) and FoxA1 depleted (a1) LNCaP cells. Control samples have been labelled with e, and d stands for the treatment. Two replicates (1, 2) have been used for each condition.

2.1.3 Chromatin immunoprecipitation with microarray

Chromatin immunoprecipitation is an antibody based targeted DNA isolation procedure that is based on the recognition of the DNA attached proteins. The procedure, when coupled with the determination of the sequences of the isolated DNA fragments, can be used to estimate the binding sites of the protein of particular interest. Depending on the antibody, the same procedure can be used in the studies of transcription factor binding sites, histone modifications, and RNA polymerase activities. The protocol consists of four steps (Aparicio et al., 2004):

1. Formaldehyde, for example, is applied to cross-link proteins with the DNA in its proximity.

2. DNA is fragmented to short sequences in order to enhance the binding site specificity and to lower the molecular size of the protein-DNA complexes.

3. An antibody is used to selectively isolate DNA fragments attached to the protein of interest. This step is called *immunoprecipitation*.

4. The proteins are detached from the DNA.

The sequence of the purified DNA fragments may be determined in various ways. One option, called *ChIP-chip*, is to use microarrays for this purpose (Wu et al., 2006). Special microarrays (*DNA tiling arrays*) have been developed for this purpose. These arrays have a high density coverage of the unique sequences of the genome, and the spot intensities of the hybridized arrays can be interpreted as quantitative measures of the corresponding fragments in the sample material. The known sequences of the probes, when aligned against the reference genome, are used to predict the chromosomal loci of the protein binding sites. A typical binding site is observed as an increased intensity of the probes, which correspond to near-by sites in the chromosome. The intensities of the probes are higher near the exact binding site and decrease towards more distant probes as the corresponding DNA is more likely cut off from the protein attached part during the fragmentation.

2.1.4 Massively parallel sequencing

DNA sequencing techniques are evolving quickly, and the conventional complementary DNA hybridisation microarrays may be often replaced with massively parallel sequencing assays (Marioni et al., 2008; Park, 2009). An excessive amount of raw data is typically produced by a sequencing experiment, hence the information is more detailed (the composition of the DNA fragments in addition to their relative concentration) than with the microarrays. Consequently, an efficient computational infrastructure is needed for the processing and storing purposes.

There is a great interest in using massively parallel DNA sequencing for the genotyping purposes (Nielsen et al., 2011; Davey et al., 2011). The advantage of the sequencing techniques is that they are less dependent on the probe design and

may thus better capture unexpected alleles and variations (Wheeler et al., 2008). Depending on the DNA preparation protocol, the same data can be used to infer much more information than the SNPs. For instance, a complete sequence may reveal chromosomal rearrangements and other mutations (Pareek et al., 2011). The sequencing methods are prone to errors, and thus a high sequencing coverage is needed at the variation site in order to obtain enough reads from the two possible alleles and to distinguish between them and the errors (Nielsen et al., 2011). Instead of a complete genome, a targeted sequencing of the exons may be used. By focusing on exons, a higher coverage or a larger set of samples can be analysed at the price of the complete genomes. The exon sequences have already been used to identify recessive mutations as variants present in both homologous chromosomes (Bilgüvar et al., 2010). Importantly, fewer samples are needed if the variants are determined at the individual level, instead of comparing the case and control frequencies of the putative chromosome regions found from the SNP data.

Massively parallel sequencing of the mRNA provides an alternative to the gene expression microarrays (Marioni et al., 2008). Analogous to the microarrays, the RNA is first converted to complementary DNA before it is sequenced. The sequence data can be used for the detection of alternative splicing as more sequence fragments are obtained from the more abundant exons. In addition, sequencing data provides sequence fragments that may overlap the exon boundaries, providing direct evidence of the splicing. Sequence overlaps that match partially with two different genes may be used to detect fusion genes caused by chromosome translocations (Ozsolak and Milos, 2010). The mRNA sequencing can be performed with a small amount of sample material, which enables transcriptional analysis of individual cells (Tang et al., 2009; Ozsolak and Milos, 2010).

Chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-seq) is an assay similar to ChIP-chip, except that the tiling microarrays are replaced with sequencing of the immunoprecipitated DNA fragments. Sequence data is converted to binding site coordinates by aligning sequence fragments against the reference genome. Bindings are predicted at those loci that have a high local enrichment of overlapping alignments. In practise, the current sequencing techniques can only measure the ends of the precipitated DNA fragments and this limits the resolution of the binding site boundaries (Zhang et al., 2008). Technical details of the data analysis are further discussed in Section 4.3.

ChIP-seq provides typically a better overlap between sample replicates than ChIP-chip, and the spatial resolution of the binding site predictions is enhanced as characterised in Park (2009) and Ho et al. (2011). We used tiling arrays in the beginning of the androgen receptor study, but the approach was replaced with Illumina's Solexa sequencer due to higher resolution and improved data quality of the ChIP-seq technology.

2.2 Biological pathways

The evolutionary selection has led to organisms with complex chemistry that can be adjusted based on the environmental factors. The relationships between cellular molecules and genes can be represented as graphs (Papin et al., 2005; Aittokallio and Schwikowski, 2006; Pisabarro et al., 2008; Jensen et al., 2009; Pe'er and Hacohen, 2011; Yosef et al., 2011). Genes, proteins, and other compounds are vertices and their relationships are represented with edges. The topologies of these graphs can be used as a basis of the mathematical models explaining the cell function (Papin et al., 2005; Bauer-Mehren et al., 2009).

The relationships between the molecules representing a certain condition, such as prostate cancer or response to a certain stimulus, such as hormones (Figure 3), are often referred to as *canonical pathways*. The canonical pathways are not exclusive but they share common entities and connectivities. Importantly, these pathway descriptions are not complete. They are used to describe the key elements and their relations, although these elements are not in isolation (Ma'ayan et al., 2005). The reduction, although helpful in the described context, produces additional challenges when the interest is in the crosstalk between the canonical pathways. We use the term *fusion pathway* for these hypothetical models that represent relationships combined from various canonical pathways. These models can be used to represent regulation between cellular functions, which has been one of the challenges in systems biology (Ma'ayan et al., 2005).

The canonical pathways representing chemical reactions of small molecules are often referred as *metabolic pathways*. Proteins associated to these pathways are typically enzymes catalysing these reactions, or their co-factors (Schilling et al., 1999; Ouzounis and Karp, 2000). The connections between the proteins related to subsequent reactions are mediated by reactants, which are often small molecules with a variety of possible targets (Croft et al., 2011; Kanehisa et al., 2012).

Signal transduction pathways are canonical pathways representing chains of reactions related to an extracellular stimuli (Jørgensen and Linding, 2010). These pathways typically represent regulation of certain biological functions, such as cell cycle adjustment or cell death. In contrast to the metabolic pathways, signal transduction cascades typically involve protein modifications such as phosphorylations and ubiquination.

Cancer is famous of its ability to activate cellular processes typically inactive in its cell type lineage (Cui et al., 2007). Typical examples include the activation of telomerase, increased motility, and epithelial-mesenchymal transition that makes epithelial cells more resistant to apoptosis and promotes their invasiveness (Hanahan and Weinberg, 2011). There is a number of ways cancer cells can change their behaviour by activating or inactivating the cell's biological processes (Vandin et al., 2012). The representation of these mechanisms (such as mutations) at the level of pathways can be used to reduce the complexity and to better identify those defects that are responsible of the pathogenesis (Vandin et al., 2012).

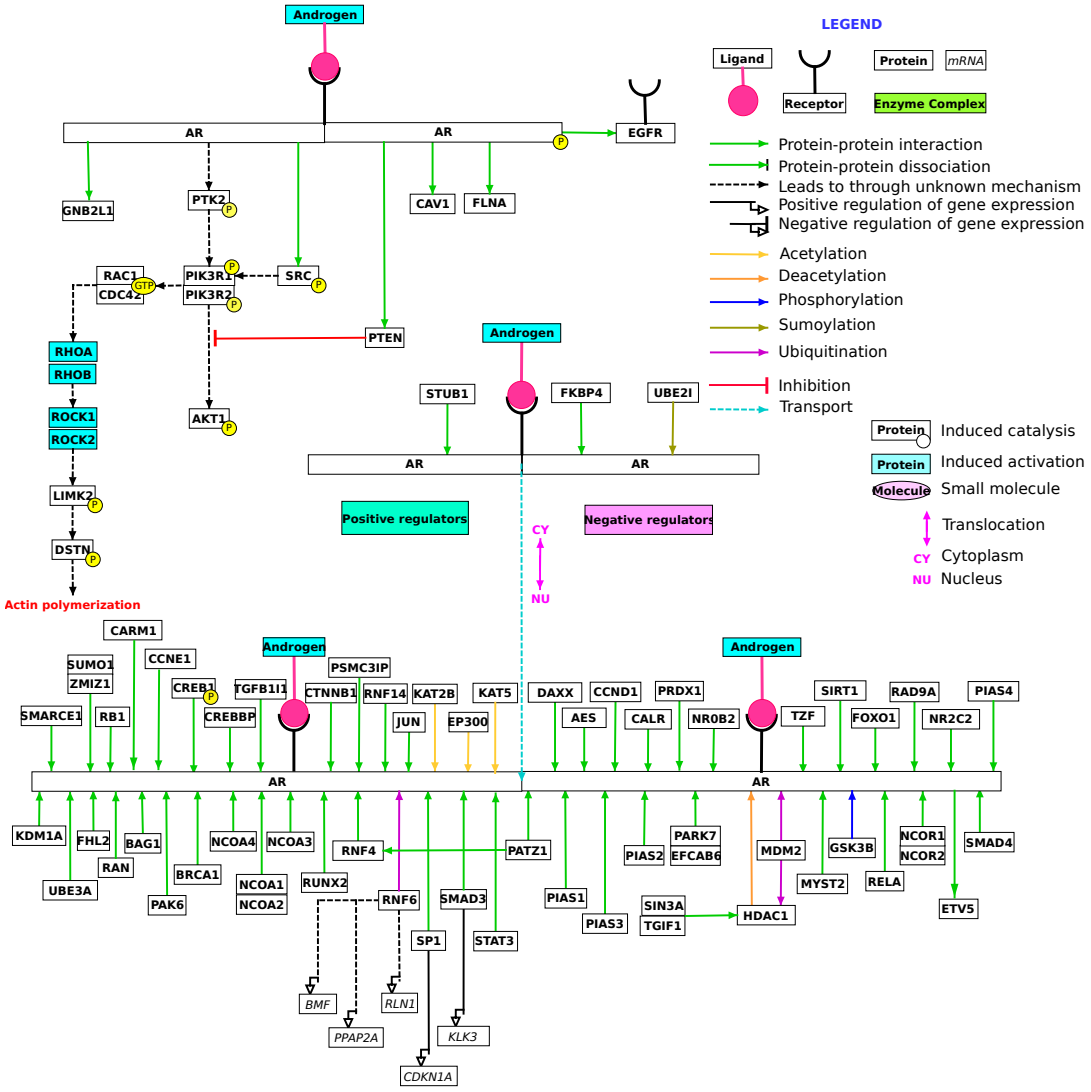

Figure 3: This canonical pathway represents androgen receptor signalling as described in WikiPathways (Hanspers et al., 2011). The lower part of the figure illustrates the interactions between the AR and other proteins modulating its activity. The promoting factors are on the left and the inhibitors on the right.

2.3 Background of cancers studied in Publications

Cancer is a disease of malignant cell growth caused by a heterogeneous combination of defects in the genome (Vogelstein and Kinzler, 2004; Gerlinger et al., 2012). The disease involves interactions between various kinds of cells within and around the tumour tissue (Hanahan and Weinberg, 2011; Clevers, 2011). Cancer is among the few diseases caused by somatic mutations (Vogelstein and Kinzler, 2004). The parental organ and the tissue from which the primary tumour has arisen has been the major determinants of the diagnosis, but remaining variation within these cancer classes causes a demand of more specific diagnostics and treatments (Fritz, 2000; Martini et al., 2011; Gerlinger et al., 2012).

Here, we demonstrate new methodologies to analyse genome-scale data in three different cancers. The biological heterogeneity of these cancers demonstrates the general properties of the proposed computational approaches and their possible applicability to a wider set of research.

2.3.1 Colorectal cancer

Colorectal cancer (CRC) covers the tumours of the colon and rectum (Muzny et al., 2012). CRC is the third most common cancer in males and the second most common cancer in females in Finland, with an annual incidence of $1415_{\male} + 1321_{\female}$ (Finnish Cancer Registry, 2012). Environmental factors are the major cause of CRCs but the individuals' heredity contributes 20%–30% to the risk of incidence (Lichtenstein et al., 2000; de la Chapelle, 2004). Some of the high-penetrance mutations are known to affect genes like *APC*, *MSH2*, *MSH6*, *MLH1*, *PMS2*, *AXIN2*, *POLD1*, *TGFBR2*, *SMAD4*, *BMPR1A* and *MUTYH*. All these examples, except the last one, have a dominant manifestation, which means that even one mutated allele is enough to cause a genetic predisposition. The mutations in *MUTYH* are typically less penetrative unless both alleles are affected (Papadopoulos and Lindblom, 1997).

The detection of new recessive mutations is an interesting challenge because the number of phenotype positive cases is limited and a mutation is likely to explain only a fraction of them. The conventional statistical tests are not sensitive enough for this purpose and thus we developed a combined approach of a rule based detection (sensitivity) of the putative sites and the data integration (specificity) based pruning of the results (Publication I).

2.3.2 Glioblastoma multiforme

Glioblastoma multiforme is a cancer that originates from the glial cells of the brain (Riemenschneider et al., 2010) and corresponds to the grade IV astrocytoma (Louis et al., 2007). Glioblastoma is typically of somatic origin. The diffuse growth within the brain tissue prevents a complete surgical removal of the tumour, and current therapies, such as radiation and chemotherapies, produce only modest responses. The median survival of the treated patients is roughly a year after

their initial diagnosis (Weller et al., 2009). Metastases on distant organs are rare and the deaths are typically caused by the primary tumour.

We focused on GBM in Publication II and Publication III because of the substantial need of new treatments and because TCGA provided its most comprehensive data set for this cancer. The data consisted of measurements of chromosome copy numbers, SNP genotypes, DNA methylation, expression profiles at the level of genes, individual exons, and microRNAs. In addition, clinical data (age, sex, time of diagnosis, time of death, treatment, etc.) was provided for the patients in the data set. All these data were used in Publication II. Publication III is based on the expression and clinical data.

2.3.3 Prostate cancer

Prostate is particularly susceptible to cancer and prostate cancer is the most common cancer in Finnish males, covering 40% of all cases (Finnish Cancer Registry, 2012). The cancer typically develops at old age and grows slowly without noticeable symptoms. Consequently, many cases are either not diagnosed or treated aggressively enough. An active surveillance may be enough for those diagnosed with a low-risk tumour. For the others, surgery, radiation, and chemotherapies are the most common options.

Œstrogen receptor like steroid receptors such as androgen receptor (AR) and glucocorticoid receptor (GR) are nuclear receptors, which, once activated, translocate from the cytosol to the nucleus and bind to the DNA (Aranda and Pascual, 2001). Once bound to the DNA, they act as transcription factors of their target genes (Aranda and Pascual, 2001). Various binding motifs of AR have been reported (Heemers and Tindall, 2007) but little is known about the exact DNA targets and the co-factors that are involved in the transcriptional regulation.

Prostate cancers are typically divided to those responding to deduction of the AR activating ligands androgen/testosterone/5α-Dihydrotestosterone and to the castration-resistant prostate cancers (CRPC), that have become refractory to the aforementioned treatments. CRPC cancer may still rely on AR mediated transcription, although it is independent of the external activators and thus the understanding of the functions of AR and its co-factors is important for these cancers as well.

RelPublication represents a hormone dependent cancer with a particular focus on the transcriptional regulation and the signalling cascades of the hormone. The publication describes how the AR response differs in the cells in the presence and in the absence of FoxA1 co-factor.

2.4 Obtaining data from biodatabases

Biomedical information has been accumulated into scientific articles over the years, but utilisation of these resources becomes infeasible when dealing with thousands of genes. For instance, PubMed returns 5738 articles about the function

of *TP53* gene and 61889 articles for the corresponding p53 protein [29.3.2012]. The invocation of these articles alone would be an overwhelming task for any human. The number of human genes (> 40000[1]) is comparable to the number of words people learn for a second language (Schmitt, 2008), not to mention that the nomenclature of them varies in context and time. Fortunately, automated tools may help with this complexity, and much of the information has been stored into various databases. The variety of these databases (Galperin and Fernández-Suárez, 2012) itself is a notable issue, but here we will focus on some of the biggest and the most relevant databases in the context of cancer research. Some databases are used to store and share original observations and the latest knowledge, whereas others act as proxies integrating their content. The relationships between the databases are often complex and the distinction between the primary sources and the proxies are often mixed. Same biological entities have been labelled with different identifiers in different databases, thus the mapping between the namespaces is an integral part of the practical bioinformatics (Huang et al., 2008).

The human reference genome forms the basis of the genome databases such as Ensembl, NCBI Entrez (Maglott et al., 2007) and UCSC Genome Browser (Rhead et al., 2010). These databases focus on the functional and structural annotations of the DNA regions. The databases can be used to fetch information about the transcripts, CpG-islands, centromeres, transcription factor binding sites, and SNPs.

Information about biochemical relationships between different compounds is gathered in various databases (Bader et al., 2006; Jensen et al., 2009). The term *pathway database* is often applied to these databases, especially when the information has been structured into canonical pathways, whereas it is used less in the context of less organised collections of molecular interactions. The classical *protein-protein interaction databases*, such as MINT (Ceol et al., 2010), IntAct (Kerrien et al., 2007), DIP (Salwinski et al., 2004), BioGRID (Stark et al., 2011), HPRD (Prasad et al., 2009), STRING (Jensen et al., 2009) and PINA (Wu et al., 2008), exemplify the latter case by focusing on the physical interactions between the proteins and protein complexes. On the other extreme, in the databases such as KEGG (Kanehisa et al., 2012) and WikiPathways (Kelder et al., 2012) where the canonical pathways are used to organise the data, the individual reactions are described in the context of these biological models. Pathway Commons (Cerami et al., 2010) provides a generic proxy for the public databases with a BioPAX (Demir et al., 2010; Strömbäck and Lambrix, 2005) interface. The current [8.3.2012] version covers databases such as Reactome (Croft et al., 2011) and HumanCyc (Romero et al., 2004), which provide metabolic pathways. Integrated Pathway Resources, Analysis and Visualization System (IPAVS) provides curated pathway information combined with the information automatically imported from other pathway databases (Sreenivasaiah et al., 2012). IPAVS web application enables direct and manual utilisation of the system. Larger

[1]Ensembl (Flicek et al., 2011) version 66.37 describes: 20563 known protein-coding genes, 536 novel protein-coding genes, 15520 pseudogenes, 11960 RNA genes, and 637 immunoglobulin/T-cell receptor gene segments.

and more automated approaches can be established on the basis of database downloads, which are supported in various formats.

The relationships between native biomolecules of the host organism and drugs can be represented much like the relationships between the other compounds (Frolkis et al., 2010). New experiments and hypotheses can be derived by combining drug target information with the other pathways (Publication III; Liikanen et al. (2011)). In these combined pathways, one can interfere with the system via the administration of the selected compounds. DrugBank is a proxy database that combines information about medicines and the genome (Knox et al., 2011). The database provides information regarding to 6711 drugs, some of which are experimental (5084) or withdrawn (69). DrugBank describes which proteins are known to interfere with the particular drug and which parts of the protein sequences are responsible for the interactions. In addition, plenty of information is provided regarding the biochemical and pharmacological properties of each compound. KEGG provides another resource of drug targets and biomarkers affecting the drug responses.

3 Aims of the studies

Publication I Development of a sensitive method that works with a limited number of samples and is capable of revealing sites with possible recessive and CRC relevant mutations. Development of an automated analysis and visualisation pipeline that can be applied and customised for other SNP data sets.

Publication II An establishment of a computational framework for the systematic analysis of a large and heterogeneous data set. Built-in support for the documentation, integrity checks, and the simple manipulation of the analysis were the key requirements for the collaborative platform. We were interested in applying this framework to TCGA GBM data set in order to find new survival associated genes and to prepare the data in a more suitable format for other studies.

Publication III Development of a software infrastructure that would use existing information to reveal relationships and causalities between the genes of interest. Automatically generated graphs are produced on demand for the genome-scale studies so that they can operate on models adjusted for their data instead of using fragmented models of canonical pathways.

RelPublication Identification of the AR binding sites and the associated cofactors in human prostate cancer cells. Characterisation of interplay between AR and FoxA1, and the prognostic value of FoxA1 activity.

4 Materials and methods

We have used an iterative development process in all projects. An automated pipeline has been established for each project. This pipeline carries out the complete analysis from the pre-processing of the data to the reporting (Publication III; Laakso et al. (2011)). Individual steps have been implemented as components, which we have been able to recycle between the projects (Johnson, 1997). The exact configuration of the analysis pipeline evolves during the iteration as new hypotheses and adjustments are made on the basis of the previous results.

4.1 Detection of recessive mutations

The recessive mutations manifest a cancer phenotype in the absence of a dominant allele. Such conditions typically arise when the same recessive allele has been inherited from both parents. In this case the homologous chromosomes are identical for the particular region, which is now called *homozygous*. One may have also inherited two different recessive alleles (these individuals are called *compound heterozygotes*) or the dominant allele has been lost. The somatic deletions causing loss of heterozygosity (LOH) have been associated with the inactivation of various tumour suppressor genes (Huang et al., 1992; Cawkwell et al., 1994; Beroukhim et al., 2006). In familial predisposition cases we are trying to measure germline DNA and avoid somatic changes, although these changes cannot be totally excluded as the sample material has been isolated from blood (controls) and tumour surrounding tissue (cases).

The resolution of a SNP-microarray is a fraction of the complete sequence, but we show that the microarrays can be used for the detection of homozygous regions. First, there has to be enough SNPs in linkage disequilibrium, in other words, their alleles tend to segregate together. Second, both alleles have to be common enough so that they provide information about the sample haplotypes. In fact, the common definition of SNP implies that the frequency of the rare allele is at least 0.01 at the population level (Mooney, 2005). Once we observe a long series of such SNPs with homozygous genotypes in a particular sample, it becomes more likely that the sample has two identical copies of the same ancestral haplotype (shared identity by descent) than two different haplotypes with the exact match of the SNP alleles. The homozygous regions are interesting because we can assume that a (possibly existing) recessively manifested mutation co-segregates together with the SNP alleles, whereas the same does not hold for the former case of heterozygous regions unless they are compound heterozygotes. The detection of putative mutation regions with possible genotype mistakes resembles the detection at LOHs in the absence of the normal tissue references, but simple algorithms can be used as there is no need to distinguish between the homozygous and heterozygous signal strengths (Dutt and Beroukhim (2007); LaFramboise (2009); Publication I).

Each chromosome is processed separately as the partitioning of the data improves its computational analysis. The genomic regions that are identical by

descent (IBD) (Thomas et al., 2008) are considered to be continuous segments in the context of the human reference genome. The segments are limited by the meiotic recombination between the homologous chromosomes. Our study focuses on the autosomes; the sex and the mitochondrial chromosomes were discarded by the analysis pipeline.

The recessive phenotypes, such as tumour suppressor deactivation mediated cancer susceptibility, which are harmful for the individual, are rare and consequently the sample material is limited. We identified 50 unrelated patients out of 1044 CRC cases, which were collected during previous studies of DNA replication errors and microsatellite instability in CRC (Aaltonen et al., 1998; Salovaara et al., 2000). For these patients the cancer was not explained by the known mutations and they all had at least one sibling with a CRC diagnosed. The genotyping was carried out successfully for 42 CRC patients and 50 blood donors using Affymetrix GeneChip® Human Mapping 100 K Set and the standard protocol (Affymetrix, 2004). Two patient samples were known to harbour a mutation in *MUTYH* gene and they were used as spike-in controls for the evaluation of the hit ranking scheme.

An ideal comparison between the cases and controls would be based on haplotypes, which can distinguish between a mutation allele and other possible alleles present homozygously. However, the resolution of the microarrays we used and the small number of samples limited our options in that. In Publication I, we used haplotype estimates only to estimate missing data, but Haplous (Karinen et al., 2012) extends our methods for the comparisons at the level of haplotypes. In our case, the distinction between different alleles is made independently for each SNP by dividing them to the wild type alleles (those with the highest frequencies) and to the rare variants (the alternative nucleotide supported by the microarray). The advantage of this frequency based assignment is that it reduces entropy in the allele sequences and simplifies the visualisation as the wild type alleles (in upper case) are likely to be followed by wild type alleles (ABCDE versus AbCde).

CRC is a complex disease that can be caused by various different mutations (Kinzler and Vogelstein, 1996). Thus, we expect that only few samples share a common recessive mutation (Aaltonen et al., 2007). A detection method that would be able to call 2–5 cases among 42 patient samples and to compare them against 50 population controls has to be more sensitive than the standard statistical tests based on the frequency comparisons. Consequently, we formulated a rule based filter that provides a list of all genomic regions which have enough (> 1) overlapping homozygous regions in patients and which are rare (none is observed) among population controls.

A majority of the accepted regions are obviously not related to the CRC susceptibility, but a ranking scheme was established to highlight the most prominent candidates. The ranking is based on the score that represents the total length of the homozygous regions contributing to the region. Longer regions are more likely to be identical by descent (Thomas et al., 2008). A higher number of samples with such overlapping regions provides a higher association to the phenotype. The top scoring regions were subjected to an annotation pipeline that fetched

the overlapping genes, which were in turn compared against the CRC candidate genes suggested by the SNPs3D text mining service (Yue et al., 2006) and by Sjöblom et al. (2006) about frequently mutated genes in breast and colorectal cancers.

SNPs3D is a database that provides information about the phenotypes associated with SNPs and genes (Yue et al., 2006). The database consists of modules providing services such as functional annotations of SNPs and construction of interaction networks of the query genes. Disease Candidate Gene module is one part of the database specialised in text mining based mappings between diseases and genes. The mappings are generated in four steps using the abstracts of the articles stored in MEDLINE. First, abstracts referring to the name of the disease are selected. Second, the frequencies of the keywords among the selected abstracts are compared against their total frequency among the MEDLINE abstracts and a list of 40 most enriched keywords is formed. Third, a score is calculated for each gene as a sum of keyword specific affinities. These affinities are products of the ratios of the disease name and the gene among the abstract containing the particular keyword. Fourth, genes are ranked based on their scores and reported.

4.2 Data analysis framework

Integrative projects using multiple data types (for example gene expression, DNA methylation, DNA copy number, genotype data) and databases are challenging for the management of the data analysis. Each data type may introduce its own set of quality check, formatting, normalisation, analysis, and reporting steps to the project, and a complete analysis may consist of hundreds or even thousands of individual steps (Laakso et al., 2011). We prepared a computational framework called Anduril that can be used to combine different computer programs on different environments together and to handle the data flow between them.

Anduril provides its own workflow configuration language (AndurilScript) that is used to bind outputs of one program, or a *component*, to the inputs of the subsequent components. Each component provides an interface describing its inputs, outputs and the parameters, but the underlying implementation is language independent. The distribution provides convenience libraries for Bash, Lua, MATLAB, Perl, Python, R, and Java, but other languages and stand-alone applications can be used as well. Aggregate components can be prepared by combining other components together with the AndurilScript. These constructions, referred as *functions*, provide recyclable routines that can be used even across the projects.

All communication between components is mediated by the files produced by the upstream components and provided to their downstream components. The workflow engine takes care of the exact locations of the files and communicates them to the components. The downstream components are launched only after the completion of their upstream neighbours, which is in contrast to some other workflow engines like Orange (Curk et al., 2005) and Ergatis (Orvis et al., 2010) that support message streams between the components. Although sometimes

slower, the file mediated communication simplifies the language and platform independent implementation of components, enables recycling of still valid results between the executions, and provides a well-organised repository of results of each step of the analysis.

Anduril has been designed for projects which may consist of thousands of individual component instances. A graphical user interface would be impractical for such cases, thus a more suitable console based user interface has been chosen instead. The terminal based user interface simplifies remote access, and a session can be left open for days depending on the total execution time. The custom language was established in order to simplify construction and modification of workflows. The current version of AndurilScript supports inheritable data types, conditions, loops, nested workflows, arrays, etc. The compile time validation of the workflow and the re-execution of modified or out-dated results enables the maintenance of workflows with thousands of steps.

An integrative bioinformatics infrastructure enables large scale studies (Almeida, 2010). Identification of prognostic genes and possible therapeutic targets in GBM is a challenge where such infrastructures can be applied (Publication II). The Cancer Genome Atlas provides a wide set of measurements and clinical information about 338 (November 2009) GBM patients and their tumours (McLendon et al., 2008). We used Anduril to establish an analysis pipeline for the DNA copy number, SNP genotype, gene expression, microRNA, and methylation data. The DNA copy numbers were estimated from the comparative genomic hybridisation array data, and the gene expressions were measured using two different microarray platforms. Affymetrix HU133A provided information at the level of genes, whereas Affymetrix Human Exon 1.0 platform was capable of detecting differences between the exons. The information of each data type was merged at the level of genes, which led to a large matrix of genes and associated results. The final matrix was represented as a web site (`http://csbi.ltdk.helsinki.fi/anduril/tcga-gbm/`) for the simultaneous accession of multiple aspects of the cancer.

The plausibility of the survival association between the result genes was tested on four (three glioma and an SV40 transformed fetal astrocyte) cell lines. A total of 11 genes (*CDKN2A*, *FLNC*, *H19*, *HIST1H4L*, *KIAA0040*, *LTF*, *NNMT*, *POSTN*, *TAGLN2*, *TIMP1*) were selected on the basis of upregulated expression and the survival association. We tested if the downregulation of these genes has an impact on the proliferation or apoptotic activity of the cells. A small interfering RNA (siRNA) silencing was performed with four different siRNAs against each gene. The proliferation of each cell line was reduced by the *MSN* targets, but the responses of the other genes were less consistent or negligible.

4.3 Analysis of AR binding sites

We studied the target genes of AR on a modified LNCaP-1F5 human prostate cancer cell line that expresses rat GR. The modified cell line was chosen because we were also interested in the interplay between AR and GR. The DNA binding responses were measured 2h after a 5a-dihydrotestosterone stimulus and the

mRNA was collected 24h after the stimulus. Another series of experiments was carried out after depletion of FoxA1 with small interfering RNA.

The transcriptional responses were analysed from the collected mRNA. Two replicates were prepared for each condition (before the stimulation, AR response, FoxA1 depleted cells before the stimulation, and the same cells after the stimulation). Illumina HumanHT-12 v3 Expression BeadChip Kit was prepared at Biomedicum Functional Genomics core facility and the data files were sent to us for the data analysis that was implemented in Anduril (Laakso et al., 2011). The signals of all samples were quantile normalised, and the mapping between the microarray probes (`illumina_humanht_12` probe set) and the genes was revised using Ensembl BioMart. A fold change was calculated for each gene between the medians of the replicates by dividing the response signal with the preceding control signal. Genes with a fold change less than 1/1.7 and greater than 1.7 were considered down- and upregulated, respectively. The threshold was selected empirically so that enough DEG were obtained for the peak assignment and the downstream analysis but simultaneously avoiding spurious results caused by the limited number of replicates. No false positives were found during the validation of the transcriptional changes but only few genes were covered in details (RelPublication Supplementary Information).

The identification of the transcription factor binding sites and their target genes was conducted in multiple steps. In the beginning, the short read sequence data was aligned against the human reference genome using ELAND algorithm (Bentley et al., 2008). Two mismatching base pairs were allowed within an accepted short read. Reads that did not align to the reference genome or had multiple alignments were ignored. The alignments of the accepted reads were used to estimate the relative abundance of the DNA fragments along the genome. Figure 4 illustrates the counts of overlapping short read alignments at two selected regions of the genome.

The sequence alignments were analysed with MACS (Zhang et al., 2008) algorithm that estimates their local enrichments across the genome. MACS is a peak calling algorithm that has been designed for the TF ChIP-seq. A variety of algorithms, which differ in their sensitivity and the widths of the peaks called exists for the same purpose (Wilbanks and Facciotti, 2010). In addition to MACS, we also analysed our samples with SPP (Kharchenko et al., 2008), which is an R package for the ChIP-seq analysis. MACS is relatively stringent in calling peaks and reports narrow and highly reliable peaks but is likely to miss some less obvious sites. SPP turned out to be even more conservative, reporting about 1/3 less sites, which is consistent with the FoxA1 observations made in Wilbanks and Facciotti (2010). The input of the peak calling programs consists of the aligned short reads of the actual TF ChIP-seq (case) and of those representing an unspecific antibody (control). An enrichment is called where the case reads have a considerably higher overlap of reads in comparison to the control set. The effects of the clonal artefacts produced by the polymerase chain reaction based amplification of the DNA are reduced by accepting only one read for each unique sequence (Wang et al., 2011). The exact peak position is adjusted between the

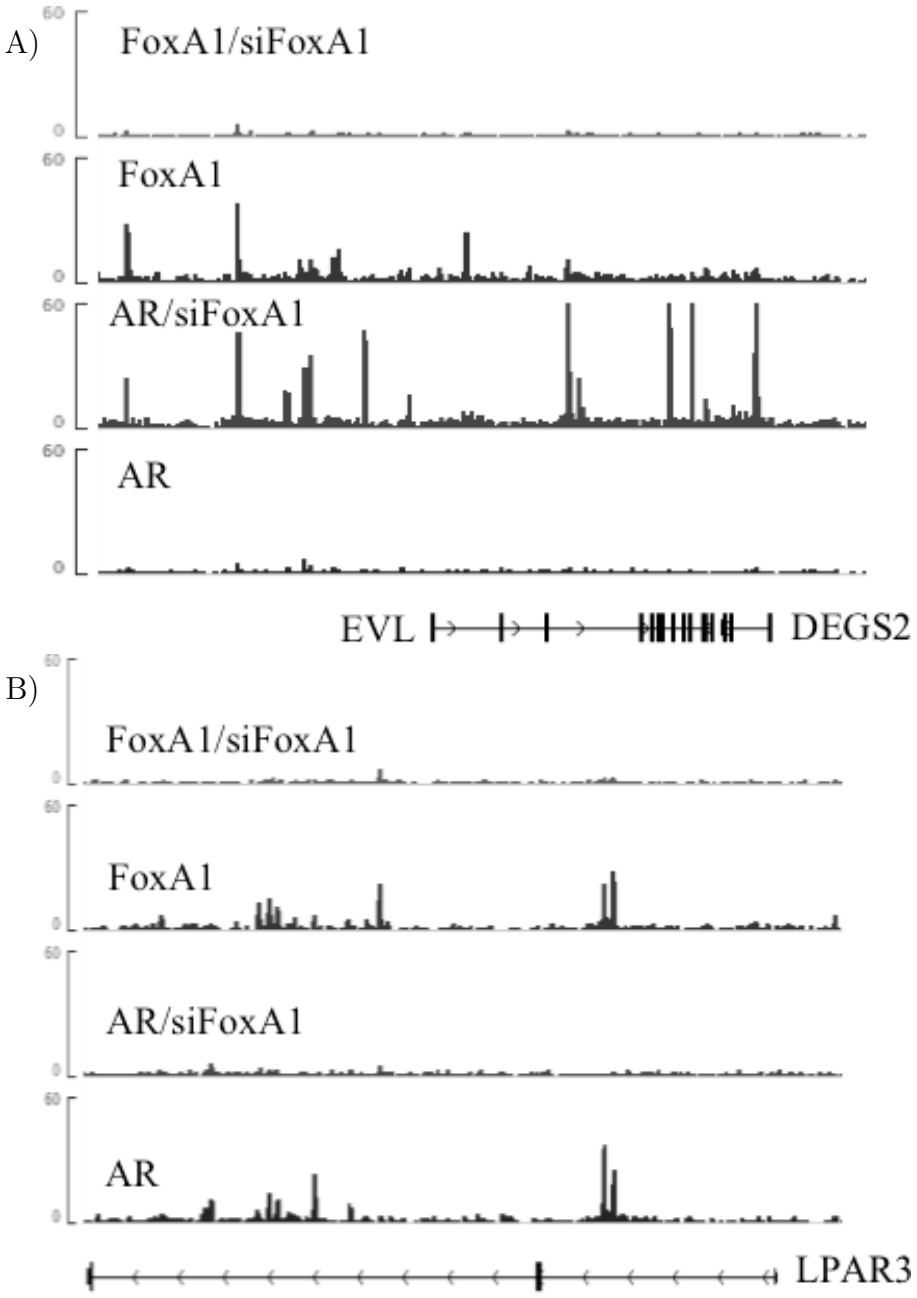

Figure 4: Examples of the local enrichments of the aligned ChIP-seq reads under various conditions. The upper most signal represents FoxA1 binding on FoxA1 depleted cells (should not be observed if the depletion is efficient), below is the same for the parental cells followed by a similar pair of AR bindings. A) depletion of FoxA1 brings AR to the new sites, some of which match well with the previous FoxA1 sites at the promoter of *EVL* gene. B) FoxA1 dependent AR sites of *LPAR3* are lost after the depletion.

enrichments on both of the DNA strands. These sites represented an overall set of binding sites p and their fold enrichments w against the control samples. ChIP-seq peaks analysed were determined as overlaps of the MACS (Zhang et al., 2008) peaks of all ChIP-seq sample replicates. Two replicates were produced for AR and FoxA1 in parental cells and AR in FoxA1 depleted cells. Four replicates of FoxA1 BSs were prepared for the FoxA1 depleted cells to compensate the weak signal and the relaxed peak calling criterion. Having the replicates, we considered the wider set of MACS peaks more comprehensive and reliable enough, although the SPP peaks were probably more accurate for the positioning of the binding site (a single bp coordinate was returned for each peak).

The assignment of the ChIP-seq peaks to their target genes produces a many-to-many relationship, where each peak may have none or many target genes. We constructed this mapping by assigning each peak to all genes within the distance of 100000 bps. Next, each gene g was assigned a score:

$$score_g^{\text{linear}} = \sum_p w_p \max\left(0, \frac{100000 - |d_{g,p}|}{100000}\right) \tag{1}$$

that combined the information about the distance d of each peak, their strength, and the number of TF BSs around the gene. Our formula is a linear version of the exponential score:

$$score_g^{\text{exponential}} = \sum_p w_p e^{\frac{-d_{g,p}}{100000}} \tag{2}$$

presented in Ouyang et al. (2009). An arbitrary cut-off of 11 was used to select a set of genes with a TF association. This cut-off provided a good overlap with the differentially expressed genes (DEG) while reducing the number of genes associated. The downstream analysis of the TF BSs was conducted for the sites that contribute to the scores of the selected genes. The rest of the peaks were filtered out in the absence of TF associated genes in their proximity.

The possible co-factors of the AR were not well known when we initiated our study but the ChIP-seq protocol enabled a detailed profiling of the AR binding sites. The DNA sequences of thousands of binding site peaks were used to describe the surroundings of the AR bindings in terms of TF binding sequences of possible co-factors. A more focused analysis was performed by combining the gene expression data with the BSs and by focusing on the sequences assigned to DEG. This BS oriented approach reduces the complexity of the traditional promoter sequences analysis that considers a possibly long upstream sequence ahead of the genes of interest (D'haeseleer, 2006).

First, the peak sequences were fed to MEME (Bailey et al., 2006) for the *de novo* motifs discovery. MEME is an iterative expectation maximisation (EM) algorithm that collects statistically unlikely but still frequent sequence motifs (Bailey, 1994; D'haeseleer, 2006). When successful, this approach is capable of revealing unexpected or unknown binding site sequences that are enriched in the given set of input sequences, as we will show later in the results

section. In this context, EM is an optimisation procedure that adjusts the motif so that the equally long fragments of the input sequences are more likely generated by the motif than the background probabilities (typically relative frequencies) of the nucleotides. The motif is represented as an n nucleotides long position weight matrix of the nucleotide probabilities. In our case, n was between 6 and 15. The probabilities of the input alignments are used as their weights when the matrix is revised for the new positional averages and for the next iteration. The preprocessing of the MEME input involved filtering of the repetitive sequences (conducted in the Ensembl database) and selection of at most 500 most reliable peaks. The repeat masking prevents the program from discovering the spurious patterns representing nucleotide repeats that are of no interest. The most reliable peaks are selected based on the fold enrichment score of the corresponding MACS peaks, and the limit of at most 500 peaks is used to guarantee a reasonable computational time (Hu et al., 2010). Neither the sequence and alignment qualities nor fold enrichments are considered during the motif construction. A more recent algorithm called Hybrid Motif Sampler is able to utilise this information while processing tens of thousands of ChIP-seq peak sequences (Hu et al., 2010). We have not yet tried the program but consider it for the future.

Second, known binding motifs were downloaded from JASPAR (Portales-Casamar et al., 2010) database. These motifs were combined with some other motifs using a method we prepared based on literature (Wang et al., 2007). Figure 5 illustrates the generic function that was used to align all motifs against the peak sequences and against sequences of the same lengths and chromosomes, but randomised positions. The alignments between the peak and the random sequences were compared in terms of frequencies of matches; scores above the threshold (0.9) dependent limit we consider to represent a reasonable match. The exact limit was calculated for each motif based on the highest (S_h) and the lowest (S_l) possible alignment score (the scores of the best and the worse possible sequence match): $\text{threshold} \cdot (S_h - S_l) + S_l$.

4.4 Interpreting new results with the existing information

We are using existing knowledge to organise the initially less structured data, such as BSs and DEG, into forms that are easier to interpret and more suitable for further predictions. We found fusion pathway models especially suitable for this purpose since these models are not restricted to the limited scopes of the canonical pathways (Publication III). The directed edges can be used to derive causalities that cannot be told from the static input data. The edge types representing the modes of interactions between the pathway entities are taken into account during the model construction in order to avoid interactions less relevant for the observations. Although useful in revealing dependencies and separating causes from consequences, predicted properties of the overall graph based on its members should be treated with caution (Gillis and Pavlidis, 2012).

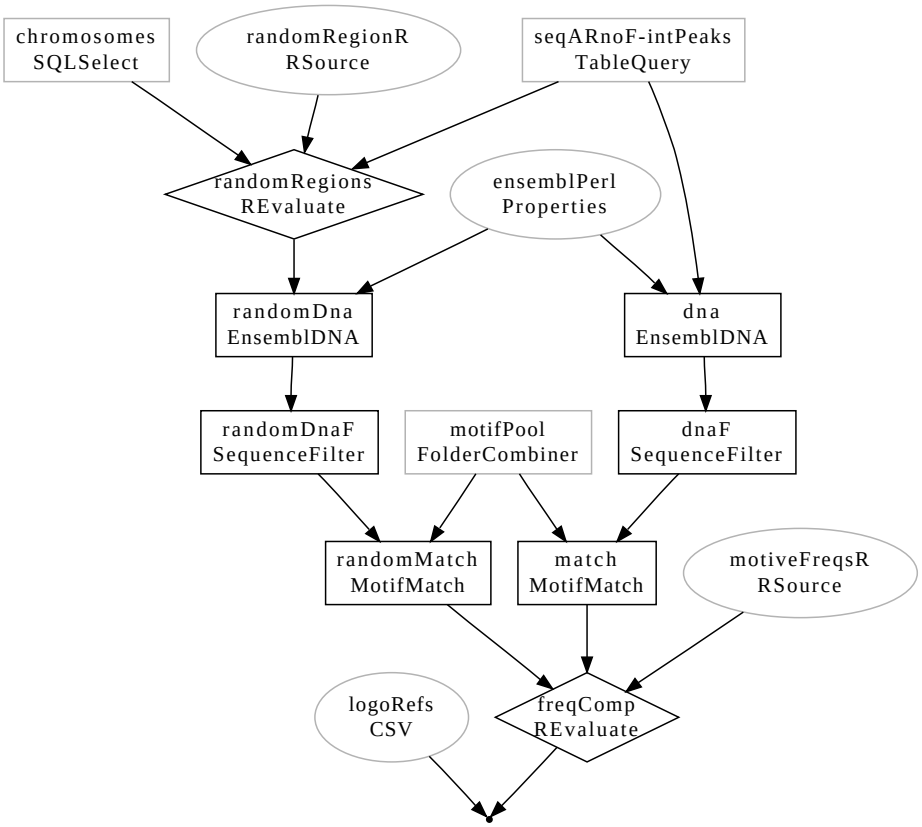

Figure 5: The structure of the Anduril workflow that compares alignments of the selected motifs (motifPool) between the given sequences (seqARnoF-intPeaks) and the random background (randomRegions). The labels of the nodes represent names of the component instances and their types. The rounded shapes are files whereas the rectangular ones represent Anduril components. The output of this workflow is a LaTeX table representing the differences.

The roles of individual genes may be context specific; their actual membership may have different confidence; and the actual outcome may be determined by a minor fraction of the participants.

The metabolic pathways have been included in our system, but we have adjusted the level of abstraction for them. Our focus has been in the genetics, and thus we replaced the reactions of small molecules with the connections between the enzymes related to the reactions producing and consuming them. The mapping logic is partly adapted from the Simple Interaction Format representation provided by Cerami et al. (2010). Enzymatic activities of the proteins are stored as Enzyme Commission numbers (Webb et al., 1992) assigned to the encoding genes, which are annotated with their Ensembl and Entrez identifiers for mutual compatibility (Philippi and Köhler, 2006).

The protein level relationships, such as phosphorylations, protein-protein interactions, etc., are represented at the level of the encoding genes. This conceptual simplification prevents us from modelling differences between the differentially

spliced isoforms and between the different states of the proteins (Philippi and Köhler, 2006). We chose this approach once we noticed that the available information, even if presented as protein identifiers, makes no difference between the isoforms. Consequently, the use of protein level links produced a trivial expansion of the current model. All the proteins of the genes were connected to all other proteins of another gene if there was a relationship between any of these proteins. The proteins and their UniProt identifiers (The UniProt Consortium, 2012) are preserved in the database and connected to their encoding genes.

Our fusion pathway construction methods (Publication III) are intended for the genome-scale studies where little is known about a large set of entities. For example, we may only know the mutation status of the genes or their expression status. On the other hand, this information may be available for a significant portion of the genome. Consequently, our models are more descriptive than predictive in comparison with kinetic models, which are more detailed (considering concentrations and reaction rates) and focused on a smaller number of entities (Ruths et al., 2008). The balancing between conflicting signals, such as concurrent promotion and inhibition, is solved by including both options with an undetermined outcome.

Each link between bioentities has a direction and a type, which specifies its meaning in terms of how the source entity is related to the target entity (for example, the source gene encodes the target protein or the source drug inhibits the biological function of the target entity). Undirected relationships, such as protein-protein interactions, are stored once for both direction. The links may have weights associated to them but suitable reliability scores are rarely provided by the source databases, and thus this feature is not used in our analysis. Resources such as Biomine (Eronen and Toivonen, 2012) and STRING (Jensen et al., 2009) may be considered for this information in future. Additional key-value pairs can be attached to the links for more detailed information. The use of these pairs may depend on the on the link type. Some of the keys have been reserved for the source databases whereas the values are used for more specific references to the origin of the link. The information about the origins of the links is useful for the tracing purposes, but patterns of key-value pairs are also used to restrict the scope of the queries. For instance, links of certain canonical pathways can be obtained by referring to the database (key) and by enumerating the pathways (values) of interest.

It is essential to upgrade database content as new information becomes available. Some data sources, such as KEGG and TCGA, are highly variable providing some changes almost each week. We have implemented our data retrieval algorithms so that they can be executed as often as needed and they will add previously unobserved items to the database. The removal of expired entities and connections is more challenging than the insertion of new entries. This is because one has to compare all previously collected information against the new information available in order to know what has been left out. Many database items are referred to various data sources, which means that one should make sure that none of them is using a resource before removing it. We have

solved this consistency requirement by establishing an automated and repeatable construction pipeline that builds the complete installation from the source code. The structural changes of the database schema and the alterations in the behaviour of the importing routines can be carried out without transformations to the already stored data. The complete build cycle of the installation script takes less than a day.

In Publication I we demonstrated integration between the literature and the result sets using SNPs3D and an article about frequent mutations in breast and colorectal cancers (Sjöblom et al., 2006). The same concept has been further developed to support simultaneous comparisons across dozens of data sources. The originally used article has been replaced with wider and more recent cancer data sets like COSMIC, TCGA and Tumorscape (Beroukhim et al., 2010). Each set of candidate genes and the associated context specific weights are now stored into a relational database where they are readily associated to the diseases, pathways, drugs, and other stored concepts. A set of algorithms has been designed that can be used to compare gene sets against each other, to describe genes with the supporting evidence, and to provide meta-ranks based on appropriate sets. For instance, a meta-study about tumour suppressor genes can be conducted by combining the related candidate sets to a meta-rank that provides the genes most frequently associated with a chromosomal deletion, mutation or transcriptional dysregulation with the highest confidence (Partanen, 2012). Having a meta-rank of study specific averages, produced in seconds, one can tell which results are most prominent candidates among a local set of results. An example of potential tumour suppressor genes is shown in Table 1. The original rank is pruned by excluding the genes that were ranked high based on chromosomal amplifications in the same tumour data sets as they were frequently deleted. Scores of each individual study are normalised by sorting them based on their original scores and by replacing the original scores with a linear rank that comes from one (the highest score) towards the last item of $1/n$. An average of linear ranks is applied to items with an equal original score and thus having an unspecified order. Zero is considered for the genes outside of the set as the study provides no evidence supporting them. The zeros are used in place of missing values because many gene sets are based on genome-wide studies, which may have actually considered these genes, although they have not been included to the result set. The measures of the genes may have been below the study specific thresholds and such genes would probably score close to zero if included. By expanding the gene sets in this way, one could leave zeros out when counting the averages for the meta-rank and the method may work better for the studies focusing on small fractions of the genome.

Gene	cosmicMetastasis	cosmicPrimary	cosmicRecurrent	tcgaOvarianGE	tcgaBreastGE	tscapeBCd	tscapeOvariand	tscapeProstated	tscapeMelanomad	score
TMPRSS5	0	0.886	0	0	0.89	0.954	0	0.507	0.797	0.448
SYNE1	0.793	0.756	0.712	0	0.853	0	0.784	0	0	0.433
PARK2	0	0	0	0	0.661	0.767	0.809	0.507	0.981	0.414
EBF2	0	0	0	0	0.848	0.998	0.906	0.972	0	0.414
PTEN	0.617	0.394	0.149	0	0.322	0.324	0	0.915	0.958	0.409
RB1	0.612	0	0.147	0	0	0.968	0.995	0.94	0	0.407
CDCA2	0	0	0	0	0.838	0.996	0.842	0.966	0	0.405
GNRH1	0	0	0	0	0.797	0.996	0.842	0.966	0	0.4
KCTD9	0	0	0	0	0.736	0.996	0.842	0.966	0	0.393
FOXO1	0	0	0	0	0.863	0.871	0.978	0.781	0	0.388
CDKN2A	0.686	0.39	0.709	0.997	0.679	0	0	0	0	0.385
TP53	0.837	0.771	0.974	0	0	0	0	0.81	0	0.377
KIT	0.816	0.676	0.903	0	0.983	0	0	0	0	0.375
DOCK5	0	0	0	0	0.512	0.996	0.869	0.972	0	0.372

Table 1: This example of tumour suppressor genes shows genes with the highest mean scores excluding the genes with frequent amplifications in breast, ovarian, or prostate cancer or in melanoma. Chromosomal deletions are taken from Tumourscape. Mutation frequencies have been collected from COSMIC and the gene expression data comes from our Anduril based TCGA workflow.

5 Results

All methods described in this dissertation are freely available on Internet. The documentation and the software downloads can be found from the following sites:

The detector of the recessive mutations (Publication I)
http://www.ltdk.helsinki.fi/sysbio/csb/downloads/CohortComparator/

Anduril workflow engine (Publication II; RelPublication)
http://csbi.ltdk.helsinki.fi/anduril/

Moksiskaan database and the related Anduril extensions (Publication III)
http://csbi.ltdk.helsinki.fi/moksiskaan/

5.1 Analysis of the CRC genotypes

We produced a highly sensitive method that is able to use SNP genotype data to narrow down the genomic search space for recessive mutations. Four important optimisations are:

1. selection of the sites that are homozygous;
2. filtering of the sites that are homozygous in the control samples;
3. annotation and property based ranking of the sites;
4. reporting each site at the level of individual homozygous samples.

The great reduction of the search space implies a change of missed mutations. We performed a probabilistic simulation of the algorithm to see how likely it is able to detect a mutation for various combinations of mutation frequencies of cases and controls (Publication I). Based on these simulations it seems that a mutation region is preserved with a reasonable probability when the probability of homozygous regions in control samples is low at that site (or they have another allele). The method is not suitable for the low penetrance phenotypes where a higher fraction of control samples may carry a mutant genotype.

The algorithm was tested by extending the set of 40 case samples with two samples having known but different mutations in *MUTYH* gene. The spike-in controls were correctly identified with partially overlapping, long, homozygous regions. The gene annotation based ranking proposed *MUTYH* site as the fourth most prominent one, which suggests that only few sequence validations would have been needed in order to detect these mutations had they been unknown.

The rule based filter produced 1874 genomic sites compliant with the recessive pattern. We used the shape of the ranking score distribution to estimate its saturation point, which left the 181 highest scoring regions for further inspection.

A mutation of our interest could have a form of a non-lethal deletion in a tumour suppressor gene. We assumed that the SNP genotype calls of homozygous deletions or unexpected alleles would be missing values, and used our method to identify short sequences of them much like we did for the homozygous regions. The analysis identified 13 possible sites. Each site had at least two cases with

sequences of at least three SNPs without a genotype call but no controls with a sequence of more than one genotype failure. Polymerase chain reaction assay validation was unable to capture deletions at these sites, suggesting that the deletions are better captured by the copy number detection methods specially designed for this purpose (Colella et al., 2007; LaFramboise, 2009).

5.2 Candidate pathways

We developed a methodology and a computational platform, Moksiskaan, for the rapid construction of fusion pathways for a given set of genes and proteins (Publication III). A centralised and uniformed database provides a useful resource for systems biology (Turenne, 2011). We demonstrated the capability of our platform in detecting crosstalk between Jak-STAT and ErbB signalling pathways (Publication III) — a detection that was later described also in WikiPathways (pathway WP437).

The biological databases are not readily compatible with each other and all details are hard to preserve when merging their content into a unifying schema (Philippi and Köhler, 2006). For example, the upper model in Figure 6 illustrates the AR signalling pathway (WP138) generated from the Moksiskaan database. The original topology was represented in Figure 3. Total of 85 out of the original 90 genes preserved their pathway association after the identity conversion, but the connectivity was almost lost. In this case, the source database represented these connections with visualised arrows without a semantic association between the genes on both ends. The bottom model in Figure 6 represents connections between the 85 genes belonging to the AR signalling pathway irrespectively of the canonical pathway that has described it in KEGG, Pathway Commons, or in WikiPathways.

Candidate pathways can be used to describe how certain biological processes such as cell motility, cell death, or cell differentiation, growth and proliferation are reflected by the transcriptional profile. We demonstrated this strategy by characterising the effects of depletion of HuR protein in the context of ductal carcinoma (Heinonen et al., 2011). The processes of interest were first expressed in terms of GO identifiers, which were then mapped to the genes with the corresponding annotations in Ensembl database. A candidate pathway was constructed so that DEG with the annotations of interest were used as seeds but all DEG were used for the status information.

One of the advantages of an integrated biodatabase is that it enables simple queries across multiple sources of information. Such queries are useful when the relationships and the roles of some experimentally observed genes or proteins are not known (Eronen and Toivonen, 2012). For instance in Karhemo et al. (2012) we had 23 genes encoding cell surface proteins with different abundances in metastasising cells in contrast to cells producing no metastases to the lungs. The roles of the genes were first characterised by fetching the downstream target genes, diseases, biological processes, and the molecular functions. Processes such as cell adhesion and migration were observed as enriched GO terms and being influenced

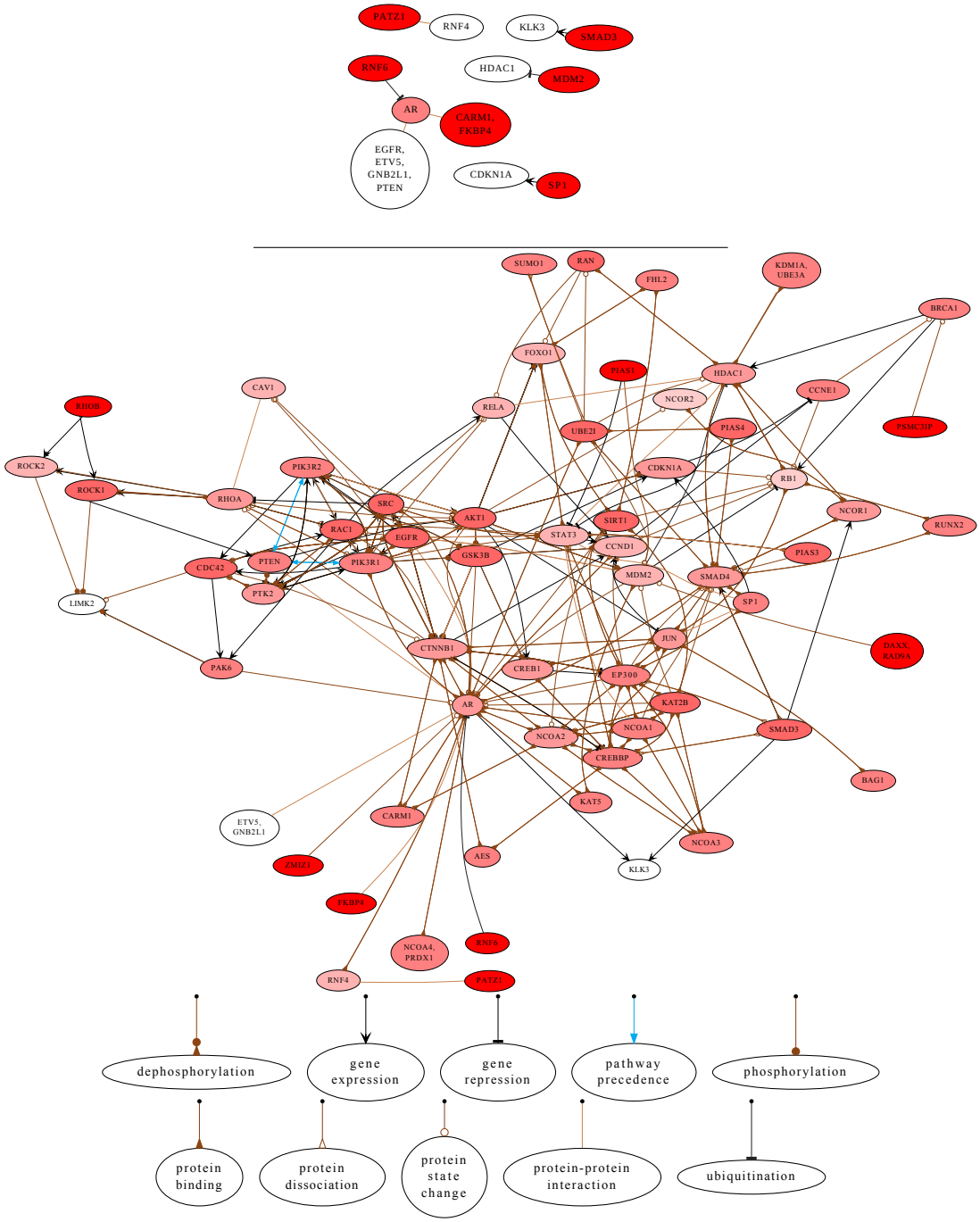

Figure 6: The upper pathway represents androgen receptor signalling as stored to the Moksiskaan database. The majority of the original connectivity (Figure 3) has been lost during the importing procedure. The lower pathway has been constructed between the members of the original pathway but the connectivity represent the union of the pathways from KEGG, Pathway Commons, and WikiPathways. The ratio of input and output connections determines the shades of red used for the genes. Red genes have only downstream neighbours whereas the white ones have none.

by the proteins (regulation terms in GO). Next, the relationships between the original genes and the other genes were investigated by retrieving the other genes connected to them and by establishing all the direct connections between the resulting genes. The initial graph was pruned by removing edges with a low Spearman's rank correlation in the TCGA breast cancer expression data and by excluding the genes without neighbours. In the final graph, the original edges were restored between the remaining genes.

5.3 Interplay between AR and FoxA1

The interplay between FoxA1 and AR appeared pivotal once we received the first results regarding AR BSs. We had executed the motif enrichment pipeline (Figure 5) and FoxA1 was at the top of the list. The co-operation of AR and FoxA1 (also known as hepatocyte nuclear factor 3, alpha) was previously described in the context of breast and prostate cancer (Gao et al., 2003; Makkonen et al., 2008; Hurtado et al., 2010). The enrichment of the FoxA1 BS motif suggested that the interplay of the TFs is not limited to few individual sites but represents a common pattern worth genome-scale experiments.

The coverage of the AR binding sites was estimated by sampling random subsets of the aligned reads and by running the MACS peak calling algorithm for them. The assumption is that the peak set and the number of peaks would not change much once there are enough reads to give a reasonable cover for all binding sites. Total of 20 iterations of sampling and peak calling were performed for two AR replicates for seven different ratios of sampling coverage. The overlap of peaks between the sample pair was also calculated for each iteration as we focused on these overlaps in the actual study, and because it illustrated the similarities between the peak sets. Ideally, if the binding sites would be the same in both replicates, and the samples would provide a perfect coverage, then these overlaps should be almost identical to both replicates. The actual numbers of peaks represented in Figure 7 show little sign of saturation, and thus we believe that we have not reached the complete set of the existing binding sites yet.

An interesting sequence enrichment that was present in parental cells but not in FoxA1 depleted cells was found from the AR binding site sequences. The sequence motif resembled the AR-motif tail followed by another AR-motif tail or the beginning of the FoxA1-motif. A similar joint AR-FoxA1-motif was also discovered in an independent study of AR and FoxA1 silencing (Wang et al., 2011). The close proximity of the half sites and the absence of the canonical motif suggests formation of a protein complex, especially because AR and FoxA1 are known to interact with each other (Gao et al., 2003) — a fact that we were able to pinpoint from the Moksiskaan database. The annotation of the protein-protein interaction between *FOXA1* and *AR* told that the evidence was obtained from PINA, that had collected it from Human Protein Reference Database, that pointed to Gao et al. (2003) via PubMed.

Our AR and FoxA1 study demonstrates how large projects can be handled with Anduril. The final form of the data analysis workflow included more than

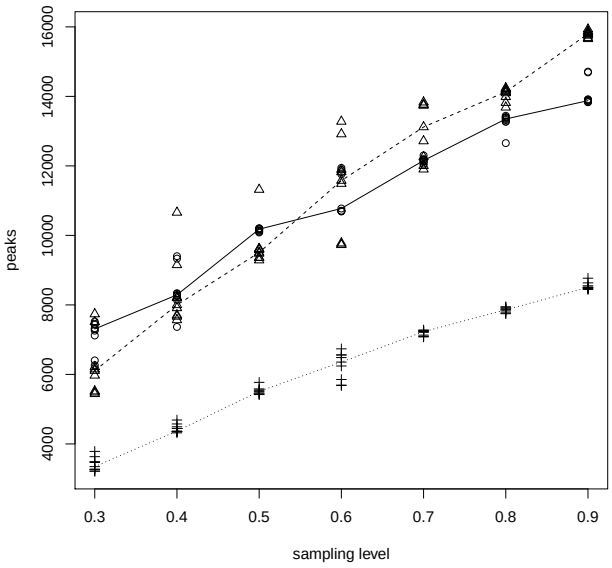

Figure 7: The number of observed peaks in MACS grows as a function of the input sequences. The circles and triangles represent peaks from two different AR replicates, and the crosses represent their overlap. The sampling level tells the ratio of reads randomly selected from the total number of reads sequenced. The lines have been drawn between the medians of the peak counts.

1500 input files and component calls, which were all maintained as a single Anduril process (Laakso et al., 2011). The execution time of this process is roughly two days when at most 12 simultaneous threads are used on a multi-core server. Most of the time is spent on the *de novo* motif discoveries. The workflow includes characterisation of 58 sets of BSs representing various intersections and overlaps between the BS sets and the expression data, which illustrates the usefulness of the code recycling provided in Anduril. The results of each 53 analyses are reported in the Section 12 of Laakso et al. (2011).

6 Discussion

The use of a component-based workflow engine enables the repeatable analysis of large data sets and helps the management of complex analyses with thousands of steps. The repeatability and the flexibility of these analyses becomes even more important in the future as new information and tools become available, and some of the current approaches shall be replaced. The quick adaptation to the enhancements in the field will be an advantage in cancer research and it enables smooth improvement of the results. An iterative approach may be applied to research projects consisting of an input data, analysis of that data, and reporting of the results. In that approach, the analysis is revised on the basis of new relevant data and the methodological progress in the field. An automated system may be used to upgrade the results and to compare the changes. A standard interface can be used to publish the latest results for the other projects so that they can revise their analyses accordingly. Having such an interlinked network of actively maintained research projects is likely to accelerate responses to new ideas and observations. The repetition of the same tasks under various conditions aids the identification of possible flaws that can be fixed and optimised. In the context of recycled components and interlinked analyses, the improvements may help a wide community.

One of the challenges in reproducing the work conducted elsewhere is the lack of detailed descriptions of all affecting parameters and approaches used in the original work. For this reason, electronic supplementary materials about the analysis are recommended for the genomic publications (Shi et al., 2010). Anduril is capable of describing the complete workflow in a detailed manner that is suitable for the article supplementary material, thus the recommendation has been followed in various publications (Publication II; Publication III; Heinonen et al. (2011); Laakso et al. (2011); Liikanen et al. (2011); Karhemo et al. (2012)).

The establishment of an integrative systems biology infrastructure is a notable effort that could benefit a wide community if shared openly and documented properly (Philippi and Köhler, 2006). We have followed common standards in our application programming interfaces and provided an extensive set of documentation, most of which is automatically generated and kept in synchronisation with the project. Although technically shareable and open, Moksiskaan database have been kept mainly at local installations because of the issues related to the licensing of the source databases. The more external code libraries and resources are integrated, the more complicated the copyright and sharing policies become (Philippi and Köhler, 2006). Data sets may be provided for academic projects but the copyright restrictions prevent researchers from establishing secondary on-line services on the basis of them. Original measurements are often published together with the gene expression microarray related articles, and funding agencies such EU and some governments are encouraging researchers to share the data they have produced. Similar strategy could be applied to the software and databases produced by the scientific labs by promoting licenses that allow their use in other services. A simple licensing scheme would be a great leap towards large scale

data integration systems, such as IPAVS, in contrast to the currently fragmented nature of the biodatabases.

The quality of the data in an integrative system is heavily dependent on that of the original sources. A joint set of information is vulnerable of false positives if an incorrect piece of information is preserved by at least one of the sources. For this purpose, special collections of curated corrections can be beneficial. These collections would represent negative evidence i.e. what should not exist in the database so that it can be distinguished from the missing information. The database schema of Moksiskaan supports curation data sets providing counter evidence regarding the bioentities and the relationships between them. In future, we may use this functionality to publish some curation data sets. Some of the flaws of the source databases can be captured during the data integration as they may violate integrity constraints of the system or cause unexpected situations when combined with the other sources of information. Although an integrative database enables parallel profiling of many data sources, it may introduce its own errors when interpreting and converting the source information to the local data structures (Moussouni et al., 2007). For the curation purposes, the traceability of the information is essential so that the users of the database can find the original sources of evidence, and the possible mistakes can be corrected where they were introduced. Even the curated databases may have a significant portion of mistakes and missing information (Cusick et al., 2008), thus periodic upgrades can be seen as an integral part of the bioinformatics.

The analysis of homozygous mutations was an example of a study with a relatively small number of samples. The conditions are even more challenging in the analysis of individual tumours required for personalised medicine. We have touched on this field by characterising enzymatic alterations in glioblastoma patients with Moksiskaan tools (Louhimo et al., 2012). In that study, we show how to use large sets of other patients as background knowledge to be able to tell the important alterations from the less interesting ones. The use of generic and adjustable analysis tools provides cost effective processing of individual patient data. Still, novel methods and a multitude of samples are needed in order to interpret the variants that are not part of the reference genome. At an individual level these non-alignable sequences may constitute a reasonable part ($\sim$9%) of the genome (Chen et al., 2012). Ultimately, the measurements of each patient are analysed using his/her own, automatically annotated, genome as a reference.

Prediction of patient survival based on the genome-scale measurements of the tumour is an interesting challenge for future research. We are planning to use Moksiskaan pathways for these prognostic purposes. We postulate that a predictor, once accurate enough, could infer therapeutic targets most optimal for increasing survival. The advantage of an *in silico* model is that the effects of various perturbations can be predicted before applying them to the patient. We can simulate the effects of all known drugs against the patient model and calculate their effects on survival estimates. The resulting information may help when comparing risks and possible therapeutic effects associated to the treatment of a cancer patient.

7 Acknowledgements

This work was carried out at the Department of Biochemistry and Developmental Biology at the Institute of Biomedicine and at the Genome-Scale Biology Program of the Research Program Unit during the years 2006–2012. Institute for Molecular Medicine Finland and CSC — IT Center for Science Ltd have kindly provided us with computational resources to a significant extend. I am grateful for the financial and educational support of Finnish Doctoral Programme in Computational Sciences FICS and its predecessor Graduate School in Computational Biology, Bioinformatics, and Biometry. I am especially thankful for Heikki Mannila, Heikki Lokki, Esko Ukkonen, and Ella Bingham for their efforts in these programmes and in helping me with my studies. Juho Rousu earns my compliments for supervising my Master's thesis (Laakso, 2007) related to Publication I. I want to thank the Cancer Society of Finland and the Association of the Nordic Cancer Registries for their assistance in statistics and cancer epidemiology.

I warmly thank Sampsa Hautaniemi for supervising my work and for his careful and rigorous guidance to the scientific community. I appreciate Sampsa's demanding but reinforcing practice in managing work in his laboratory. Students are not simply expected to do their best but they are given a fair chance to succeed in terms of time and other resources. The demand of publications and qualifications has not prevented us from conducting our tasks properly and from learning things well. It is largely because of the resources spent in testing, refactoring, maintenance, and packaging of our software that we have been able to work with as many and as large projects as we do today.

The past six years in Computational Systems Biology Laboratory have been exiting. I have honestly enjoyed the work not only because of the suitable facilities and inspiring projects but because of the nice colleagues occupying the laboratory. Together we have been through a countless number of interesting discussions and debates and together we have been establishing the ecosystem of bioinformatics tools in Anduril. Thus, I want to thank all the current and past laboratory members including: Amjad Alkodsi, Alejandra Cervera, Emanuela Henao Diaz, Javier Núñez-Fontarnau, Anna-Maria Lahesmaa-Korpinen, Chengyu Liu, Minna Miettinen, Kari Nousiainen, Lilli Saarinen, Miko Valori, and Mikko Kivelä.

The methodologies we have developed, such as Anduril, are readily applicable for industrial purposes. I thank Sampsa, Erkka Valo, Kristian Ovaska, Lauri Lyly, Ping Chen, Riku Louhimo, Rony Lindell, Sirkku Karinen, Viljami Aittomäki, Ville Rantanen, and Vladimir Rogojin for their enthusiasm in commercialising these technologies and in establishing Significo Research Ltd for this purpose.

This dissertation has been reviewed by Tero Aittokallio and Tapio Visakorpi. I express my gratitude to both reviewers for their critical and instructive comments. I appreciate the help Tiia Pelkonen provided in proof-reading and correcting my language.

I am proud that I have been given a change to work with Lauri Aaltonen and with his excellent laboratory. Special thanks to Auli Karhu, Rainer Lehtonen, Sari Tuupanen, Iina Niittymäki, and Eevi Kaasinen for their advice in genetics. I

have enjoyed using Riku Katainen's Rikurator program to visualise sequencing data and I am happy that he accepted to integrate Moksiskaan into the program.

Our long collaboration with Olli Jänne and Biswajyoti Sahu is a good example of a fecund interaction between the bioinformatics and experimentalists. Together, we have been performing tens of iterations of experiments, analyses, and yet another set of experiments or an alternated approach.

I exalt Juha Klefström, Johanna Englund, Mikko Myllynen, and Topi Tervonen for the biological validations and scrutiny of the tumour suppressor gene prediction algorithm. The study of the epithelial proteins and their functions has been a rousing benchmark for the data integration based ranks.

It has been a pleasure to investigate functions of HuR-protein with the great help of Ari Ristimäki and Mira Heinonen. I own my compliment to Ari for his elegant introductions to pathology and epithelial cancer.

I am thankful for Pirjo Laakkonen and Piia-Riitta Karhemo for their research in metastasis related proteins (Karhemo et al., 2012) and for sharing that data with us. The integration of that information with other cancer studies and patient survival information enables various lines of research we are still busy with.

I want to express my gratitude for Henk Stunnenberg and George Reid for the visits to their sequencing laboratories in the Nijmegen Centre for Molecular Life Sciences, Nijmegen, NL and in the Institute of Molecular Biology, Mainz, GE, respectively. Both visits were great opportunities to apply and to test Moksiskaan and Anduril against novel data sets and to learn how the sequencing data is produced in practise.

I am looking forward for the hypergraph based biodatabases such as the one Spyro Mousses, Preston Lee, Toni Farley, and Jeff Kiefer are developing in the Translational Genomics Research Institute, Phoenix, US. I am happy that I got a change to visit Spyro's laboratory and I hope that we will co-operate on many projects in future.

Open source software and public databases have played an important role in this work, which would have been otherwise impossible for us. I want to acknowledge all the people behind these programs and services for their invaluable contribution to the scientific work presented. I eulogise donors of tissue samples (including the ancestries of cell lines) enabling the research. To these anonymous yet indispensable individuals I dedicate this book.

In the end, my compliments to my beloved family, Noora and Vilja, for their patience during the long days I have spent at work. I am indebted to you and to my parents, Eila and Risto, for the indefatigable encouragement and aid that they have provided.

Marko Laakso
Helsinki, 5.9.2012

References

Aaltonen, L., Johns, L., Järvinen, H., Mecklin, J., and Houlston, R. (2007). Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clinical Cancer Research*, 13(1):356–361.

Aaltonen, L., Salovaara, R., Kristo, P., Canzian, F., Hemminki, A., Peltomäki, P., Chadwick, R., Kääriäinen, H., Eskelinen, M., Järvinen, H., et al. (1998). Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *New England Journal of Medicine*, 338(21):1481–1487.

Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3):243–255.

Almeida, J. (2010). Computational ecosystems for data-driven medical genomics. *Genome Medicine*, 2(9):67.

Aparicio, O., Geisberg, J., Struhl, K., et al. (2004). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Current Protocols in Cell Biology*, pages Unit–17.

Aranda, A. and Pascual, A. (2001). Nuclear hormone receptors and gene expression. *Physiological Reviews*, 81(3):1269–1304.

Bader, G., Cary, M., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(suppl 1):D504–D506.

Bailey, T. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. AAAI Press.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(Web Server issue):W369.

Bauer-Mehren, A., Furlong, L., and Sanz, F. (2009). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular Systems Biology*, 5(1).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Bentley, D., Balasubramanian, S., Swerdlow, H., Smith, G., Milton, J., Brown, C., Hall, K., Evers, D., Barnes, C., Bignell, H., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Beroukhim, R., Lin, M., Park, Y., Hao, K., Zhao, X., Garraway, L., Fox, E., Hochberg, E., Mellinghoff, I., Hofer, M., et al. (2006). Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Computational Biology*, 2(5):e41.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., Mc Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y.-J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tabernero, J., Baselga, J., Tsao, M.-S., Demichelis, F., Rubin, M. A., Janne, P. A., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True,

L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R., and Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905.

Bilgüvar, K., Öztürk, A., Louvi, A., Kwan, K., Choi, M., Tatlı, B., Yalnızoğlu, D., Tüysüz, B., Çağlayan, A., Gökben, S., et al. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*, 467(7312):207–210.

Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Cawkwell, L., Lewis, F., and Quirke, P. (1994). Frequency of allele loss of DCC, p53, RBI, WT1, NF1, NM23 and APC/MCC in colorectal cancer assayed by fluorescent multiplex polymerase chain reaction. *British Journal of Cancer*, 70(5):813.

Ceol, A., Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(suppl 1):D532–D539.

Cerami, E., Gross, B., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G., and Sander, C. (2010). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Research*.

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(18):e123–e123.

Chen, R., Mias, G., Li-Pook-Than, J., Jiang, L., Lam, H., Chen, R., Miriami, E., Karczewski, K., Hariharan, M., Dewey, F., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307.

Clevers, H. (2011). The cancer stem cell: premises, promises and challenges. *Nature Medicine*, 17(3):313–319.

Colella, S., Yau, C., Taylor, J., Mirza, G., Butler, H., Clouston, P., Bassett, A., Seller, A., Holmes, C., and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, 35(6):2013.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(suppl 1):D691.

Cui, Q., Ma, Y., Jaramillo, M., Bari, H., Awan, A., Yang, S., Zhang, S., Liu, L., Lu, M., O'Connor-McCourt, M., et al. (2007). A map of human cancer signaling. *Molecular Systems Biology*, 3(1).

Curk, T., Demsar, J., Xu, Q., Leban, G., Petrovic, U., Bratko, I., Shaulsky, G., and Zupan, B. (2005). Microarray data mining with visual programming. *Bioinformatics*, 21:396–398.

Cusick, M., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A., Simonis, N., Rual, J., Borick, H., Braun, P., Dreze, M., et al. (2008). Literature-curated protein interaction datasets. *Nature Methods*, 6(1):39–46.

Dancey, J., Bedard, P., Onetto, N., and Hudson, T. (2012). The genetic basis for cancer treatment decisions. *Cell*, 148(3):409–420.

Davey, J., Hohenlohe, P., Etter, P., Boone, J., Catchen, J., and Blaxter, M. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7):499–510.

de la Chapelle, A. (2004). Genetic predisposition to colorectal cancer. *Nature Reviews Cancer*, 4(10):769–780.

Demir, E., Cary, M., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., et al. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942.

D'haeseleer, P. (2006). How does DNA sequence motif discovery work? *Nature Biotechnology*, 24(8):959–961.

Dutt, A. and Beroukhim, R. (2007). Single nucleotide polymorphism array analysis of cancer. *Current Opinion in Oncology*, 19(1):43.

Easton, D., Pooley, K., Dunning, A., Pharoah, P., Thompson, D., Ballinger, D., Struewing, J., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093.

Eronen, L. and Toivonen, H. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13(1):119.

Evans, C. (2011). Scientists develop new database that provides comprehensive view of glioblastoma multiforme genome. The Cancer Genome Atlas Research Briefs http://cancergenome.nih.gov/researchhighlights/researchbriefs/newdatabase.

Finnish Cancer Registry (2012). Ajantasaiset perustaulukot — Syöpäjärjestöt. http://www.cancer.fi/syoparekisteri/tilastot/ajantasaiset-perustaulukot/.

Flicek, P., Amode, M., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2011). Ensembl 2011. *Nucleic Acids Research*, 39(suppl 1):D800.

Forbes, S., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J., Futreal, P., and Stratton, M. (2008). The catalogue of somatic mutations in cancer (COSMIC). *Current Protocols in Human Genetics*, Chapter 10.

Frazer, K., Ballinger, D., Cox, D., Hinds, D., Stuve, L., Gibbs, R., Belmont, J., Boudreau, A., Hardenbol, P., Leal, S., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

Fritz, A. (2000). *International classification of diseases for oncology: ICD-O*. World Health Organization.

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D., Liu, P., Gautam, B., Ly, S., Guo, A., et al. (2010). SMPDB: the small molecule pathway database. *Nucleic Acids Research*, 38(suppl 1):D480–D487.

Galperin, M. and Fernández-Suárez, X. (2012). The 2012 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 40(D1):D1–D8.

Gao, N., Zhang, J., Rao, M., Case, T., Mirosevich, J., Wang, Y., Jin, R., Gupta, A., Rennie, P., and Matusik, R. (2003). The role of hepatocyte nuclear factor-3α (forkhead box a1) and androgen receptor in transcriptional regulation of prostatic genes. *Molecular Endocrinology*, 17(8):1484–1507.

Gardina, P., Clark, T., Shimada, B., Staples, M., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., et al. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, 7(1):325.

Gerlinger, M., Rowan, A., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892.

Gillis, J. and Pavlidis, P. (2012). "guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3):e1002444.

Gräf, S., Nielsen, F., Kurtz, S., Huynen, M., Birney, E., Stunnenberg, H., and Flicek, P. (2007). Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*, 23(13):i195–i204.

Hanahan, D. and Weinberg, R. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

Hanspers, K., Pandey, A., Pico, A., Waagmeester, A., et al. (2011). Androgen receptor signaling pathway (Homo sapiens). http://www.wikipathways.org/index.php?title=Pathway:WP138&oldid=44951.

Heemers, H. and Tindall, D. (2007). Androgen receptor (AR) coregulators: a diversity of functions converging on and regulating the AR transcriptional complex. *Endocrine Reviews*, 28(7):778–808.

Heinonen, M., Hemmes, A., Salmenkivi, K., Abdelmohsen, K., Vilén, S., Laakso, M., Leidenius, M., Salo, T., Hautaniemi, S., Gorospe, M., et al. (2011). Role of RNA binding protein HuR in ductal carcinoma in situ of the breast. *The Journal of Pathology*, 224(4):529–539.

Ho, J., Bishop, E., Karchenko, P., Nègre, N., White, K., and Park, P. (2011). ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC Genomics*, 12(1):134.

Hoheisel, J. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nature reviews genetics*, 7(3):200–210.

Houlston, R. and Peto, J. (2004). The search for low-penetrance cancer susceptibility alleles. *Oncogene*, 23(38):6471–6476.

Hu, M., Yu, J., Taylor, J., Chinnaiyan, A., and Qin, Z. (2010). On the detection and refinement of transcription factor binding sites using chip-seq data. *Nucleic Acids Research*, 38(7):2154–2167.

Huang, W., Sherman, B., Stephens, R., Baseler, M., Lane, H., Lempicki, R., et al. (2008). DAVID gene ID conversion tool. *Bioinformation*, 2(10):428.

Huang, Y., Boynton, R., Blount, P., Silverstein, R., Yin, J., Tong, Y., McDaniel, T., Newkirk, C., Resau, J., Sridhara, R., et al. (1992). Loss of heterozygosity involves multiple tumor suppressor genes in human esophageal cancers. *Cancer Research*, 52(23):6525.

Hurtado, A., Holmes, K., Ross-Innes, C., Schmidt, D., and Carroll, J. (2010). FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genetics*, 43(1):27–33.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Jensen, L., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., et al. (2009). STRING 8 — a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl 1):D412–D416.

Johnson, R. (1997). Components, frameworks, patterns. *ACM SIGSOFT Software Engineering Notes*, 22(3):10–17.

Jørgensen, C. and Linding, R. (2010). Simplistic pathways or complex networks? *Current Opinion in Genetics & Development*, 20(1):15–22.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114.

Karhemo, P., Ravela, S., Laakso, M., Ritamo, I., Tatti, O., Mäkinen, K., Goodison, S., Stenman, U., Hölttä, E., Hautaniemi, S., Valmu, L., Lehti, K., and Laakkonen, P. (2012). An optimized isolation of biotinylated cell surface proteins reveals novel players in cancer metastasis. *Journal of Proteomics*. In print.

Karinen, S., Saarinen, S., Lehtonen, R., Rastas, P., Vahteristo, P., Aaltonen, L., and Hautaniemi, S. (2012). Rule-based induction method for haplotype comparison and identification of candidate disease loci. *Genome Medicine*, 4(3):21.

Kathiresan, S., Voight, B., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P., Anand, S., Engert, J., Samani, N., Schunkert, H., et al. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics*, 41(3):334–341.

Kelder, T., van Iersel, M., Hanspers, K., Kutmon, M., Conklin, B., Evelo, C., and Pico, A. (2012). Wikipathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., et al. (2007). Intact — open source resource for molecular interaction data. *Nucleic Acids Research*, 35(suppl 1):D561–D565.

Kharchenko, P., Tolstorukov, M., and Park, P. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12):1351–1359.

Kinzler, K. and Vogelstein, B. (1996). Lessons from hereditary review colorectal cancer. *Cell*, 87:159–170.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2011). DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, 39(suppl 1):D1035.

Laakso, M. (2007). Computational Identification of Recessive Mutations in Cancers Using High Throughput SNP-arrays. Master's thesis, University of Helsingin, Finland. http://urn.fi/URN:NBN:fi-fe20071065.

Laakso, M., Sahu, B., Ovaska, K., Jänne, O., and Hautaniemi, S. (2011). Androgen Receptor and FoxA1 Interaction Study (Anduril report). http://www.ltdk.helsinki.fi/sysbio/csb/FOXA1.pdf.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, 37(13):4181–4193.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Levsky, J. and Singer, R. (2003). Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science*, 116(14):2833–2838.

Lichtenstein, P., Holm, N., Verkasalo, P., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer — analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, 343(2):78–85.

Liikanen, I., Monsurrò, V., Ahtiainen, L., Raki, M., Hakkarainen, T., Diaconu, I., Escutenaire, S., Hemminki, O., Dias, J., Cerullo, V., et al. (2011). Induction of interferon pathways mediates in vivo resistance to oncolytic adenovirus. *Molecular Therapy*, 19(10):1858–1866.

Louhimo, R., Aittomäki, V., Faisal, A., Laakso, M., Chen, P., Ovaska, K., Valo, E., Lahti, L., Rogojin, V., Kaski, S., and Hautaniemi, S. (2012). Systematic use of computational methods allows stratifying treatment responders in glioblastoma multiforme. Submitted. http://csbi.ltdk.helsinki.fi/camda/.

Louis, D., Ohgaki, H., Wiestler, O., and Cavenee, W. (2007). *WHO classification of tumours of the central nervous system*. World Health Organization classification of tumours. International Agency for Research on Cancer.

Ma'ayan, A., Jenkins, S., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N., Weng, G., Ram, P., Rice, J., et al. (2005). Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science*, 309(5737):1078.

Madsen, B., Villesen, P., and Wiuf, C. (2007). A periodic pattern of SNPs in the human genome. *Genome Research*, 17(10):1414–1419.

Maglott, D., Ostell, J., Pruitt, K., and Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 35(suppl 1):D26–D31.

Makkonen, H., Jääskeläinen, T., Pitkänen-Arsiola, T., Rytinki, M., Waltering, K., Mättö, M., Visakorpi, T., and Palvimo, J. (2008). Identification of ETS-like transcription factor 4 as a novel androgen receptor target in prostate cancer cells. *Oncogene*, 27(36):4865–4876.

Mao, X., Young, B., and Lu, Y. (2007). The application of single nucleotide polymorphism microarrays in cancer research. *Current Genomics*, 8(4):219.

Marioni, J., Mason, C., Mane, S., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.

Martini, M., Vecchione, L., Siena, S., Tejpar, S., and Bardelli, A. (2011). Targeted therapies: how personal should we go? *Nature Reviews Clinical Oncology*, 9(2):87–97.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E., Brat, D., Mastrogianakis, G., Olson, J., Mikkelsen, T., Lehman, N., Aldape, K., et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068.

Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics*, 6(1):44–56.

Moussouni, F., Berti-Équille, L., Rozé, G., Loréal, O., and Guérin, E. (2007). QDex: a database profiler for generic bio-data exploration and quality aware integration. In *Web Information Systems Engineering — WISE 2007 Workshops*, pages 5–16. Springer.

Muzny, D., Bainbridge, M., Chang, K., Dinh, H., Drummond, J., Fowler, G., Kovar, C., Lewis, L., Morgan, M., Newsham, I., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487:330–337.

Nicolae, D., Wen, X., Voight, B., and Cox, N. (2006). Coverage and characteristics of the Affymetrix GeneChip Human Mapping 100K SNP set. *PLoS Genetics*, 2(5):e67.

Nielsen, R., Paul, J., Albrechtsen, A., and Song, Y. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451.

Orvis, J., Crabtree, J., Galens, K., Gussman, A., Inman, J., Lee, E., Nampally, S., Riley, D., Sundaram, J., Felix, V., et al. (2010). Ergatis: a web interface and scalable software system for bioinformatics workflows. *Bioinformatics*, 26(12):1488–1492.

Ouyang, Z., Zhou, Q., and Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, 106(51):21521–21526.

Ouzounis, C. and Karp, P. (2000). Global properties of the metabolic map of *Escherichia coli*. *Genome Research*, 10(4):568–576.

Ovaska, K., Laakso, M., and Hautaniemi, S. (2008). Fast Gene Ontology based clustering for microarray experiments. *BioData Mining*, 1(11).

Ozsolak, F. and Milos, P. (2010). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98.

Papadopoulos, N. and Lindblom, A. (1997). Molecular basis of HNPCC: mutations of MMR genes. *Human Mutation*, 10(2):89–99.

Papin, J., Hunter, T., Palsson, B., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, 6(2):99–111.

Pareek, C., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52(4):413–435.

Park, P. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680.

Partanen, J. (2012). *Epithelial integrity as a tumor suppressor mechanism — The interplay between Lkb1 and c-Myc in breast cancer development*. PhD thesis, Faculty of Medicine, University of Helsinki, Finland. http://urn.fi/URN:ISBN:978-952-10-8007-4.

Pe'er, D. and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, 144(6):864–873.

Philippi, S. and Köhler, J. (2006). Addressing the problems with life-science databases for traditional uses and systems biology. *Nature Reviews Genetics*, 7(6):482–488.

Pisabarro, A., Pérez, G., Lavín, J., and Ramírez, L. (2008). Genetic networks for the functional study of genomes. *Briefings in Functional Genomics & Proteomics*, 7(4):249–263.

Podpečan, V., Lavrač, N., Mozetič, I., Novak, P., Trajkovski, I., Langohr, L., Kulovesi, K., Toivonen, H., Petek, M., Motaln, H., et al. (2011). SegMine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinformatics*, 12(1):416.

Portales-Casamar, E., Thongjuea, S., Kwon, A., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(suppl 1):D105.

Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human protein reference database — 2009 update. *Nucleic Acids Research*, 37(suppl 1):D767–D772.

Quackenbush, J. et al. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427.

Rhead, B., Karolchik, D., Kuhn, R., Hinrichs, A., Zweig, A., Fujita, P., Diekhans, M., Smith, K., Rosenbloom, K., Raney, B., et al. (2010). The UCSC genome browser database: update 2010. *Nucleic Acids Research*, 38(suppl 1):D613–D619.

Riemenschneider, M., Jeuken, J., Wesseling, P., and Reifenberger, G. (2010). Molecular diagnostics of gliomas: state of the art. *Acta Neuropathologica*, pages 1–18.

Romero, P., Wagg, J., Green, M., Kaiser, D., Krummenacker, M., and Karp, P. (2004). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6(1):R2.

Ruths, D., Muller, M., Tseng, J., Nakhleh, L., and Ram, P. (2008). The Signaling Petri Net-Based Simulator: A Non-Parametric Strategy for Characterizing the Dynamics of Cell-Specific Signaling Networks. *PLoS Computational Biology*, 4(2):e1000005.

Salovaara, R., Loukola, A., Kristo, P., Kääriäinen, H., Ahtola, H., Eskelinen, M., Härkönen, N., Julkunen, R., Kangas, E., Ojala, S., et al. (2000). Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *Journal of Clinical Oncology*, 18(11):2193.

Salwinski, L., Miller, C., Smith, A., Pettit, F., Bowie, J., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(suppl 1):D449–D451.

Schilling, C., Schuster, S., Palsson, B., and Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnology Progress*, 15(3):296–303.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12(3):329.

Shi, L., Campbell, G., Jones, W., Campagne, F., Wen, Z., Walker, S., Su, Z., Chu, T., Goodsaid, F., Pusztai, L., et al. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28(8):827.

Sjöblom, T., Jones, S., Wood, L., Parsons, D., Lin, J., Barber, T., Mandelker, D., Leary, R., Ptak, J., Silliman, N., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268.

Sreenivasaiah, P., Rani, S., Cayetano, J., Arul, N., et al. (2012). IPAVS: integrated pathway resources, analysis and visualization system. *Nucleic Acids Research*, 40(D1):D803–D808.

Stark, C., Breitkreutz, B., Chatr-aryamontri, A., Boucher, L., Oughtred, R., Livstone, M., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al. (2011). The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(suppl 1):D698.

Strömbäck, L. and Lambrix, P. (2005). Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, 21(24):4401–4407.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Eberta, B., Gillettea, M., Paulovichg, A., Pomeroyh, S., Goluba, T., Landera, E., and Mesirova, J. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Journal of Computer Science & Systems Biology*, 102(43):15545–15550.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382.

Tarca, A., Draghici, S., Khatri, P., Hassan, S., Mittal, P., Kim, J., Kim, C., Kusanovic, J., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75.

The UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40(D1):D71–D75.

Thomas, A., Camp, N., Farnham, J., Allen-Brady, K., and Cannon-Albright, L. (2008). Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Annals of Human Genetics*, 72(2):279–287.

Turenne, N. (2011). Role of a web-based software platform for systems biology. *Journal of Computer Science & Systems Biology*, 4:035–041.

Van der Ploeg, M. (2009). Cytochemical nucleic acid research during the twentieth century. *European Journal of Histochemistry*, 44(1):7–42.

Vandin, F., Upfal, E., and Raphael, B. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Research*, 22(2):375–385.

Vogelstein, B. and Kinzler, K. (2004). Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M., Ohgi, K., et al. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by erna. *Nature*, 474(7351):390–394.

Wang, Q., Li, W., Liu, X., Carroll, J., Jänne, O., Keeton, E., Chinnaiyan, A., Pienta, K., and Brown, M. (2007). A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. *Molecular Cell*, 27(3):380–392.

Webb, E. et al. (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Number Ed. 6. Academic Press.

Weller, M., Felsberg, J., Hartmann, C., Berger, H., Steinbach, J., Schramm, J., Westphal, M., Schackert, G., Simon, M., Tonn, J., et al. (2009). Molecular predictors of progression-free and overall survival in patients with newly diagnosed glioblastoma: a prospective translational study of the German Glioma Network. *Journal of Clinical Oncology*, 27(34):5743.

Wheeler, D., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y., Makhijani, V., Roth, G., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876.

WHO (2011). *Global status report on noncommunicable diseases 2010*. World Health Organization, Geneva.

Wilbanks, E. and Facciotti, M. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 5(7):e11471.

Wilkes, T., Laux, H., and Foy, C. (2007). Microarray data quality-review of current developments. *OMICS: A Journal of Integrative Biology*, 11(1):1–13.

Wu, J., Smith, L., Plass, C., and Huang, T. (2006). ChIP-chip comes of age for genome-wide functional analysis. *Cancer Research*, 66(14):6899.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T., and Hautaniemi, S. (2008). Integrated network analysis platform for protein-protein interactions. *Nature Methods*, 6(1):75–77.

Yosef, N., Zalckvar, E., Rubinstein, A., Homilius, M., Atias, N., Vardi, L., Berman, I., Zur, H., Kimchi, A., Ruppin, E., et al. (2011). ANAT: A Tool for Constructing and Analyzing Functional Protein Networks. *Science's STKE*, 4(196):pl1.

Yue, P., Melamud, E., and Moult, J. (2006). SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics*, 7(1):166.

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nussbaum, C., Myers, R., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137.

Conflicts of interest

Marko Laakso is one of the founders, board members, and shareholders of Significo Research Ltd (http://www.significo.fi/), which provides life science data analysis and business intelligence services. The company is selling analysis services partly based on the methods, such as Anduril and Moksiskaan, described in this thesis.